Submitted to the Annals of Statistics

ARK: ROBUST KNOCKOFFS INFERENCE WITH COUPLING

BY YINGYING FAN[1,a], LAN GAO[2,c] AND JINCHI LV[1,b]

[1]*Data Sciences and Operations Department, Marshall School of Business, University of Southern California,*
[a]*fanyingy@marshall.usc.edu;* [b]*jinchilv@marshall.usc.edu*

[2]*Department of Business Analytics and Statistics, Haslam College of Business, The University of Tennessee ,* [c]*lgao13@utk.edu*

We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and k-familywise error rate (k-FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or k-FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework. Additionally, we formally connect our concept of knockoff variable coupling to a type of Wasserstein distance.

1. Introduction. The knockoffs inference framework (Barber and Candès, 2015; Candès et al., 2018; Barber and Candès, 2019) is a powerful innovative tool for feature selection with controlled error rates. In particular, the model-X knockoffs (Candès et al., 2018) achieves the false discovery rate (FDR) control at a predetermined level in finite samples without requiring any specific model assumptions on how the response depends on the features, making it an attractive option for feature selection in a wide range of statistical applications. The fundamental idea of the knockoffs procedure is to construct knockoff variables that are exchangeable in distribution with the original features but are independent of the response conditional on the original variables. These knockoff variables serve as a control group for the original features, allowing researchers to identify relevant original features for the response. The model-X knockoffs inference has gained increasing popularity since its inception and there have been flourishing developments and extensions of the knockoffs framework and spirits, such as the k-familywise error rate (k-FWER) control with knockoffs (Janson and Su, 2016), power analysis for knockoffs procedure (Fan et al., 2020a; Spector and Janson, 2022; Wang and Janson, 2022; Weinstein et al., 2020; Fan et al., 2020b), derandomized knockoffs (Ren, Wei and Candès, 2021; Ren and Barber, 2022), knockoffs inference for time series data (Chi et al., 2023), kernel knockoffs procedure (Dai, Lyu and Li, 2022), and FDR control by data splitting or creating mirror variables (Li and Maathuis, 2021; Dai et al., 2022; Cao, Sun and Yao, 2021; Guo et al., 2022).

A key assumption in the model-X knockoffs inference is that the joint distribution of features is known. However, such information is almost never available in practice. There

MSC2020 subject classifications: Primary 62G35, 62F07, 62E17; secondary 62G10, 62F35.
Keywords and phrases: Knockoffs inference, Wasserstein distance, False discovery rate control, Familywise error rate control, Coupling, Robustness, High dimensionality.

has been overwhleming empirical evidence that the model-X knockoffs framework is robust to misspecified or estimated feature distributions (Candès et al., 2018; Sesia, Sabatti and Candès, 2019; Jordon, Yoon and van der Schaar, 2018; Lu et al., 2018; Zhu et al., 2021; Romano, Sesia and Candès, 2020). Yet, the theoretical characterization of its robustness is still largely missing. A notable exception is the recent work of Barber, Candès and Samworth (2020), where it was formally and elegantly shown that the knockoffs data matrix collecting the knockoff variables can be generated from a distribution, which we name as the *working* distribution for the ease of presentation, that is different from the true underlying feature distribution, and that the resulting FDR inflation can be measured by the empirical Kullback–Leibler (KL) divergence between the true conditional distribution $X_j|X_{-j}$ and the working conditional distribution. Here, $X_j \in \mathbb{R}$ stands for the jth feature, $X_{-j} \in \mathbb{R}^{p-1}$ stands for the feature vector with the jth feature removed, and p is the feature dimensionality. Two important assumptions in their analyses for ensuring the asymptotic FDR control are 1) the working distribution should be learned independently from the training data used for feature selection and 2) the empirical KL divergence between the two knockoffs data matrices (of diverging dimensionalities) generated from the working and true distributions, respectively, needs to vanish as the sample size increases. Although their results are general and apply to arbitrary dependence structure of the response on features, these two assumptions do not always describe the practical implementation. Our results in the current paper are free of the two assumptions discussed above.

To put more content into our statements above, especially the one about assumption 2), let us consider the scenario where the true feature matrix has independent and identically distributed (i.i.d.) entries from the t-distribution with ν degrees of freedom, but we misspecify it and use the Gaussian distribution as a working distribution to generate the knockoff variable matrix $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times p}$, where n is the sample size. It can be calculated that the empirical KL divergence between $\widehat{\mathbf{X}}$ and the model-X knockoff variable matrix $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ defined in Barber, Candès and Samworth (2020) has mean and variance both at order $\frac{np}{\nu(\nu+p)}$. Thus, only when $\nu^2 \gg n\min(n,p)$ (which is equivalent to $\frac{np}{\nu(\nu+p)} \to 0$), the FDR inflation as derived therein can vanish asymptotically. In contrast, our theory shows that as long as $\nu^2 \gg s^4(\log p)^{4+4/\gamma}$ for some $\gamma \in (0,1)$ with $s \ll n^{1/2}$ a sparsity parameter, the knockoffs procedure based on the working distribution can achieve the asymptotic FDR control. More details for our results and model assumptions are summarized formally in Section 4.1. We provide additional comparisons of our results with those of Barber, Candès and Samworth (2020) in various parts of the paper where more specifics can be discussed. We emphasize and acknowledge that Barber, Candès and Samworth (2020) established general robustness results without specific model assumptions, while some of our results rely on certain specific model assumptions. The main point we advocate here is that a different notion of closeness than the KL divergence can be advantageous in studying the robustness of the model-X knockoffs. We also formally connect our concept of closeness to a type of Wasserstein distance. We provide detailed comparison with some other existing work in the literature in Section 6.

The major goal of our paper is to establish a general theory on the robustness of the model-X knockoffs framework for the FDR and k-FWER control. We approach the problem by studying the performance of the *approximate knockoffs (ARK)* procedure, an algorithm that is most popularly implemented in practice when applying the knockoffs framework. The ARK procedure differs from the model-X knockoffs in that the former generates the knockoff feature matrix from a working distribution that can be misspecified or learned from the *same* training data for feature selection. By showing that the ARK procedure achieves the asymptotic FDR and k-FWER control as sample size increases, we can verify the robustness of the model-X knockoffs. An important idea in our technical analyses is *coupling*, where we pair the ARK procedure with the model-X knockoffs procedure in such a way that random

variables in these two paired procedures are close in realizations with high probability. Hereafter, we will refer to the model-X knockoffs as the *perfect knockoffs* procedure to emphasize its difference from the approximate knockoffs procedure. It is important to emphasize that we require the realizations of random variables in the paired procedures to be close, instead of the corresponding distributions being close. This is a major distinction from the assumption in Barber, Candès and Samworth (2020). Our new notion of closeness allows us to justify the robustness of the model-X knockoffs in some broader contexts not covered by studies in the existing literature. We also emphasize that although our conditions are imposed on the perfect knockoff variables, we do *not* need to know or construct them in implementation; the existence of such variables is sufficient for our theoretical robustness analyses.

We present our theory by first laying out general conditions on the existence of the coupled perfect knockoff statistics and their closeness to the approximate knockoff statistics in Section 2, and then provide examples justifying these conditions in Sections 3 and 4. More specifically, our theory has three layers, related to different stages in applying the knockoffs inference procedure. Our preliminary theory in Section 2 directly makes assumptions on the quality of the approximate knockoff statistics (cf. (3)) by requiring the existence and closeness of their coupled perfect knockoff statistics. Then under some regularity conditions imposed on the distribution of these perfect knockoff statistics, we prove that the FDR and k-FWER are controlled asymptotically using the approximate knockoff statistics. This lays the theoretical foundation for our subsequent analyses in Sections 3 and 4.

The second layer of our theory, presented in Section 3, delves deeper and replaces the coupling condition imposed on the knockoff statistics in Section 2 with a coupling condition on the approximate knockoff variables generated from some mispecified or estimated feature distribution. Similar in nature to the coupling condition in Section 2, this new condition assumes that there exist perfect knockoff variables that can be coupled with approximate knockoff variables so that their realizations are close to each other with high probability. Since knockoff statistics are known functions of knockoff variables, such alternative condition intuitively and naturally leads to the verification of the coupling condition on knockoff statistics in Section 2. Indeed, we showcase using two commonly analyzed knockoff statistics, namely the marginal correlation statistics and the regression coefficient difference (RCD) statistics, that the coupling condition on knockoff variables can guarantee the coupling condition on knockoff statistics. We also verify that for each of these two constructions of knockoff statistics, the other regularity conditions in our preliminary theory in Section 2 also hold, ensuring the asymptotic FDR and k-FWER control. Notably, our theory also reveals that, the marginal correlation is of "low accuracy," and needs more stringent conditions than RCD to achieve asymptotic FDR control. This message is consistent with Niu et al. (2024) when studying the conditional randomization test using the model-X framework.

The last layer of our theory is presented in Section 4 and showcases three specific constructions of the coupled perfect knockoff variables. By imposing conditions on the misspecified or estimated feature distribution, we construct explicitly the coupled perfect knockoff variables and prove that the coupling conditions in the first and second layers of our general theory are satisfied. This gives us a complete theory with conditions imposed on the working distribution for generating knockoff variables and verifies the robustness of the model-X knockoffs inference procedure. Our theory allows high dimensionality of features and allows *in-sample* estimation of the feature distribution.

The rest of the paper is organized as follows. Section 2 first introduces the approximate knockoffs procedure and then presents the general conditions and theory for the asymptotic FDR control. We also introduce the coupling idea, a key technique in our theoretical analyses. We illustrate our general theory using two commonly used constructions of knockoff statistics in Section 3. Section 4 further provides three specific constructions of the coupled perfect

knockoff variables. We present companion theory for robust k-FWER control in Section 5. We provide detailed discussions on some most related works in Section 6, and present some simulation examples in Section 7. We conclude our paper by summarizing the key results and discussing some future research directions in Section 8. All the proofs and technical details are provided in the Supplementary Material.

To facilitate the technical presentation, let us introduce some notation that will be used throughout the paper. We use $a_n \ll b_n$ or $a_n = o(b_n)$ to represent $a_n/b_n \to 0$, $a_n \gg b_n$ to represent $a_n/b_n \to \infty$, and $a_n \lesssim b_n$ or $a_n = O(b_n)$ to represent $a_n \leq Cb_n$ for an absolute constant $C > 0$. Let $a \wedge b$ and $a \vee b$ be the minimal and maximal values of a and b, respectively. For a vector $\mathbf{x} \in \mathbb{R}^p$, denote by $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$, and $\|\mathbf{x}\|_0$ the ℓ_1-norm, ℓ_2-norm, and ℓ_0-norm, respectively. For $1 \leq j \leq p$, $\mathbf{x}_j$ is the jth component of $\mathbf{x}$ and $\mathbf{x}_{-j}$ is a subvector of $\mathbf{x}$ with the jth component removed. For a matrix $\mathbf{M} \in \mathbb{R}^{n\times p}$, denote by $\mathbf{M}_{i,j}$ the (i,j)th entry of $\mathbf{M}$, $\mathbf{M}_j$ the jth column of $\mathbf{M}$, and $\mathbf{M}_{A_1,A_2}$ a submatrix of $\mathbf{M}$ consisting of $(\mathbf{M}_{i,j})_{i\in A_1, j\in A_2}$ for sets $A_1 \subset \{1,\cdots,n\}$ and $A_2 \subset \{1,\cdots,p\}$. Let $\|\mathbf{M}\|_{\max}$ and $\|\mathbf{M}\|_2$ be the maximum norm and spectral norm of a matrix $\mathbf{M}$, respectively. For $1 \leq j \leq p$, $-j$ represents the set $\{1,\cdots,p\} \setminus \{j\}$, and denote by $|\mathcal{A}|$ the cardinality of set $\mathcal{A}$. For a positive definite matrix $\mathbf{\Sigma}$, let $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ be the smallest and largest eigenvalues of $\mathbf{\Sigma}$, respectively.

2. Preliminary results on robust knockoffs inference via coupling.

2.1. *Model setup and model-X knockoffs framework.* Assume that we have n i.i.d. observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from the population (X, Y), where $X = (X_1,\cdots,X_p)^T$ is the p-dimensional feature vector and $Y \in \mathbb{R}$ is a scalar response. Here, the feature dimensionality p can diverge with the sample size n. Adopting the matrix notation, the n i.i.d. observations can be written as the data matrix $\mathbf{X} = (\mathbf{X}_{i,j}) \in \mathbb{R}^{n\times p}$ collecting the values of all the features and vector $\mathbf{y} = (y_1,\cdots,y_n)^T \in \mathbb{R}^n$ collecting the values of the response. A feature X_j is defined as null (or irrelevant) if and only if it is independent of the response conditional on all the remaining features; that is, $Y \perp\!\!\!\perp X_j | X_{-j}$, where X_{-j} is a subvector of X with the jth component removed. Denote by $\mathcal{H}_0 = \{1 \leq j \leq p : X_j \text{ is a null feature}\}$ the set of null features and $\mathcal{H}_1 = \mathcal{H}_0^c$ that of nonnull (or relevant) features. To ensure the model identifiability and interpretability, we follow Candès et al. (2018) and assume that $\mathcal{H}_1$ exists and is unique. Further assume that the subset of relevant features is sparse such that $p_1 = |\mathcal{H}_1| = o(n \wedge p)$, where $|\mathcal{A}|$ stands for the cardinality of a given set. The goal is to select as many relevant features as possible while controlling some error rate measure at the prespecified target level.

A commonly used measure for evaluating the feature selection performance is FDR (Benjamini and Hochberg, 1995), where for an outcome $\widehat{S}$ of some feature selection procedure, the FDR is defined as

$$\text{FDR} = \mathbb{E}[\text{FDP}] \text{ with } \text{FDP} = |\widehat{S} \cap \mathcal{H}_0|/|\widehat{S}|. \tag{1}$$

The model-X knockoffs framework provides a flexible way for controlling the FDR at some prespecified target level in finite samples (Candès et al., 2018), allowing arbitrary dimensionality of X and arbitrary dependence between response Y and feature vector X. A key step of the model-X knockoffs inference (Candès et al., 2018) is to generate the model-X knockoff variables $\widetilde{X} = (\widetilde{X}_1,\cdots,\widetilde{X}_p)^T$ such that $\widetilde{X} \perp\!\!\!\perp Y | X$ and

$$(X, \widetilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \widetilde{X}) \text{ for each subset } S \subset \{1,\cdots,p\}, \tag{2}$$

where $(X, \widetilde{X})_{\text{swap}(S)}$ is obtained by swapping the components X_j and $\widetilde{X}_j$ in $(X, \widetilde{X})$ for each $j \in S$.

The construction of the model-X knockoff variables, which we will refer to as the *perfect* knockoff variables in future presentation, requires the exact knowledge of the distribution

of feature vector X. For example, Algorithm 1 in Candès et al. (2018) provided a general approach to generating the perfect knockoff variables when such information is available. However, the exact knowledge of feature distribution is usually unavailable in real applications. Thus, in practical implementation, the problem becomes identifying the relevant subset $\mathcal{H}_1$ with the approximate knockoff variables generated from a feature distribution that can be different from the true underlying one; we name the practical procedure as the *approximate* knockoffs and formally present it in the next section for completeness. As stated in the Introduction, we study the robustness of the model-X knockoffs procedure by investigating the feature selection performance of the approximate knockoffs procedure.

2.2. *Approximate knockoffs and a roadmap of our analysis.* In practice, the approximate knockoffs inference procedure below is implemented popularly for controlling the FDR.

1) *Generating approximate knockoff variables.* Since the true underlying feature distribution $F(\cdot)$ is generally unavailable, we generate the knockoff variables from some user-specified feature distribution $\widehat{F}(\cdot)$, which can depend on the sample $(\mathbf{X}, \mathbf{y})$, using the same algorithm proposed for generating the perfect knockoff variables (e.g., Algorithm 1 in Candès et al. (2018)). Denote by $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_{i,j}) \in \mathbb{R}^{n\times p}$ the resulting approximate knockoff variable matrix.
2) *Constructing approximate knockoff statistics.* Pretend that $\widehat{\mathbf{X}}$ were perfect knockoff variable matrix and follow the same procedure as in Candès et al. (2018) to calculate the knockoff statistics $\widehat{W}_j$ with $j = 1, \cdots, p$. Specifically, we first compute the feature importance statistics
$$(Z_1, \cdots, Z_p, \widehat{Z}_1, \cdots, \widehat{Z}_p)^T = t((\mathbf{X}, \widehat{\mathbf{X}}), \mathbf{y}),$$
where $t(\cdot)$ is a measurable function of input $((\mathbf{X}, \widehat{\mathbf{X}}), \mathbf{y})$, and Z_j and $\widehat{Z}_j$ measure the importance of the jth feature and its approximate knockoff counterpart relative to the response, respectively. Then the approximate knockoff statistic $\widehat{W}_j$ for the jth feature is defined as
$$\widehat{W}_j = f_j(Z_j, \widehat{Z}_j), \tag{3}$$
where $f_j(\cdot, \cdot)$ is an antisymmetric function satisfying $f_j(x, y) = -f_j(y, x)$. See Barber and Candès (2015) for examples and characterizations on the valid construction of knockoff statistics.
3) *Selecting relevant features*. Calculate a data-driven threshold T for the knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$ and select the set of important features as $\widehat{S} = \{1 \le j \le p : \widehat{W}_j \ge T\}$. Denoting $\widehat{\mathcal{W}} = \{|\widehat{W}_1|, \cdots, |\widehat{W}_p|\}$, the threshold for FDR control is defined as
$$T = \min\left\{t \in \widehat{\mathcal{W}} : \frac{\#\{j : \widehat{W}_j \le -t\}}{\#\{\widehat{W}_j \ge t\} \vee 1} \le q\right\} \tag{4}$$
where $q \in (0, 1)$ is the prespecified level for the FDR.

It is seen that the only difference of the algorithm above from the perfect knockoffs procedure (Candès et al., 2018) is how the knockoff variable matrix $\widehat{\mathbf{X}}$ is generated. The perfect knockoffs procedure based on the true feature distribution $F(\cdot)$ has been shown to control the FDR at the target level (Candès et al., 2018). For the approximate knockoffs inference, however, it is reasonable to expect some inflation in the FDR control, and the inflation level depends on the qualities of both the approximate knockoff variable matrix $\widehat{\mathbf{X}}$ and the resulting knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$. A desired property is that as the approximate knockoff statistics "approach" the perfect knockoff statistics, the level of inflation also vanishes. One contribution of our paper is to formally introduce a notion of *closeness* measuring the qualities

of the approximate knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$ and knockoff variable matrix $\widehat{\mathbf{X}}$. As will be discussed in Section 3.4, our closeness measure is closely related to a type of Wasserstein distance.

We provide a roadmap of our technical analyses. Our theory has three layers, corresponding reversely to the steps in the approximate knockoffs procedure described above. To put it into more content, note that the set of selected features $\widehat{S}$ is defined directly as a function of the approximate knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$. Hence, given $\{\widehat{W}_j\}_{j=1}^p$, feature selection can be conducted without the knowledge of $\widehat{\mathbf{X}}$ or the feature distribution $F(\cdot)$. For this reason, our layer 1 analysis concerns the quality of $\{\widehat{W}_j\}_{j=1}^p$ for achieving the asymptotic FDR control; see Section 2.3 for a characterization on qualified knockoff statistics. The second layer of our analysis studies the quality of $\widehat{\mathbf{X}}$ and is built on the first layer. We characterize what kind of $\widehat{\mathbf{X}}$ can lead to qualified knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$ satisfying the conditions established in our layer 1 analysis; see Section 3 for such analysis in our layer 2. Our layer 3 analysis is built on the first two layers and goes all the way to the root of the knockoffs inference; we provide specific examples and conditions on $\widehat{F}(\cdot)$ for ensuring that $\widehat{\mathbf{X}}$ satisfies conditions in our layer 2 analysis. The key idea empowering our theoretical investigation is variable coupling behind the approximate knockoffs (ARK) procedure; we formally introduce such idea in the next subsection for laying out preliminary results for our subsequent in-depth analysis.

2.3. *Layer 1 analysis: knockoff statistics coupling.* An important observation is that the perfect knockoff variables in the model-X knockoffs framework are not unique. Consequently, the knockoff statistics are not unique either. Indeed, even with the same algorithm (e.g., Algorithm 1 in Candès et al. (2018)), the knockoff variables generated from different runs of the algorithm are only identically distributed. Our coupling idea is deeply rooted on such observation. Let us introduce some additional notation to facilitate our formal presentation of the general theory. Following the model-X knockoffs framework, for a realization of the perfect knockoff variable matrix $\widetilde{\mathbf{X}}$ generated from the true feature distribution $F(\cdot)$, we let

$$(Z_1^*, \cdots, Z_p^*, \widetilde{Z}_1, \cdots, \widetilde{Z}_p)^T = t((\mathbf{X}, \widetilde{\mathbf{X}}), \mathbf{y})$$

and define the perfect knockoff statistics $\widetilde{W}_j = f_j(Z_j^*, \widetilde{Z}_j)$ for $1 \le j \le p$, where functions $t(\cdot)$ and $f_j(\cdot)$ are identical to the ones in the approximate knockoffs procedure in Section 2.2.

We now establish preliminary theory on the asymptotic FDR control for the approximate knockoffs inference procedure, with regularity conditions imposed on the $\widehat{W}_j$ values.

CONDITION 1 (Coupling accuracy). *There exist perfect knockoff statistics $\{\widetilde{W}_j\}_{j=1}^p$ such that for some sequence $b_n \to 0$,*

$$\mathbb{P}\big(\max_{1\le j\le p} |\widehat{W}_j - \widetilde{W}_j| \ge b_n\big) \to 0. \tag{5}$$

Conditions on the convergence rate b_n for ensuring the asymptotic FDR control will be specified in the subsequent assumptions. Condition 1 above couples each realization of the approximate knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$ with a realization of the perfect knockoff statistics $\{\widetilde{W}_j\}_{j=1}^p$, and they need to be sufficiently close to each other with high probability. Note that the existence of such $\{\widetilde{W}_j\}_{j=1}^p$ is required only for the theory, whereas the implementation uses only $\{\widehat{W}_j\}_{j=1}^p$. We will provide examples in later sections verifying the existence of such coupled $\{\widetilde{W}_j\}_{j=1}^p$. The two conditions below are on the quality of the perfect knockoff statistics $\{\widetilde{W}_j\}_{j=1}^p$ and the signal strength in the data as measured by $\widetilde{W}_j$'s.

CONDITION 2 (Average concentration of $\widetilde{W}_j$). *There exist deterministic quantities* $\{w_j\}_{j=1}^p$ *such that* $p^{-1}\sum_{j=1}^p \mathbb{P}(|\widetilde{W}_j - w_j| \geq \delta_n) = o(p^{-1})$, *where* $\delta_n \to 0$ *is a sequence satisfying* $\delta_n \geq b_n$.

CONDITION 3 (Signal strength). *Let* $\mathscr{A}_n = \{j \in \mathcal{H}_1 : w_j \geq 5\delta_n\}$. *It holds that* $a_n = |\mathscr{A}_n| \to \infty$ *and* $w_j > -\delta_n$ *for* $j \in \mathscr{A}_n^c$.

As discussed in Barber and Candès (2015) and Candès et al. (2018), a desired property of the knockoff statistics is to have a large and positive value for $\widetilde{W}_j$ if $j \in \mathcal{H}_1$, and a small and symmetric around zero value for $\widetilde{W}_j$ if $j \in \mathcal{H}_0$. Conditions 2 and 3 together formalize this property. Condition 2 requires that each perfect knockoff statistic $\widetilde{W}_j$ is concentrated around some population parameter w_j with rate δ_n in an average probability sense. By design, $\widetilde{W}_j$'s and w_j's are feature importance measures, and Conditions 2 and 3 characterize the desired properties they need to possess. Note that there is no requirement that each individual w_j with $j \in \mathcal{H}_1$ is positive and large; we only need that there exist enough number (i.e., a_n) of w_j's with $j \in \mathcal{H}_1$ that are positive and large enough. Implicitly, $a_n \to \infty$ requires that the number of relevant features $|\mathcal{H}_1|$ diverges with sample size as well. The condition $\delta_n \geq b_n$ requires that the coupling accuracy b_n should not exceed the order of concentration error so that $\widehat{W}_j$'s are as good as $\widetilde{W}_j$'s for estimating the population quantities w_j's.

Define $p_0 = |\mathcal{H}_0|$ and $G(t) = p_0^{-1}\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$. By Candès et al. (2018), the perfect knockoff statistics $\widetilde{W}_j$ with $j \in \mathcal{H}_0$ are symmetrically distributed around zero. It follows that $G(t) = p_0^{-1}\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \leq -t)$. We need to impose the technical conditions below on the distribution of the perfect knockoff statistics for our robustness analysis.

CONDITION 4 (Weak dependence among nulls). *For some constants* $0 < \gamma < 1$, $0 < c_1 < 1$, $C_1 > 0$, *and a positive sequence* $m_n = o(a_n)$, *it holds that*

$$\mathrm{Var}\Big(\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j > t)\Big) \leq C_1 m_n p_0 G(t) + o\big((\log p)^{-1/\gamma}[p_0 G(t)]^2\big) \tag{6}$$

uniformly over $t \in (0, G^{-1}(\frac{c_1 q a_n}{p})]$.

CONDITION 5 (Distribution of $\widetilde{W}_j$). *Assume that* $G(t)$ *is a continuous function. For the same constants* γ *and* c_1 *as in Condition 4, it holds that as* $n \to \infty$,

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p})]} \frac{G(t-b_n) - G(t+b_n)}{G(t)} \to 0 \tag{7}$$

and

$$a_n^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + b_n\Big) \to 0. \tag{8}$$

Condition 4 ensures that the random variable $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \geq t)$ has a standard deviation negligible compared to its mean, and thus can concentrate around its mean $\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$. Condition 1 together with (7) in Condition 5 can guarantee that $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t) \approx \sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \geq t)$ in probability, via an application of Markov's inequality. Combining these two results we can prove that $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t) \approx \sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \geq t) \approx \sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$ and similarly $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \leq -t) \approx \sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \leq -t)$ uniformly over

$0 < t \le G^{-1}(\frac{c_1 q a_n}{p})$ with asymptotic probability one. In view of the definition of T in (4), assumption (8) ensures that the numerator in the ratio in (4) is mainly contributed by null features, which together with Conditions 1–4 proves that threshold T falls into the range $(0, G^{-1}(\frac{c_1 q a_n}{p})]$ with asymptotic probability one; See Lemma 4. Thus, $\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widehat{W}_j \ge T) \approx \sum_{j\in\mathcal{H}_0} \mathbb{1}(\widehat{W}_j \le -T)$ with asymptotic probability one by the symmetry of $\{\widetilde{W}_j\}_{j\in\mathcal{H}_0}$. Consequently, the FDR of the approximate knockoffs procedure is asymptotically the same as that of the perfect knockoffs procedure, where the latter has been proved to be controlled at the target level. This ensures that the FDR of the approximate knockoffs procedure can be controlled asymptotically, as formally stated in Theorem 1 below.

Condition 4 above can be easily satisfied if $\widetilde{W}_j$'s with $j \in \mathcal{H}_0$ are independent of each other. At the presence of dependence, it imposes an assumption on the strength of correlation among the indicator functions $\mathbb{1}(\widetilde{W}_j > t)$ with $j \in \mathcal{H}_0$. The ratio $\frac{G(t-b_n)-G(t+b_n)}{G(t)}$ in Condition 5 above is closely related to the hazard rate function in survival analysis if $G(t)$ has a probability density function. Loosely speaking, assumption (7) is satisfied for $b_n = o((\log p)^{1/\gamma})$ if the hazard rate function has enough smoothness and is more or less bounded uniformly over the range $t \in (0,\, G^{-1}(\frac{c_1 q a_n}{p})]$; it imposes an important condition on coupling accuracy b_n. Assumption (8) is satisfied if 1) only a fast vanishing fraction of $\widetilde{W}_j$'s for important features take negative values with nonvanishing probabilities, or 2) $\widetilde{W}_j$'s for important features all take positive values with high probability.

We are now ready to present our first general theorem on the FDR control for the approximate knockoffs inference procedure.

THEOREM 1. *Under Conditions 1–5, we have*

$$\limsup_{n\to\infty} \mathrm{FDR} \le q. \tag{9}$$

3. Layer 2 analysis: knockoff variables coupling.

3.1. *Characterization of approximate knockoff variables.* Section 2 establishes preliminary theoretical results on the asymptotic FDR control for the approximate knockoffs inference. The key assumption is Condition 1. Since the knockoff statistics are intermediate results calculated from the knockoff variables, it is important to provide a characterization on the quality of the approximate knockoff variable matrix $\widehat{\mathbf{X}}$ that can guarantee Condition 1. The assumption below is imposed for such a purpose.

CONDITION 6. *For $\widehat{\mathbf{X}}$ constructed from the approximate knockoffs procedure, there exists a perfect knockoff data matrix $\widetilde{\mathbf{X}}$ and an asymptotically vanishing sequence Δ_n such that*

$$\mathbb{P}\Big(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \ge \Delta_n\Big) \to 0, \tag{10}$$

where $\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} := \max_{1\le j\le p} n^{-1/2}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2$, and $\widehat{\mathbf{X}}_j$ and $\widetilde{\mathbf{X}}_j$ are the jth columns of the approximate and perfect knockoff variable matrices $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$, respectively.

Condition 6 above couples each approximate knockoff variable $\widehat{\mathbf{X}}_j$ with a perfect knockoff variable $\widetilde{\mathbf{X}}_j$. Similar to Condition 1, we need the realizations instead of the distributions of $\widehat{\mathbf{X}}_j$ and $\widetilde{\mathbf{X}}_j$ to be close, which is a major distinction from the assumption in Barber, Candès and Samworth (2020). Such distinction allows $\widehat{\mathbf{X}}$ to be constructed using sample $(\mathbf{X}, \mathbf{y})$ *without* data splitting under relaxed estimation accuracy assumptions, as will be illustrated in the next

two subsections. Later in Section 4, we will provide extensive analysis on the coupling order Δ_n using some specific examples of feature distributions.

We next show that the closeness between $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ can lead to the closeness between $\widehat{W}_j$'s and $\widetilde{W}_j$'s as required by Condition 1. Since different construction of the knockoff statistics depends on the feature matrix differently, we showcase the theory using two constructions of the knockoff statistics: the marginal correlation knockoff statistics and the regression coefficient difference (RCD) knockoff statistics.

For clarity, we include Table 1 to summarize the sets of assumptions on the model setting, feature distribution, the knockoff statistics, and the corresponding rates for the coupling accuracy in our layer 2 analysis.

TABLE 1
Summary of key conditions and results for asymptotic FDR control in Layer 2 analysis.

Model setting	Nonparametric model (14) in Section 3.2	Linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in Section 3.3
Feature distribution	$X \overset{d}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$	sub-Gaussian
Sparsity assumption	Condition 9 on $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}$	Sparse $\boldsymbol{\beta}$; sparse precision matrix for covariates; Condition 11
Knockoff statistics	Marginal correlation	RCD with debiased Lasso
$\widehat{W}_j$ coupling accuracy	Δ_n	$\Delta_n s\sqrt{(\log p)/n}$
Δ_n requirement for FDR control	$\Delta_n\sqrt{n}(\log p)^{1/2+1/\gamma} \to 0$	$\Delta_n s(\log p)^{1+1/\gamma} \to 0$

3.2. *Marginal correlation knockoff statistics.* Marginal correlation is a commonly analyzed measure on variable importance for feature screening due to its simplicity. Given $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ satisfying Condition 6, the approximate knockoff statistics based on the marginal correlation difference are defined as

$$\widehat{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widehat{\mathbf{X}}_j^T\mathbf{y}|) \text{ for } 1 \le j \le p, \tag{11}$$

and the coupled perfect knockoff statistics are given by

$$\widetilde{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widetilde{\mathbf{X}}_j^T\mathbf{y}|) \text{ for } 1 \le j \le p. \tag{12}$$

Observe that $\widetilde{W}_j - \widehat{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\widehat{\mathbf{X}}_j^T\mathbf{y}| - |\widetilde{\mathbf{X}}_j^T\mathbf{y}|)$ and thus under Condition 6, we have that with asymptotic probability one,

$$\max_{1\le j\le p} |\widehat{W}_j - \widetilde{W}_j| \le \Delta_n. \tag{13}$$

This result is summarized formally in Lemma 5 in Section A.2 of the Supplementary Material.

We consider the flexible nonparametric regression model

$$Y = f(X_{\mathcal{H}_1}) + \varepsilon, \tag{14}$$

where f is some unknown regression function, $X_{\mathcal{H}_1} = (X_j)_{j\in\mathcal{H}_1}$ contains all the relevant features for response Y, and ε is the model error satisfying $\varepsilon \perp\!\!\!\perp X$ and $\mathbb{E}(\varepsilon) = 0$. Assume that feature vector $X = (X_1, \cdots, X_p)^T \overset{d}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the positive definite covariance matrix. Moreover, let the distribution of the perfect knockoff variables $\widetilde{X} = (\widetilde{X}_1, \cdots, \widetilde{X}_p)^T$ satisfy that

$$(X, \widetilde{X}) = (X_1, \cdots, X_p, \widetilde{X}_1, \cdots, \widetilde{X}_p) \overset{d}{\sim} N\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - rI_p \\ \boldsymbol{\Sigma} - rI_p & \boldsymbol{\Sigma} \end{pmatrix}\right), \tag{15}$$

where $r > 0$ is a constant such that the above covariance matrix is positive definite. Here, we consider the equicorrelated construction (Candès et al., 2018) for simpler presentation and the diagonal matrix rI_p can be replaced with a general version $\mathrm{diag}(r_1, \cdots, r_p)$ with possibly distinct diagonal entries $\{r_j\}_{j=1}^p$. Note that the Gaussian distribution assumption is imposed mainly to verify the general Conditions 4 and 5. If one assumes directly these two conditions, the Gaussian distribution assumption can be removed.

Furthermore, we make the additional technical assumptions below on the generative model (14) to verify the conditions in our layer 1 analysis presented in Section 2.

CONDITION 7. *Y is a sub-Gaussian random variable with sub-Gaussian norm $\|Y\|_{\psi_2}$.*

CONDITION 8. *Define $\mathscr{A}_n = \{j \in \mathcal{H}_1 : (\mathbb{E}Y^2)^{-1/2}(|\mathbb{E}(X_jY)| - |\mathbb{E}(\widetilde{X}_jY)|) \geq 5\delta_n\}$ with*

$$\delta_n = C_{X,Y}\sqrt{n^{-1}\log p}, \tag{16}$$

where $C_{X,Y} := \max_{1\leq j\leq p}\left\{\frac{16\sqrt{2}\|X_j\|_{\psi_2}\|Y\|_{\psi_2}}{(\mathbb{E}Y^2)^{1/2}} \vee \frac{8\sqrt{2}|w_j|\|Y\|_{\psi_2}^2}{\mathbb{E}Y^2}\right\}$. It holds that $a_n := |\mathscr{A}_n| \to \infty$ and $C_{X,Y}$ is a positive constant that is independent of p and n.

Denote by $(\mathbf{\Sigma}^{-1})_j$ the jth column of matrix $\mathbf{\Sigma}^{-1}$, $\mathbf{\Sigma}_{i,j}$ the (i,j)th entry of matrix $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\mathcal{H}_1,j}$ a vector given by $(\mathbf{\Sigma}_{i,j})_{i\in\mathcal{H}_1}$. Recall the definition $G(t) = p_0^{-1}\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t) = p_0^{-1}\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \leq -t)$.

CONDITION 9. *For some sequence $m_n = o(a_n)$, matrices $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}$ are sparse in the sense that $\max_{1\leq j\leq p}\|(\mathbf{\Sigma}^{-1})_j\|_0 \leq m_n$ and $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\mathbf{\Sigma}_{\mathcal{H}_1,j} \neq \mathbf{0}) \leq m_n$. In addition, $C_1 < r < \min_{1\leq j\leq p}\mathbf{\Sigma}_{j,j} \leq \max_{1\leq j\leq p}\mathbf{\Sigma}_{j,j} < C_2$ for some constants $C_1 > 0$ and $C_2 > 0$.*

CONDITION 10. *It holds that $|\mathcal{H}_1|^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}(\widetilde{W}_j < -t) \leq G(t)$ for all $t \in (0, C_3\sqrt{\frac{\log p}{n}})$ with $C_3 > 0$ some large constant.*

Under Conditions 7–10, we can verify that Conditions 2–5 are satisfied. This together with Condition 6 and our general theorem on the FDR control (cf. Theorem 1) leads to the theorem below.

THEOREM 2. *Assume that Conditions 6–10 are satisfied. In addition, assume that for some constant $0 < \gamma < 1$, $(\log p)^{1/\gamma}m_n/a_n \to 0$ and the coupling accuracy Δ_n in Condition 6 satisfies $\sqrt{n}\Delta_n(\log p)^{1/2+1/\gamma} \to 0$. Then for the approximate knockoffs inference based on the marginal correlation, we have*

$$\limsup_{n\to\infty}\mathrm{FDR} \leq q.$$

Let us make a few remarks on the conditions and result presented in Theorem 2 above. Condition 8 verifies the signal strength assumption in Condition 3 in the specific context of model (14) and marginal correlation knockoff statistics. We show in Lemma 6 in Section A.2 of the Supplementary Material that Condition 2 holds with $\delta_n = O(\sqrt{n^{-1}\log p})$. Since we assume Gaussian feature distribution in this section, the dependence among the indicator functions as required by Condition 4 is determined by covariance matrix $\mathbf{\Sigma}$. Hence, Condition 9 is imposed to justify the validity of Condition 4. It is worth mentioning that

the sparse dependence structure assumed in Condition 9 can be replaced with a general assumption that the conditional distribution $X_{\mathcal{H}_0}|X_{\mathcal{H}_1}$ has sparse pairwise dependency and the sequence $\{h_j(t; X_{\mathcal{H}_1}) := \mathbb{E}(\mathbb{1}(\widetilde{W}_j \geq t)|X_{\mathcal{H}_1})\}_{j\in\mathcal{H}_0}$ has sparse pairwise correlation for each given $t > 0$. Condition 10 is a technical assumption that is intuitive and requires that on average, the probability of a relevant feature having a negative valued $\widetilde{W}_j$ is smaller than the corresponding probability of an irrelevant feature. Such condition is compatible with our requirement that relevant features should have positive and larger magnitude for $\widetilde{W}_j$.

Note that in this example, $w_j = \mathbb{E}\widetilde{W}_j$ and the concentration rate δ_n as in Condition 2 is $\delta_n \sim \sqrt{(\log p)/n}$. The assumption $\sqrt{n}\Delta_n(\log p)^{1/2+1/\gamma} \to 0$ in Theorem 2 requires that $\Delta_n \ll n^{-1/2}(\log p)^{-1/2-1/\gamma}$, and hence, $\Delta_n \ll \delta_n$. In view of (13), the requirement of $\Delta_n \ll \delta_n$ indeed restricts that the quality of $\widehat{W}_j$'s, as measured by Δ_n in the current example, is of an order smaller than δ_n. This also suggests that an independent sample of size $N \gg n$ may be needed to learn the covariate distribution for constructing the approximate knockoff variables in order to achieve the desired accuracy of $\max_{1\leq j\leq p}|\widehat{W}_j - \widetilde{W}_j| \leq \Delta_n \ll n^{-1/2}(\log p)^{-1/2-1/\gamma}$.

It is worth mentioning that the bound obtained in (13) may be improved under additional model assumptions. For instance, if additionally the covariates $\{X_j\}_{j=1}^p$ are independent, then under Condition 6 we can show that $\max_{1\leq j\leq p}|\widehat{W}_j - \widetilde{W}_j| \leq C\Delta_n\sqrt{n^{-1}\log p}$ (see Lemma 20 in Section B.16 of the Supplementary Material for details). The improved result is because of the elimination of spurious correlation between null and signal covariates. In this case, the condition on Δ_n in Theorem 2 is relaxed to $\Delta_n(\log p)^{1+1/\gamma} \to 0$.

The above discussions suggest that knockoff statistics based on marginal correlation are of low quality in the sense that they are less robust to estimation error and model misspecification. Indeed, we will see in the next section that some other popularly used knockoff statistics such as RCD can achieve asymptotic FDR control under much relaxed assumptions.

3.3. *Regression coefficient difference with debiased Lasso.* A popularly used construction of the knockoff statistics is RCD. We present our results under the following linear regression model for simplification; the extension to the generalized linear model (GLM) can be found in Section C of the Supplementary Material. We consider

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_j)_{1\leq j\leq p} \in \mathbb{R}^p$ is the true regression coefficient vector, $\boldsymbol{\varepsilon} \overset{d}{\sim} N(\mathbf{0}, \sigma^2 I_n)$ is the model error vector, and $\boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbf{X}$. Assume that feature vector $X = (X_1, \cdots, X_p)^T$ has mean $\mathbf{0}_p \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Denote by $\boldsymbol{\beta}^{\text{aug}} = (\boldsymbol{\beta}^T, \mathbf{0}_p^T)^T \in \mathbb{R}^{2p}$ the augmented true parameter vector.

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_j)_{1\leq j\leq 2p} \in \mathbb{R}^{2p}$ be the debiased Lasso estimator (Zhang and Zhang (2014)) based on the augmented design matrix $\widehat{\mathbf{X}}^{\text{aug}} := [\mathbf{X}, \widehat{\mathbf{X}}]$, where $\widehat{\mathbf{X}}$ is the approximate knockoff variable matrix. Assume that Condition 6 is satisfied and $\widetilde{\mathbf{X}}$ is the coupled perfect knockoffs variable matrix. Similarly, define $\widetilde{\mathbf{X}}^{\text{aug}} := [\mathbf{X}, \widetilde{\mathbf{X}}]$. Then $\widehat{\boldsymbol{\beta}}$ can be coupled with the debiased Lasso estimator denoted as $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_j)_{1\leq j\leq 2p} \in \mathbb{R}^{2p}$ based on $\widetilde{\mathbf{X}}^{\text{aug}}$. Then the RCD knockoff statistics can be defined as

$$\widehat{W}_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}| \tag{17}$$

$$\text{and} \quad \widetilde{W}_j = |\widetilde{\beta}_j| - |\widetilde{\beta}_{j+p}| \tag{18}$$

for the approximate and perfect knockoffs procedures, respectively, for $1 \leq j \leq p$.

We provide the explicit definition of the debiased Lasso estimator to assist future presentation. For $1 \le j \le 2p$, the debiased Lasso estimator is a one-step bias correction from some initial estimator $\widehat{\boldsymbol{\beta}}^{\text{init}} = (\widehat{\beta}_j^{\text{init}})_{1\le j\le 2p} \in \mathbb{R}^{2p}$ and is defined as

$$\widehat{\beta}_j = \widehat{\beta}_j^{\text{init}} + \frac{\widehat{\mathbf{z}}_j^T\big(\mathbf{y} - \widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}}\big)}{\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}}, \tag{19}$$

where $\widehat{\mathbf{z}}_j$ is the score vector defined as

$$\widehat{\mathbf{z}}_j = \widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}\widehat{\boldsymbol{\gamma}}_j \tag{20}$$

with $\widehat{\boldsymbol{\gamma}}_j := \arg\min_{\mathbf{b}}\big\{(2n)^{-1}\|\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\mathbf{b}\|_2^2 + \lambda_j\|\mathbf{b}\|_1\big\}$ and $\{\lambda_j\}_{j=1}^{2p}$ the nonnegative regularization parameters. We construct the initial estimator as

$$\widehat{\boldsymbol{\beta}}^{\text{init}} := \arg\min_{\mathbf{b}}\Big\{(2n)^{-1}\|\mathbf{y} - \widehat{\mathbf{X}}^{\text{aug}}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1\Big\} \tag{21}$$

with $\lambda = C\sqrt{n^{-1}\log(2p)}$ the regularization parameter and $C > 0$ some constant.

Analogously, the coupled debiased Lasso estimator $\widetilde{\boldsymbol{\beta}}$ can be defined componentwisely as

$$\widetilde{\beta}_j = \widetilde{\beta}_j^{\text{init}} + \frac{\widetilde{\mathbf{z}}_j^T\big(\mathbf{y} - \widetilde{\mathbf{X}}^{\text{aug}}\widetilde{\boldsymbol{\beta}}^{\text{init}}\big)}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}} \quad \text{for } 1 \le j \le 2p, \tag{22}$$

where

$$\widetilde{\boldsymbol{\beta}}^{\text{init}} = (\widetilde{\beta}_j^{\text{init}})_{1\le j\le 2p} := \arg\min_{\mathbf{b}}\Big\{(2n)^{-1}\|\mathbf{y} - \widetilde{\mathbf{X}}^{\text{aug}}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1\Big\} \tag{23}$$

and

$$\widetilde{\mathbf{z}}_j = \widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j \text{ with } \widetilde{\boldsymbol{\gamma}}_j := \arg\min_{\mathbf{b}}\Big\{(2n)^{-1}\|\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\mathbf{b}\|_2^2 + \lambda_j\|\mathbf{b}\|_1\Big\}. \tag{24}$$

It is important to emphasize that the *same* regularization parameters λ and λ_j's in defining $\widehat{\boldsymbol{\beta}}$ should be used as in defining $\widetilde{\boldsymbol{\beta}}$ in (22) so that their constructions differ only by the used feature matrix; this plays a key role in applying our coupling technique. Indeed, we prove in Lemma 11 in Section A.3 of the Supplementary Material that the coupling technique together with Condition 6 and some other regularity conditions ensures that with asymptotic probability one,

$$\max_{1\le j\le 2p}|\widetilde{\beta}_j - \widehat{\beta}_j| \lesssim \Delta_n s\sqrt{n^{-1}\log p}. \tag{25}$$

The above result guarantees that $\widehat{W}_j$'s and $\widetilde{W}_j$'s are also uniformly close over $1 \le j \le p$ with $\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j| \lesssim \Delta_n s\sqrt{n^{-1}\log p}$. As long as $s\Delta_n \to 0$, this upper bound has a smaller order than the concentration rate δ_n of $\widetilde{W}_j$ (cf. Condition 2), because here $\delta_n \sim \sqrt{n^{-1}\log p}$ as shown in our Lemma 12 in Section A.3. As commented after Theorem 2, the assumption that the coupling rate of $\max_{1\le j\le p}|\widetilde{W}_j - \widehat{W}_j|$ is of a smaller order than the concentration rate δ_n plays a key role in establishing our theory on the asymptotic FDR control.

We next introduce some additional notation and formally present the regularity conditions specific to this section. Observe that by symmetry, the augmented feature vector with the perfect knockoff variables has covariance matrix

$$\boldsymbol{\Sigma}^A = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \boldsymbol{D} \\ \boldsymbol{\Sigma} - \boldsymbol{D} & \boldsymbol{\Sigma} \end{pmatrix}, \tag{26}$$

where $\boldsymbol{D}$ is a diagonal matrix such that matrix $\boldsymbol{\Sigma}^A$ is positive definite. Let $\boldsymbol{\Omega}^A = (\boldsymbol{\Sigma}^A)^{-1}$ and $\boldsymbol{\gamma}_j = (\gamma_{j,l})_{l\neq j}$ with $\boldsymbol{\gamma}_{j,l} = -\boldsymbol{\Omega}^A_{j,l}/\boldsymbol{\Omega}^A_{j,j}$. It has been shown in Peng et al. (2009) that the residuals

$$e_j = \widetilde{X}^{\text{aug}}_j - \widetilde{X}^{\text{aug}}_{-j}\boldsymbol{\gamma}_j$$

satisfy that $\text{Cov}(e_j, \widetilde{X}^{\text{aug}}_{-j}) = \boldsymbol{0}$, $\text{Var}(e_j) = 1/\boldsymbol{\Omega}^A_{j,j}$, and $\text{Cov}(e_j, e_l) = \frac{\boldsymbol{\Omega}^A_{j,l}}{\boldsymbol{\Omega}^A_{j,j}\boldsymbol{\Omega}^A_{l,l}}$. For $1 \leq j \leq 2p$, denote by $\mathcal{S}_j = \text{supp}(\boldsymbol{\gamma}_j) \cup \text{supp}(\widetilde{\boldsymbol{\gamma}}_j) \cup \text{supp}(\widehat{\boldsymbol{\gamma}}_j)$. Let $J = \text{supp}(\boldsymbol{\beta}^{\text{aug}}) \cup \text{supp}(\widetilde{\boldsymbol{\beta}}^{\text{init}}) \cup \text{supp}(\widehat{\boldsymbol{\beta}}^{\text{init}})$ and $s := \|\boldsymbol{\beta}^{\text{aug}}\|_0 = \|\boldsymbol{\beta}\|_0 = o(n)$. We make the technical assumptions below.

CONDITION 11. *a) For some constant $C_4 > 0$, $\mathbb{P}(|J| \leq C_4 s) \to 1$.*
b) For some sequence $m_n \lesssim s$, it holds that $\max_{1\leq j\leq 2p} \|\boldsymbol{\Omega}^A_j\|_0 \leq m_n$ and $\mathbb{P}(\max_{1\leq j\leq 2p} |\mathcal{S}_j| \leq C_5 m_n) \to 1$ with some constant $C_5 > 0$.
c) $\max_{1\leq j\leq 2p} \|\boldsymbol{\gamma}_j\|_2 \leq C_6$ and $C_7 < \lambda_{\min}(\boldsymbol{\Omega}^A) \leq \lambda_{\max}(\boldsymbol{\Omega}^A) < C_8$ with some positive constants C_6, C_7, and C_8.

CONDITION 12 (Restrictive eigenvalues). *Assume that with probability $1 - o(1)$,*

$$\min_{\|\delta\|_0 \leq C_9 s} \frac{\delta^T [\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} \delta}{n\|\delta\|_2^2} \geq \kappa_1 \tag{27}$$

for some large enough constant $C_9 > 0$ and a constant $\kappa_1 > 0$.

CONDITION 13. *The features X_j's and errors e_j's are sub-Gaussian with sub-Gaussian norms $\|X_j\|_{\psi_2} \leq \phi$ and $\|e_j\|_{\psi_2} \leq \phi$ for some constant $\phi > 0$.*

CONDITION 14. *Let $\mathscr{A}_n = \{j \in \mathcal{H}_1 : |\beta_j| \gg \sqrt{n^{-1}\log p}\}$ and it holds that $a_n := |\mathscr{A}_n| \to \infty$.*

We are now ready to state our results on the FDR control for the approximate knockoffs inference based on the debiased Lasso coefficients.

THEOREM 3. *Assume that Conditions 6 and 10–14 hold, $m_n/a_n \to 0$, and $\frac{m_n^{1/2} s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} + \Delta_n s(\log p)^{1+1/\gamma} \to 0$ for some constant $0 < \gamma < 1$. Then we have*

$$\limsup_{n\to\infty} \text{FDR} \leq q.$$

Similarly as discussed in the last section, Condition 11 is used to verify the weak dependence assumption in Condition 4. Condition 6 and the two regularity Conditions 12–13 are imposed for verifying the coupling accuracy Condition 1. Condition 14 contributes to verifying the general signal strength requirement in Condition 3.

3.4. *Connection of Condition 6 with Wasserstein distance.* We detour slightly and discuss the connection of Condition 6 with a type of Wasserstein distance and state a conjecture of ours; it is safe to skip this section and proceed to Section 4 for knockoff variable coupling.

First recall that the knockoff variable matrix is generated in a rowwise fashion independent of each other. Given a row $\mathbf{x}$ of the original data matrix $\mathbf{X}$, denote by $\widehat{\mu}_{\mathbf{x}}$ the estimated or misspecified conditional distribution for generating the corresponding row in the approximate knockoff variable matrix $\widehat{\mathbf{X}}$, and denote by $\widetilde{\mu}_{\mathbf{x}}$ its oracle counterpart based on the true feature distribution. Conditional on the original data matrix $\mathbf{X}$, let $\widehat{\mu}^n = \widehat{\mu}_{\mathbf{x}_1} \times \widehat{\mu}_{\mathbf{x}_2} \times \cdots \times \widehat{\mu}_{\mathbf{x}_n}$ and

$\widetilde{\mu}^n = \widetilde{\mu}_{\mathbf{X}_1} \times \widetilde{\mu}_{\mathbf{X}_2} \times \cdots \times \widetilde{\mu}_{\mathbf{X}_n}$, where $\mathbf{x}_i$ is the ith row of the original data matrix $\mathbf{X}$. Define the conditional $(1,2)$-Wasserstein distance between $\widehat{\mu}^n$ and $\widetilde{\mu}^n$ as

$$\mathbb{W}_{1,2}(\widehat{\mu}^n, \widetilde{\mu}^n | \mathbf{X}) = \inf_{\eta \in \Gamma(\widehat{\mu}^n, \widetilde{\mu}^n)} \mathbb{E}_{(\mathrm{vec}(\widehat{\mathbf{X}}), \mathrm{vec}(\widetilde{\mathbf{X}})) \overset{d}{\sim} \eta} [\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} | \mathbf{X}], \tag{28}$$

where $\Gamma(\widehat{\mu}^n, \widetilde{\mu}^n)$ is the set consisting of all couplings of $\widehat{\mu}^n$ and $\widetilde{\mu}^n$, $\|\cdot\|_{1,2}$ is the matrix $(1,2)$-norm as defined in Condition 6, and $\mathrm{vec}(\widehat{\mathbf{X}})$ stands for vectorization of $\widehat{\mathbf{X}}$ by rows, similarly for $\mathrm{vec}(\widetilde{\mathbf{X}})$.

PROPOSITION 1. *Assume that there exists a deterministic sequence $c_n \to 0$ and a coupling $\eta^* \in \Gamma(\widehat{\mu}^n, \widetilde{\mu}^n)$ such that*

$$\mathbb{P}_{\boldsymbol{X}}(\mathbb{W}_{1,2}(\widehat{\mu}^n, \widetilde{\mu}^n | \mathbf{X}) \geq c_n) \to 0, \tag{29}$$

$$\mathbb{E}_{(\mathrm{vec}(\widehat{\mathbf{X}}), \mathrm{vec}(\widetilde{\mathbf{X}})) \overset{d}{\sim} \eta^*} [\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} | \mathbf{X}] \leq C_{\boldsymbol{X}} \mathbb{W}_{1,2}(\widehat{\mu}^n, \widetilde{\mu}^n | \mathbf{X}), \tag{30}$$

where $C_{\boldsymbol{X}} \geq 1$ depends only on $\mathbf{X}$ with well-defined expectation $\mathbb{E}_{\boldsymbol{X}}[C_{\boldsymbol{X}}] < \infty$, and $\mathbb{P}_{\boldsymbol{X}}$ and $\mathbb{E}_{\boldsymbol{X}}$ are probability and expectation taken with respect to $\mathbf{X}$, respectively. Then as $n \to \infty$, Condition 6 is satisfied with Δ_n chosen such that $\mathbb{E}_{\boldsymbol{X}}[C_{\boldsymbol{X}}] c_n \Delta_n^{-1} \to 0$.

It is seen that assumption (29) and the existence of η^* in Proposition 1 provide sufficient conditions ensuring Condition 6. We next verify the existence of η^* in a special scenario.

In Section 4.2, we present a concrete construction for coupling of the approximate and perfect knockoff variable matrices under Gaussian distribution, given by (37) and (38), respectively. The lemma below is based on such constructions. The proof of Lemma 1 is postponed to Section B.1 of the Supplementary Material.

LEMMA 1 (Gaussian Coupling). *Consider Gaussian knockoffs in Section 4.2. Let η^* be the conditional coupling measure used for generating (37) and (38). Define $\widehat{\mathbf{D}} := (2rI_p - r^2\widehat{\boldsymbol{\Omega}})^{1/2}$ and $\mathbf{D} := (2rI_p - r^2\boldsymbol{\Omega})^{1/2}$, where $\boldsymbol{\Omega}$, $\widehat{\boldsymbol{\Omega}}$, and r are the same as defined in Section 4.2. Let $\mathbf{D}_j$ and $\widehat{\mathbf{D}}_j$ be the jth columns of $\mathbf{D}$ and $\widehat{\mathbf{D}}$, respectively. If $\|\widehat{\mathbf{D}}_j\|_2 \|\mathbf{D}_j\|_2 - \widehat{\mathbf{D}}_j^T \mathbf{D}_j \leq C\|\widehat{\mathbf{D}}_j - \mathbf{D}_j\|_2^2$ for all $j = 1, \cdots, p$ with a constant $C \in (0, 1/2)$, then (30) is satisfied with $C_{\boldsymbol{X}} = \frac{2}{1-2C}\big(1 + \sqrt{2n^{-1}}\big)(r^2 \vee 1)$.*

The condition $\|\widehat{\mathbf{D}}_j\|_2 \|\mathbf{D}_j\|_2 - \widehat{\mathbf{D}}_j^T \mathbf{D}_j \leq C\|\widehat{\mathbf{D}}_j - \mathbf{D}_j\|_2^2$ can be satisfied if the covariates are close to independent, i.e., $\boldsymbol{\Omega}$ close to diagonal. In particular, when $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}$ are both diagonal, it holds that $\widehat{\mathbf{D}}_j^T \mathbf{D}_j - \|\widehat{\mathbf{D}}_j\|_2 \|\mathbf{D}_j\|_2 = 0 \leq \|\widehat{\mathbf{D}}_j - \mathbf{D}_j\|_2^2$. We conjecture that for more general $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}$, the coupling measure used for generating (37) and (38) could still satisfy (30). Proving the existence of η^* in the general scenario is highly challenging and left for future research.

4. Layer 3 analysis: construction of coupled knockoff variables. In this section, we present three specific constructions for the coupled perfect knockoff variables and verify that they satisfy Condition 6 with the desired convergence rate.

4.1. *Knockoffs for multivariate t-distribution.* In this example, we will construct knockoffs for multivariate t-distributed features by leveraging only information of the first two moments; the knowledge of the t-distribution will *not* be utilized in the approximate knockoffs construction. Assume that the underlying true feature distribution for $X = (X_1, \cdots, X_p)^T$ is

the multivariate centered t-distribution $t_\nu(\mathbf{0}, \mathbf{\Omega}^{-1})$ with unknown parameters ν and $\mathbf{\Omega}^{-1}$. We construct the approximate knockoff variables from the Gaussian distribution with the attempt to match the first two moments of feature vector X. It is seen that the working distribution $\widehat{F}$ is misspecified. It has been a common practice to use the multivariate Gaussian distribution to construct knockoff variables in practice; see, e.g., Candès et al. (2018); Bai et al. (2021).

Assume that there is an effective estimator $\widehat{\mathbf{\Theta}}$ for the precision matrix $\mathbf{\Theta} := [\mathrm{Cov}(X)]^{-1} = \frac{\nu-2}{\nu}\mathbf{\Omega}$ constructed using data matrix $\mathbf{X}$. We construct the approximate knockoffs variable matrix $\widehat{\mathbf{X}}$ from the misspecified Gaussian distribution as

$$\widehat{\mathbf{X}} = \mathbf{X}(I_p - r\widehat{\mathbf{\Theta}}) + \mathbf{Z}(2rI_p - r^2\widehat{\mathbf{\Theta}})^{1/2}, \tag{31}$$

where r is a constant such that $2rI_p - r^2\widehat{\mathbf{\Theta}}$ is positive definite, and $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is independent of $(\mathbf{X}, \mathbf{y})$ and consists of i.i.d. standard Gaussian entries.

Before suggesting our coupled perfect knockoff variables, it is necessary to review some properties of the multivariate t-distribution. Note that an alternative representation of the ith row of $\mathbf{X}$ is $\mathbf{x}_i = \frac{\eta_i}{\sqrt{Q_i/\nu}}$, where $\nu > 0$ is the degrees of freedom, $\eta_i \overset{d}{\sim} N(\mathbf{0}, \mathbf{\Omega}^{-1})$, $Q_i \overset{d}{\sim} \chi^2_\nu$, and $\eta_i \perp\!\!\!\perp Q_i$. Here, χ^2_ν is the chi-square distribution with ν degrees of freedom. When ν is large, the distribution of $\mathbf{x}_i$ is close to the Gaussian distribution $N(\mathbf{0}, (\frac{\nu-2}{\nu}\mathbf{\Omega})^{-1})$. Using this alternative representation, the design matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathrm{diag}(\frac{1}{\sqrt{\mathbf{Q}/\nu}})\boldsymbol{\eta}, \tag{32}$$

where $\boldsymbol{\eta}$ is the matrix with rows $\{\eta_i\}_{i=1}^n$, and $\mathrm{diag}(\frac{1}{\sqrt{\mathbf{Q}/\nu}}) = \mathrm{diag}(\frac{1}{\sqrt{Q_1/\nu}}, \cdots, \frac{1}{\sqrt{Q_n/\nu}})$.

We are ready to introduce our construction of the coupled perfect knockoff variable matrix

$$\widetilde{\mathbf{X}} = \mathbf{X}(I_p - r\mathbf{\Omega}) + \mathrm{diag}(\frac{1}{\sqrt{\mathbf{Q}/\nu}})\mathbf{Z}(2rI_p - r^2\mathbf{\Omega})^{1/2}, \tag{33}$$

where $\mathbf{Q}$, ν, and $\mathbf{\Omega}$ are identical to the ones in (32), and r and $\mathbf{Z}$ are identical to the ones in (31). Thus, $\mathbf{Z}$ is independent of $\mathbf{Q}$ and $\boldsymbol{\eta}$. In view of (32), we can see that

$$\begin{aligned}(\mathbf{X}, \widetilde{\mathbf{X}}) = &\mathrm{diag}(\frac{1}{\sqrt{\mathbf{Q}/\nu}})\big(\boldsymbol{\eta}, \boldsymbol{\eta}(I_p - r\mathbf{\Omega}) + \mathbf{Z}(2rI_p - r^2\mathbf{\Omega})^{1/2}\big)\\ :=&\mathrm{diag}(\frac{1}{\sqrt{\mathbf{Q}/\nu}})(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}}),\end{aligned}$$

where $(\boldsymbol{\eta}, \widetilde{\boldsymbol{\eta}})$ have i.i.d. rows that follow the Gaussian distribution $N(\mathbf{0}, \mathbf{\Sigma}^{\mathrm{aug}})$ with

$$\mathbf{\Sigma}^{\mathrm{aug}} = \begin{pmatrix} \mathbf{\Omega}^{-1} & \mathbf{\Omega}^{-1} - rI_p \\ \mathbf{\Omega}^{-1} - rI_p & \mathbf{\Omega}^{-1} \end{pmatrix}. \tag{34}$$

Thus, this verifies that $\widetilde{\mathbf{X}}$ forms a perfect knockoff variable matrix for $\mathbf{X}$.

The proposition below verifies that the coupling assumption in Condition 6 holds.

PROPOSITION 2. *Assume that $C_l \le \|\mathbf{\Omega}^{-1}\|_2 \le C_u$ and $\|(2rI_p - r^2\mathbf{\Omega})^{-1}\|_2 \le C_u$ for some constants $C_u > 0$ and $C_l > 0$. Assume further that $\mathbf{\Omega}$ and $\widehat{\mathbf{\Theta}}$ are both sparse in the sense that $\max_{1\le j\le p}(\|\mathbf{\Omega}_j\|_0 + \|\widehat{\mathbf{\Theta}}_j\|_0) \le \rho_n$ almost surely with $\rho_n(n^{-1}\log p)^{1/2} \to 0$ and $\rho_n\nu^{-1/2} \to 0$, and that there exists a constant $C > 0$ such that*

$$\mathbb{P}\big(\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 \ge C\rho_n(n^{-1}\log p)^{1/2}\big) \to 0. \tag{35}$$

Then as $\nu \ge 9$ and $\log p = o(n^{1-4/\nu})$, we have that for some constant $C > 0$,

$$\mathbb{P}\Big(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \le C\big(\rho_n(n^{-1}\log p)^{1/2} + \nu^{-1/2}\big)\Big) \to 1. \tag{36}$$

The assumed convergence rate of $\rho_n(n^{-1}\log p)^{1/2}$ for precision matrix estimation in (35) has been verified in many existing works (e.g., Cai, Liu and Luo (2011), Fan, Liao and Liu (2016), and Fan and Lv (2016)) under the sparsity assumption. Proposition 2 above indicates that the knockoffs procedure can potentially achieve the asymptotic FDR control even when the working distribution is misspecified but with the first two moments matched.

We next compare our results to those in Barber, Candès and Samworth (2020). For simplicity, let us further assume that $\mathbf{\Omega} = I_p$ and is known. Then $X \overset{d}{\sim} t_\nu(\mathbf{0}, I_p)$ and the constructed approximate knockoff variables $\widehat{X} \overset{d}{\sim} N(\mathbf{0}, \frac{\nu}{\nu-2} I_p)$. We set $r = 1$ in (31) and (33) when constructing the approximate and perfect knockoff matrices, and hence the augmented covariance matrix in (34) is given by $\mathbf{\Sigma}^{\text{aug}} = I_{2p}$. In such case, Proposition 2 guarantees that

$$\mathbb{P}\left(\max_{1\le j\le p} n^{-1/2} \|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2 \le C\nu^{-1/2} \right) \to 1.$$

This implies that Condition 6 is satisfied with $\Delta_n = C\nu^{-1/2}$. Observe that $X_j = \frac{Z_j}{\sqrt{\mathcal{X}_\nu^2/\nu}}$ with $Z_j \overset{d}{\sim} N(0,1)$ and the denominator satisfies that for an absolute constant $C > 0$ and $\nu \gg \log(np)$,

$$\mathbb{P}\left(|\mathcal{X}_\nu^2/\nu - 1| \ge C\sqrt{\nu^{-1}\log(np)} \right) = O((np)^{-C^2/8}).$$

These indicate that the multivariate t-distribution is asymptotically close to the standard Gaussian distribution when $\nu \gg \log(np)$. Thus, under Conditions 10–12 and 14 for the setting of the linear model, if we construct the knockoff statistics as RCD based on the debiased Lasso, we can prove similarly as Theorem 3 that

$$\limsup_{n\to\infty} \text{FDR} \le q,$$

when $\nu^{1/2} \gg s(\log p)^{1+1/\gamma}$ and $\frac{s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$ for some $0 < \gamma < 1$.

Barber, Candès and Samworth (2020) also derived an upper bound on the FDR inflation. Directly applying their result and calculating the KL divergence in their upper bound under the specific model setting stated above, we can obtain the lemma below.

LEMMA 2. *By applying Theorem 1 in Barber, Candès and Samworth (2020), it requires at least* $\nu^2 \gg n\min(n,p)$ *for* $\limsup_{n\to\infty} \text{FDR} \le q$.

The intuition behind Lemma 2 above is that Theorem 1 in Barber, Candès and Samworth (2020) requires the empirical KL divergence $\max_{j\in\mathcal{H}_0} \widehat{KL}_j$ converging to zero in probability, where

$$\widehat{KL}_j = \sum_{i=1}^n \left[\frac{\mathbf{X}_{i,j}^2(\nu-2)}{2\nu} - \frac{\nu+p}{2}\log\left(1 + \frac{\mathbf{X}_{i,j}^2}{\nu + \|\mathbf{X}_{i,-j}\|_2^2}\right) \right.$$
$$\left. - \left(\frac{\widehat{\mathbf{X}}_{i,j}^2(\nu-2)}{2\nu} - \frac{\nu+p}{2}\log\left(1 + \frac{\widehat{\mathbf{X}}_{i,j}^2}{\nu + \|\mathbf{X}_{i,-j}\|_2^2}\right)\right)\right].$$

Here, $\mathbf{X} = (\mathbf{X}_{i,j}) \in \mathbb{R}^{n\times p}$ consists of i.i.d. rows sampled from $t_\nu(\mathbf{0}, I_p)$, while $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_{i,j}) \in \mathbb{R}^{n\times p}$ consists of i.i.d. rows sampled from $N(\mathbf{0}, I_p)$. As shown in the proof of Lemma 2 in Section B.2 of the Supplementary Material, $\widehat{KL}_j$ is a sum of i.i.d. random variables with positive mean of order $\frac{Cp}{\nu(\nu+p)}$. Hence, $\widehat{KL}_j$ is concentrated at $\frac{Cnp}{\nu(\nu+p)}$ and to ensure that

TABLE 2
Summary of key conditions and results in Layer 3 analysis

Covariate distribution	$t_\nu(\mathbf{0}, \mathbf{\Omega^{-1}})$	$N(\mathbf{0}, \mathbf{\Omega^{-1}})$	Nonparanormal
Source of error in constructing $\widehat{\mathbf{X}}$	Misspecified distribution and estimated $\mathbf{\Omega}$	Estimated $\mathbf{\Omega}$	Estimated $\mathbf{\Omega}$
Verified coupling rate Δ_n	$\rho_n(n^{-1}\log p)^{1/2} + \nu^{-1/2}$	$\rho_n\sqrt{\frac{\log p}{n}}$	$\rho_n\sqrt{\frac{\log p}{n}} + \sqrt{\frac{p\rho_n(\log n)^3}{n}}$
$p \gg n$?	Yes	Yes	No
Marginal correlation knockoff statistics	Out-sample estimation needed for general $\mathbf{\Omega}$; In-sample estimation allowed for diagonal $\mathbf{\Omega}$		
RCD with debiased Lasso	In-sample estimation allowed for general $\mathbf{\Omega}$ with sparsity		

$\widehat{KL}_j \xrightarrow{d} 0$, we need at least $\frac{np}{\nu(\nu+p)} \to 0$, or equivalently, $\nu^2 \gg n\min(n,p)$. Such condition is stronger than our requirement $\nu^{1/2} \gg s(\log p)^{1+1/\gamma}$ derived from the coupling technique when $s = o(\sqrt{n})$ and $p \geq n$.

4.2. *Gaussian knockoffs.* We now study the commonly used example of Gaussian knockoffs with the correctly specified distribution family. Assume that feature vector $X = (X_1, \cdots, X_p)^T \overset{d}{\sim} N(\mathbf{0}, \mathbf{\Omega}^{-1})$ with unknown precision matrix $\mathbf{\Omega}$, and we have an effective estimate $\widehat{\mathbf{\Omega}}$ that may be constructed using in-sample observations. A popularly used approximate knockoff variable matrix is

$$\widehat{\mathbf{X}} = \mathbf{X}(I_p - r\widehat{\mathbf{\Omega}}) + \mathbf{Z}(2rI_p - r^2\widehat{\mathbf{\Omega}})^{1/2}, \tag{37}$$

where $r > 0$ is some constant such that $2rI_p - r^2\widehat{\mathbf{\Omega}}$ is positive definite, and $\mathbf{Z} = (Z_{i,j}) \in \mathbb{R}^{n\times p}$ is independent of $(\mathbf{X}, \mathbf{y})$ with i.i.d. entries $Z_{i,j} \overset{d}{\sim} N(0,1)$. Note that the approximate knockoff variable matrix in (37) uses the correctly specified distribution family for $\mathbf{X}$ (i.e., the Gaussian distribution).

We couple $\widehat{\mathbf{X}}$ with the perfect knockoff variable matrix

$$\widetilde{\mathbf{X}} = \mathbf{X}(I_p - r\mathbf{\Omega}) + \mathbf{Z}(2rI_p - r^2\mathbf{\Omega})^{1/2}, \tag{38}$$

where importantly, $\mathbf{Z}$ and r are identical to those used in constructing $\widehat{\mathbf{X}}$. We present the result below regarding the accuracy of the approximate knockoff variables.

PROPOSITION 3. *Assume that $C_l \leq \|\mathbf{\Omega}^{-1}\|_2 \leq C_u$ and $\|(2rI_p - r^2\mathbf{\Omega})^{-1}\|_2 \leq C_u$ for some constants $C_u > 0$ and $C_l > 0$. Assume further that precision matrix $\mathbf{\Omega}$ and its estimator $\widehat{\mathbf{\Omega}}$ are both sparse in the sense that $\max_{1\leq j\leq p}\|(\mathbf{\Omega}_j\|_0 + \|\widehat{\mathbf{\Omega}}_j\|_0) \leq \rho_n$ almost surely with $\rho_n(n^{-1}\log p)^{1/2} \to 0$, and that there exists a constant $C > 0$ such that*

$$\mathbb{P}\big(\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \geq C\rho_n(n^{-1}\log p)^{1/2}\big) \to 0. \tag{39}$$

Then we have that for some constant $C > 0$,

$$\mathbb{P}\big(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \leq C\rho_n(n^{-1}\log p)^{1/2}\big) \to 1. \tag{40}$$

Proposition 3 above implies that Condition 6 is satisfied with coupling accuracy $\Delta_n = C\rho_n(n^{-1}\log p)^{1/2}$, where ρ_n represents the sparsity level of $\mathbf{\Omega}$ and its estimator. We discuss the implication on FDR control utilizing the previously studied two knockoff statistics, namely the marginal correlation and RCD statistics, by applying Theorems 2–3, and then compare with the relevant results in Barber, Candès and Samworth (2020).

First consider the linear model and the RCD knockoff statistics based on the debiased Lasso. It follows from Theorem 3 that under Conditions 10–12 and 14, we have

$\limsup_{n\to\infty} \mathrm{FDR} \le q$ provided that $s\rho_n(\log p)^{3/2+1/\gamma} = o(\sqrt{n})$ for some $0 < \gamma < 1$. Our technical analyses do *not* require data splitting or an independent pretraining sample. In comparison, the results in Barber, Candès and Samworth (2020) require an independent unlabeled pretraining data set with sample size N to estimate the unknown precision matrix. Specific to the model setting considered in this section, their results indicate that $\limsup_{n\to\infty} \mathrm{FDR} \le q$ when $N \gg n\rho_n(\log p)^2$. This again shows the advantage of our coupling technique in the robustness analyses.

Next we move to the marginal correlation statistics. In view of (40), (13), and Theorem 2, it is seen that in-sample estimation generally cannot meet the required condition of $\sqrt{n}\Delta_n(\log p)^{1/2+1/\gamma} \to 0$ in Theorem 2, and hence there is no guarantee of asymptotic FDR control even using our coupling idea. This message is consistent with Niu et al. (2024), where the model-X framework for conditional independence test is investigated; see Section 6 for more detailed discussion. In the special case of independent features as discussed at the end of Section 3.2, in-sample estimation can achieve asymptotic FDR control if $(\log p)^{1+1/\gamma} \max_j |\widehat{\sigma}_j^{-2} - \sigma_j^{-2}| = o_p(1)$, where $\widehat{\sigma}_j^2$ and σ_j^2 are estimated and true variance for the jth feature, respectively.

We next compare with the relevant results in Barber, Candès and Samworth (2020) for marginal correlation statistics. It is discussed in their Section 3.2.1 that the KL divergence in their FDR inflation upper bound can be replaced with some E_j defined on summary statistics, such as

$$E_j = E_j(\mathbf{X}_j^T\mathbf{y}, \widehat{\mathbf{X}}_j^T\mathbf{y}), \text{ with } E_j(a,b) = \log\Big(\frac{\mathbb{P}((\mathbf{X}_j^T\mathbf{y}, \widehat{\mathbf{X}}_j^T\mathbf{y}) = (a,b)|\mathbf{X}_{-j}, \widehat{\mathbf{X}}_{-j}, \mathbf{y})}{\mathbb{P}((\mathbf{X}_j^T\mathbf{y}, \widehat{\mathbf{X}}_j^T\mathbf{y}) = (b,a)|\mathbf{X}_{-j}, \widehat{\mathbf{X}}_{-j}, \mathbf{y})}\Big), \tag{41}$$

and that their FDR inflation upper bound remains to hold. An independent pretraining sample is required for generating their $\widehat{\mathbf{X}}_j$'s. Note that E_j above depends on the "closeness" of $\widehat{\mathbf{X}}_j^T\mathbf{y}$ to $\mathbf{X}_j^T\mathbf{y}$. For the FDR inflation in their upper bound to asymptotically vanish, it is required that $\max_j |E_j| = o_p(1)$. It is unclear how $\max_j |E_j| = o_p(1)$ can be translated into the explicit bound on the estimation accuracy of $\mathbf{\Omega}$ when the covariate dependence is most general. In the simpler case of independent covariates, the condition reduces to $\max_j |E_j| = O_p((\log p)\max_{1\le j\le p} |\widehat{\sigma}_j^{-2} - \sigma_j^{-2}|)$. Comparing to our condition of $(\log p)^{1+1/\gamma} \max_j |\widehat{\sigma}_j^{-2} - \sigma_j^{-2}| = o_p(1)$ discussed above, the additional term of $(\log p)^{1/\gamma}$ is the price we pay for in-sample estimation.

4.3. *Nonparanormal knockoffs.* We further investigate a much more general distribution family, that is, the Gaussian copula distributions. Assume that $X = (X_1, \cdots, X_p)^T$ has marginal distributions $X_j \overset{d}{\sim} F_j(\cdot)$ and satisfies that $(\Phi^{-1}(F_1(X_1)), \cdots, \Phi^{-1}(F_p(X_p)))^T \overset{d}{\sim} N(\mathbf{0}, \mathbf{\Omega}^{-1})$, where the diagonal entries of $\mathbf{\Omega}^{-1}$ are all one. Further assume that we have effective estimators $\widehat{F}_j$ for F_j and $\widehat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$. Define $\widehat{\mathbf{V}} = (\widehat{\mathbf{V}}_{i,j}) \in \mathbb{R}^{n\times p}$ with $\widehat{\mathbf{V}}_{i,j} = \Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))$ and $\widetilde{\mathbf{V}} = (\widetilde{\mathbf{V}}_{i,j}) \in \mathbb{R}^{n\times p}$ with $\widetilde{\mathbf{V}}_{i,j} = \Phi^{-1}(F_j(\mathbf{X}_{i,j}))$. Let $\widehat{\mathbf{U}} = (\widehat{\mathbf{U}}_{i,j}) \in \mathbb{R}^{n\times p}$ be given by

$$\widehat{\mathbf{U}} = \widehat{\mathbf{V}}(I_p - r\widehat{\mathbf{\Omega}}) + \mathbf{Z}(2rI_p - r^2\widehat{\mathbf{\Omega}})^{1/2}, \tag{42}$$

where $r > 0$ is some constant such that $2rI_p - r^2\widehat{\mathbf{\Omega}}$ is positive definite, and $\mathbf{Z} = (\mathbf{Z}_{i,j}) \in \mathbb{R}^{n\times p}$ is independent of $(\mathbf{X}, \mathbf{y})$ with i.i.d. entries $Z_{i,j} \overset{d}{\sim} N(0,1)$. We construct the approximate knockoff variable matrix as $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_{i,j}) \in \mathbb{R}^{n\times p}$ with

$$\widehat{\mathbf{X}}_{i,j} = \widehat{F}_j^{-1}(\Phi(\widehat{\mathbf{U}}_{i,j})). \tag{43}$$

It is seen that this example also uses the correctly specified distribution family for X, i.e., the Gaussian copula.

We suggest to construct the coupled perfect knockoff variable matrix as $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_{ij})$ with

$$\widetilde{\mathbf{X}}_{i,j} = F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})), \tag{44}$$

where $\widetilde{\mathbf{U}}_{i,j}$ represents the (i,j)th entry of matrix

$$\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}}(I_p - r\boldsymbol{\Omega}) + \mathbf{Z}(2rI_p - r^2\boldsymbol{\Omega})^{1/2} \tag{45}$$

with $\mathbf{Z}$ and r identical in values to the ones used in (42). The proposition below characterizes the coupling rate between $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$.

PROPOSITION 4. *Assume that (39) is satisfied and both $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}$ are sparse in the sense that $\max_{1\le j\le p}(\|\boldsymbol{\Omega}_j\|_0 + \|\widehat{\boldsymbol{\Omega}}_j\|_0) \le \rho_n$ with $p\rho_n = o(n/(\log n)^3)$ almost surely. Assume further that for $1 \le j \le p$, the distribution estimators satisfy $\frac{1}{2n} \le \widehat{F}_j(x) \le 1 - \frac{1}{2n}$ for each $x \in \operatorname{supp}(X_j)$, $\operatorname{supp}(X_j) \subset [-b, b]$ for some constant $b > 0$, and there exists a constant $M > 0$ such that*

$$\mathbb{P}\Big(\max_{1\le j\le p} \sup_{x\in[2Mn^{-1}\log n, 1-2Mn^{-1}\log n]} \big|\widehat{F}_j^{-1}(x) - F_j^{-1}(x)\big| \ge (Mn^{-1}\log n)^{1/2}\Big) \to 0, \tag{46}$$

$$\mathbb{P}\Big(\max_{1\le j\le p} \sup_{x\in(F_j^{-1}(2Mn^{-1}\log n), F_j^{-1}(1-2Mn^{-1}\log n))} \frac{|\widehat{F}_j(x) - F_j(x)|}{F_j(x)[1-F_j(x)]} \ge (Mn^{-1}\log n)^{1/2}\Big) \to 0, \tag{47}$$

$$\mathbb{P}\Big(\max_{1\le j\le p} \sup_{x,y\in(0,1)} \frac{\big|\widehat{F}_j^{-1}(x) - \widehat{F}_j^{-1}(y)\big|}{|x-y| + (n^{-1}(\log n)|x-y|)^{1/2} + n^{-1}\log n} \ge M\Big) \to 0. \tag{48}$$

Then we have

$$\mathbb{P}\Big(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \le C\Big(\rho_n\sqrt{\frac{\log p}{n}} + \sqrt{\frac{p\rho_n(\log n)^3}{n}}\Big)\Big) \to 1. \tag{49}$$

REMARK 1. *When estimators $\{\widehat{F}_j\}_{j=1}^p$ are the empirical distribution functions and $p = o(n)$, it can be shown that (46), (47), and (48) can be satisfied when the density function f_{X_j} is uniformly bounded on the support.*

See, e.g., Liu, Lafferty and Wasserman (2009); Liu et al. (2012) for the estimation of nonparanormal distributions, and we opt not to discuss it here due to the space constraint. We also remark that the bounded support assumption of $\operatorname{supp}(X_j) \subset [-b, b]$ is to simplify the technical proofs and may be removed by applying the truncation technique and letting b slowly diverge with n. Since such technical relaxation is not the main focus of the current paper, we choose not to explore it here.

5. Robust knockoffs for k-FWER control. Model-X knockoffs framework has also been explored for the purpose of k-FWER control (Lehmann and Romano, 2005), where the goal is to control

$$k\text{-FWER} = \mathbb{P}(|\widehat{S} \cap \mathcal{H}_0| \ge k) \tag{50}$$

below a prespecified target level $q \in (0,1)$. Given the approximate knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$, the set of selected features is $\widehat{S} = \{1 \le j \le p : \widehat{W}_j \ge T_v\}$, where the threshold is defined as

$$T_v = \sup\Big\{t \in \widehat{\mathcal{W}} : \#\{j : -\widehat{W}_j \ge t\} = v\Big\} \tag{51}$$

with v the largest integer such that

$$\sum_{i=k}^{\infty} 2^{-(i+v)} \binom{i+v-1}{i} \le q. \tag{52}$$

When the true feature distribution is known, Janson and Su (2016) showed that the perfect knockoffs inference procedure provides precise finite-sample control on the k-FWER. We now establish the companion theory for the approximate knockoffs inference procedure.

Denote by $\widehat{V} = |\widehat{S} \cap \mathcal{H}_0|$ the number of false discoveries. Similar to the FDR analysis, we assume that the number of relevant features $|\mathcal{H}_1| \to \infty$ as $n \to \infty$. Further, we consider the scenario where k diverges very slowly with n. Our layer 1 theory will again build on the key Condition 1 that there exist coupled perfect knockoff statistics that are sufficiently close to the approximate knockoff statistics. However, different from the FDR study where Conditions 2–5 are needed, we assume instead the two technical conditions below and their interpretations are similar to Conditions 4–5. Recall the definition that $G(t) = p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \ge t)$ and $p_0 = |\mathcal{H}_0|$.

CONDITION 15 (Weak dependence). *For constants $0 < \gamma < 1$ and $C > 0$, and a positive sequence $m_n = o(k)$, it holds that*

$$\mathrm{Var}\Big(\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widetilde{W}_j > t)\Big) \le C m_n p_0 G(t) + o\big((\log k)^{-1/\gamma} [p_0 G(t)]^2\big) \tag{53}$$

uniformly over $t \in (G^{-1}(\frac{3k}{2p}), G^{-1}(\frac{k}{2p}))$.

CONDITION 16. *Assume that $G(t)$ is a continuous function. It holds that as $n \to \infty$,*

$$\sup_{t \in \left(G^{-1}(\frac{3k}{2p}), G^{-1}(\frac{k}{2p})\right)} \frac{G(t-b_n) - G(t+b_n)}{G(t)} \to 0 \tag{54}$$

and

$$k^{-1} \sum_{j \in \mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{3k}{2p})\Big) \to 0 \tag{55}$$

Now we are ready to present our general theorem on the k-FWER control for the approximate knockoffs procedure.

THEOREM 4. *Assume that Conditions 1, 15, and 16 are satisfied, $k \to \infty$, and $m_n/k \to 0$ as $n \to \infty$. Then for each $\varepsilon > 0$, we have*

$$\limsup_{n \to \infty} \mathbb{P}(\widehat{V} \ge k(1+\varepsilon)) \le q. \tag{56}$$

The main idea for proving Theorem 4 is to compare the approximate knockoff statistics $\{\widehat{W}_j\}_{j=1}^p$ with their coupled perfect counterparts $\{\widetilde{W}_j\}_{j=1}^p$ and show that the approximate

threshold T_v satisfies $|T_v - \widetilde{T}_v| \le b_n$ as long as $\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j| \le b_n$, where $\widetilde{T}_v$ is the corresponding threshold from the perfect knockoff statistics. Moreover, we can show that for each $\varepsilon > 0$, with high probability it holds that $\widetilde{T}_{v+M_v+1} < \widetilde{T}_v - 2b_n \le \widetilde{T}_{v+M_v}$ for some integer $M_v \le k\varepsilon$. Therefore, the probability of the approximate knockoffs inference procedure making at least k false discoveries can be related to that of the k-FWER control with the perfect knockoff statistics, which establishes the desired result in Theorem 4.

Similar to the layer 2 FDR analysis in Section 3, we showcase the general theory using two constructions of the knockoff statistics: the marginal correlation and the RCD knockoff statistics, under the coupling accuracy assumption in Condition 6.

With the marginal correlation knockoff statistics, under the same model setting of Section 3.2, the following result on the k-FWER control can be established.

THEOREM 5. *Assume the same model setting* (14) *as in Section 3.2 and the marginal correlation knockoff statistics* (11). *Further, assume that Conditions 6, 9, and 10 are satisfied,* $k \to \infty$, $m_n/k \to 0$, *and* $\Delta_n\sqrt{n\log p} \to 0$. *Then for each* $\varepsilon > 0$, *we have*

$$\limsup_{n\to\infty} \mathbb{P}(\widehat{V} \ge k(1+\varepsilon)) \le q.$$

Analogously, with the RCD knockoff statistics, under the same setting of Section 3.3, we have the parallel theorem for the k-FWER control below.

THEOREM 6. *Assume the same linear model setting as in Section 3.3 and the RCD knockoff statistics* (17). *Further, assume that Conditions 6, 10, and 11–13 are satisfied,* $k \to \infty$, $m_n/k \to 0$, *and* $\frac{m_n^{1/2}s(\log p)^{3/2}(\log k)^{1/\gamma}}{\sqrt{n}} + \Delta_n s\log p \to 0$ *for some constant* $0 < \gamma < 1$. *Then for each* $\varepsilon > 0$, *we have*

$$\limsup_{n\to\infty} \mathbb{P}(\widehat{V} \ge k(1+\varepsilon)) \le q. \tag{57}$$

6. Connection with literature. We now provide more detailed comparison with three additional existing works Fan et al. (2020a); Fan et al. (2020b); Niu et al. (2024).

Fan et al. (2020b) investigated the power and robustness of knockoffs inference in the linear model setting where the features follow a latent factor model with parametric idiosyncratic noise. In-sample estimation is allowed for constructing their approximate knockoff variables. Condition 4 therein for robustness analysis is essentially a preliminary form of our knockoff variable coupling condition under their parametric model assumption, and Condition 6 therein is loosely comparable to the proved results in our Theorem 1; these two conditions are model specific and directly assumed therein without theoretical justification. Fan et al. (2020a) provided theoretical guarantee for the asymptotic FDR control for the approximate knockoffs procedure under an assumption that the FDR function is Lipschitz with respect to feature covariance matrix when the feature distribution is jointly Gaussian. In their paper, the feature distribution and model sparsity are learned by balanced sample splitting, and the dependence of response Y on covariates in X can be nonlinear and arbitrary. Their Lipschitz assumption on FDR function is comparable to the proved results in our Theorem 1.

Niu et al. (2024) studied the robustness of the conditional randomization test (CRT) and demonstrated that, when the feature distribution is learned in sample, type-I error control cannot be attained for arbitrary test statistics. In their Section 3, a test statistic that is closely related to the marginal correlation test was investigated and it was shown that its type-I error can be arbitrarily inflated when in-sample feature distribution is learned. This message is similar to ours in the sense that marginal correlation statistics have low accuracy (see Section 3.2). Niu et al. (2024) also established an interesting double-robustness phenomenon: errors

in fitting the distribution of the features can be compensated for by using a test statistic that more accurately captures the distribution of the response given the features. Since for FDR or k-FWER control, there is only one source of error caused by estimated/misspecified covariate distribution, the double-robustness may not be a relevant property in our study.

Comparing to these existing works, a major innovation of our paper is the introduction of a new closeness measure for evaluating the qualities of the approximate knockoff variables and knockoff statistics. This new measure is closely related to the $(1, 2)$-Wasserstein distance. The coupling idea for knockoffs robustness analysis and the $(1, 2)$-Wasserstein distance are both new to the literature; they equip us with a much more powerful tool for better understanding the practical robustness of the model-X framework. Indeed, as revealed in our analysis, the robustness of model-X procedure goes beyond the scenarios already revealed in the literature. The connection to the $(1, 2)$-Wasserstein distance also suggests that the robustness of model-X can be a general phenomenon beyond the covariate distribution examples provided in our current paper.

There exist some other less related works in the literature that contribute to relaxing the assumption of fully known feature distribution in the model-X knockoffs framework. For instance, Huang and Janson (2020) relaxed such assumption via assuming the existence of sufficient statistic for the model and proposing an alternative conditional exchangeability for knockoffs given the sufficient statistic.

7. Simulation studies. In this section, we examine the finite-sample performance of the approximate knockoffs inference using the approximate or misspecified feature distribution through some simulation examples.

7.1. *Approximate feature distribution.* Our first simulation example considers Gaussian feature vector $X \overset{d}{\sim} N(\mathbf{0}, \mathbf{\Omega}^{-1})$, where the precision matrix $\mathbf{\Omega} = (\omega_{ij}) \in \mathbb{R}^{p\times p}$ is unknown and sparse with entries $\omega_{ij} = 0.2^{|i-j|}$ for $|i - j| < 10$ and $\omega_{ij} = 0$ for $|i - j| \geq 10$. We apply the James–Stein-type shrinkage estimator for the covariance matrix (as in the R Package 'knockoff') and examine the FDR control of the approximate knockoffs inference procedure with estimated covariance matrix. In-sample estimation is used for learning the feature covariance matrix. We consider two settings: the linear regression model and logistic regression model.

SETTING 1. *Assume that $Y = X\boldsymbol{\beta} + \varepsilon$, where ε is a random error with $\varepsilon \overset{d}{\sim} N(0, 1)$. Let the coefficient $\boldsymbol{\beta} \in \mathbb{R}^p$ be sparse with 50 nonzero components, where the nonzero locations are randomly selected and each nonzero coefficient is randomly generated from $\{\pm 3\}$.*

SETTING 2. *Assume that the response Y depends on X through a logistic regression model. Let the regression coefficient $\boldsymbol{\beta} \in \mathbb{R}^p$ be sparse with 30 nonzero components, where the nonzero locations are randomly selected and each nonzero coefficient is randomly generated from $\{\pm 3\}$.*

We consider the construction of knockoff statistics using the debiased Lasso regression coefficient difference. We set $p = 400$ and $n \in \{150, 250, 350, 500\}$. From the numerical results in Table 3, it is seen that for a few settings of sample size, the FDR is marginally inflated above the target level $q = 0.2$ due to the estimated feature distribution and Monte Carlo error. Overall, the approximate knockoffs inference procedure demonstrates robust FDR control across various values of sample size n, which verifies our theoretical analysis that the knockoff statistics based on the debiased Lasso regression coefficient difference can guarantee the asymptotic FDR control with in-sample learned feature distribution.

TABLE 3
FDR control for the approximate knockoffs procedure using estimated feature distribution under Settings 1 and 2, with a targeted FDR level $q = 0.2$. Results are based on 100 replications.

Setting 1					Setting 2				
n	150	250	350	500	n	150	250	350	500
FDR	0.186	0.211	0.203	0.189	FDR	0.142	0.205	0.207	0.205

7.2. *Misspecified feature distribution.* In the second simulation example, we consider a feature vector $X \in \mathbb{R}^p$ generated from a multivariate t-distribution $t_\nu(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$. To examine the effect of misspecified feature distribution, we generate knockoff variables using the misspecified Gaussian distribution $N(\mathbf{0}, \frac{\nu}{\nu-2}\mathbf{\Sigma})$ with matched first two moments and explore the FDR control of approximate knockoffs inference procedure as the number of degrees of freedom ν changes. We fix the sample size as $n = 300$ and the dimensionality as $p = 400$. Again the linear model in Setting 1 and logistic model in Setting 2 are considered. We investigate the FDR control using knockoff statistics constructed from the debiased Lasso coefficient difference.

The number of degrees of freedom ν determines the closeness between the approximate and coupled perfect knockoff procedures, as demonstrated in layer 3 analysis in Section 4.1. We examine the behavior of the approximate knockoffs procedure for $\nu \in \{5, 10, 20, 50\}$. It is observed from Table 4 that the approximate knockoffs procedure can have slightly inflated FDR for a small value of $\nu = 5$, while achieving desired FDR control almost always for larger values of $\nu = 10$, 20, and 50. This again verifies our theoretical analysis.

TABLE 4
FDR control for the approximate knockoffs procedure using misspecified feature distribution under Settings 1 and 2, with a targeted FDR level $q = 0.2$. Results are based on 100 replications.

Setting 1					Setting 2				
ν	5	10	20	50	ν	5	10	20	50
FDR	0.238	0.190	0.206	0.195	FDR	0.175	0.162	0.186	0.169

8. Discussions. We have investigated in this paper the robustness of the model-X knockoffs framework introduced in Candès et al. (2018) by characterizing the feature selection performance of the approximate knockoffs (ARK) procedure, a popularly implemented version of the model-X knockoffs framework in practice. The approximate knockoffs procedure differs from the model-X knockoffs procedure in that it uses the misspecified or estimated feature distribution to generate the knockoff variables *without* the use of sample splitting. We have proved formally that the approximate knockoffs procedure can achieve the asymptotic FDR and k-FWER control as the sample size diverges in the high-dimensional setting. A key idea empowering our technical analysis is coupling, where we pair statistics in the approximate knockoffs procedure with those in the model-X knockoffs procedure so that they are close in realizations with high probability. The knockoff variable coupling has been investigated under some specific distribution assumptions in the current work. An interesting future study is to investigate the coupling idea under a broader class of or even general feature distributions.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor, and the Editor for their constructive comments that improved the quality of this paper.

Funding. YF was supported in part by NIH Grant 1R01GM131407 and NSF grant 2310981. JL was supported in part by NSF Grants EF-2125142 and DMS-2324490.

SUPPLEMENTARY MATERIAL

Supplement to "ARK: Robust Knockoffs Inference with Coupling"

The Supplementary Material Fan, Gao and Lv (2024) contains all the proofs and technical details, and an extension of the analysis in Section 3.3 to the setting of the generalized linear model.

REFERENCES

BAI, X., REN, J., FAN, Y. and SUN, F. (2021). KIMI: knockoff Inference for Motif Identification from molecular sequences with controlled false discovery rate. *Bioinformatics* **37** 759–766.

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43** 2055–2085.

BARBER, R. F. and CANDÈS, E. J. (2019). A knockoff filter for high-dimensional selective inference. *Ann. Statist.* **47** 2504–2537. https://doi.org/10.1214/18-AOS1755 MR3988764

BARBER, R. F., CANDÈS, E. J. and SAMWORTH, R. J. (2020). Robust inference with knockoffs. *Ann. Statist.* **48** 1409–1431. https://doi.org/10.1214/19-AOS1852 MR4124328

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57** 289–300.

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. https://doi.org/10.1198/jasa.2011.tm10155 MR2847973

CAI, T. T. and LIU, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association* **111** 229–240.

CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B* **80** 551–577. https://doi.org/10.1111/rssb.12265 MR3798878

CAO, Y., SUN, X. and YAO, Y. (2021). Controlling the False Discovery Rate in Transformational Sparsity: split Knockoffs. *arXiv preprint arXiv:2103.16159v3*.

CHI, C. M., FAN, Y., ING, C. K. and LV, J. (2023). High-Dimensional Knockoffs Inference for Time Series Data. *arXiv preprint arXiv:2112.09851*.

DAI, X., LYU, X. and LI, L. (2022). Kernel knockoffs selection for nonparametric additive models. *Journal of the American Statistical Association* 1–13.

DAI, C., LIN, B., XING, X. and LIU, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association* 1–38.

DING, P. (2016). On the conditional distribution of the multivariate t distribution. *The American Statistician* **70** 293–295.

FAN, Y., GAO, L. and LV, J. (2024). Supplement to "ARK: Robust Knockoffs Inference with Coupling".

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19** C1–C32.

FAN, Y. and LV, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics* **44** 2098–212.

FAN, Y., DEMIRKAYA, E., LI, G. and LV, J. (2020a). RANK: large-scale inference with graphical nonlinear knockoffs. *J. Amer. Statist. Assoc.* **115** 362–379. https://doi.org/10.1080/01621459.2018.1546589 MR4078469

FAN, Y., LV, J., SHARIFVAGHEFI, M. and UEMATSU, Y. (2020b). IPAD: stable interpretable forecasting with knockoffs inference. *J. Amer. Statist. Assoc.* **115** 1822–1834. https://doi.org/10.1080/01621459.2019.1654878 MR4189760

GIVENS, C. R. and SHORTT, R. M. (1984). A class of Wasserstein metrics for probability distributions. *Michigan Mathematical Journal* **31** 231–240.

GUO, X., REN, H., ZOU, C. and LI, R. (2022). Threshold selection in feature screening for error rate control. *Journal of the American Statistical Association* 1–13.

HUANG, D. and JANSON, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics* **48** 3021–3042.

JANSON, L. and SU, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975. https://doi.org/10.1214/16-EJS1129 MR3486422

JORDON, J., YOON, J. and VAN DER SCHAAR, M. (2018). KnockoffGAN: generating knockoffs for feature selection using generative adversarial networks. In *International Conference on Learning Representations*.

LEHMANN, E. L. and ROMANO, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. https://doi.org/10.1214/009053605000000084 MR2195631

LI, J. and MAATHUIS, M. H. (2021). GGM knockoff filter: false discovery rate control for Gaussian graphical models. *Journal of the Royal Statistical Society Series B* **83** 534–558.

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. https://doi.org/10.1214/13-AOS1169 MR3161453

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10**.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. https://doi.org/10.1214/12-AOS1037 MR3059084

LU, Y., FAN, Y., LV, J. and NOBLE, W. S. (2018). DeepPINK: reproducible feature selection in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS 2018)* **31**.

NIU, Z., CHAKRABORTY*, A., DUKES, O. and KATSEVICH, E. (2024). Reconciling model-X and doubly robust approaches to conditional independence testing. *Annals of Statistics* **to appear**.

PENG, J., WANG, P., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104** 735–746.

REN, Z. and BARBER, R. F. (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

REN, Z., WEI, Y. and CANDÈS, E. (2021). Derandomizing knockoffs. *Journal of the American Statistical Association* 1–11.

ROMANO, Y., SESIA, M. and CANDÈS, E. (2020). Deep knockoffs. *Journal of the American Statistical Association* **115** 1861–1872.

SAULIS, L. (1992). Probabilities of large deviations for random vectors. *Theory of Probability & Its Applications* **36** 494–507.

SESIA, M., SABATTI, C. and CANDÈS, E. J. (2019). Gene hunting with hidden Markov model knockoffs. *Biometrika* **106** 1–18. https://doi.org/10.1093/biomet/asy033 MR3912377

SPECTOR, A. and JANSON, L. (2022). Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics* **50** 252–276.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. https://doi.org/10.1214/14-AOS1221 MR3224285

VERSHYNIN, R. (2018). *High-Dimensional Probability*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108231596 MR3837109

WANG, W. and JANSON, L. (2022). A high-dimensional power analysis of the conditional randomization test and knockoffs. *Biometrika* **109** 631–645.

WEINSTEIN, A., SU, W. J., BOGDAN, M., BARBER, R. F. and CANDÈS, E. J. (2020). A Power Analysis for Model-X Knockoffs with ℓ_p-Regularized Statistics. *arXiv preprint arXiv:2007.15346v2*.

ZHANG, H. and CHEN, S. X. (2021). Concentration inequalities for statistical inference. *Commun. Math. Res.* **37** 1–85. https://doi.org/10.4208/cmr.2020-0041 MR4220305

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B* **76** 217–242. https://doi.org/10.1111/rssb.12026 MR3153940

ZHU, Z., FAN, Y., KONG, Y., LV, J. and SUN, F. (2021). DeepLINK: deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences of the United States of America* **118** e2104683118.

Supplement to "ARK: Robust Knockoffs Inference with Coupling"

Yingying Fan, Lan Gao and Jinchi Lv

This Supplementary Material contains the proofs of Theorems 1–6, Propositions 1–4, and some key technical lemmas. All the notation is the same as defined in the main body of the paper. Section A presents the Proofs of Theorems 1–6 and Propositions 1–4. We provide the proofs of the key lemmas and additional technical details in Section B. In Section C, we extend the analysis in Section 3.3 for knockoff statistics constructed with the regression coefficient difference to the setting of the generalized linear model (GLM). Throughout the Supplement, C stands for some positive constant whose value may change from line to line.

APPENDIX A: PROOFS OF THEOREMS 1–6 AND PROPOSITIONS 1–4

A.1. Proof of Theorem 1. It has been shown in Candès et al. (2018) that the model-X knockoffs inference procedure achieves the exact FDR control when the perfect knockoff statistics are employed. Note that the approximate knockoff statistics $\{\widehat{W}_j\}$ are expected to provide a reliable approximation to the perfect knockoff statistics $\{\widetilde{W}_j\}$, as assumed in Condition 1. The main idea of the proof is to establish the FDR control for the approximate knockoffs inference procedure through a comparison of the approximate knockoff statistics and a certain realization of the perfect knockoff statistics. The two lemmas below provide a sketch of the proof and can be established under Conditions 1–5.

LEMMA 3. *Assume that Conditions 1, 4, and 5 are satisfied. When $a_n \to \infty$ and $m_n/a_n \to 0$, we have that for some constant $0 < c_1 < 1$,*

$$\sup_{t \in \left(0, G^{-1}(\frac{c_1 q a_n}{p})\right]} \left| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \geq t)}{\sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \geq t)} - 1 \right| = o_p(1), \tag{A.1}$$

$$\sup_{t \in \left(0, G^{-1}(\frac{c_1 q a_n}{p})\right]} \left| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \leq -t)}{\sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \leq -t)} - 1 \right| = o_p(1). \tag{A.2}$$

LEMMA 4. *Under Conditions 1–5, we have that for some constant $0 < c_1 < 1$, $\mathbb{P}(T \leq G^{-1}(\frac{c_1 q a_n}{p})) \to 1$.*

We present the proofs of Lemmas 3 and 4 in Sections B.3 and B.4, respectively. Now we are ready to prove Theorem 1. Let us define two events $\mathcal{B}_1 = \{T \leq G^{-1}(\frac{c_1 q a_n}{p})\}$ and

$$\mathcal{B}_{2,\epsilon} = \Big\{ \sup_{t \in (0, G^{-1}(\frac{c_1 q a_n}{p})]} \Big(\Big| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \geq t)}{\sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \geq t)} - 1 \Big| \vee \Big| \frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \leq -t)}{\sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \leq -t)} - 1 \Big| \Big) \leq \epsilon \Big\}$$

for $\epsilon > 0$. Lemmas 3 and 4 above have shown that $\mathbb{P}(\mathcal{B}_1^c) \to 0$ and $\mathbb{P}(\mathcal{B}_{2,\epsilon}^c) \to 0$ for each $\epsilon > 0$. In addition, it holds naturally that $0 \leq \mathrm{FDP} \leq 1$. Then it follows that

$$\begin{aligned} \mathrm{FDR} &\leq \mathbb{E}\left(\frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \geq T)}{1 \vee \sum_{j=1}^p \mathbb{1}(\widehat{W}_j \geq T)} \cdot \mathbb{1}(\mathcal{B}_1) \mathbb{1}(\mathcal{B}_{2,\epsilon}) \right) + \mathbb{P}(\mathcal{B}_1^c) + \mathbb{P}(\mathcal{B}_{2,\epsilon}^c) \\ &= \mathbb{E}\left(\frac{\sum_{j \in \mathcal{H}_0} \mathbb{1}(\widehat{W}_j \geq T)}{1 \vee \sum_{j=1}^p \mathbb{1}(\widehat{W}_j \geq T)} \cdot \mathbb{1}(\mathcal{B}_1) \mathbb{1}(\mathcal{B}_{2,\epsilon}) \right) + o(1). \end{aligned} \tag{A.3}$$

In view of the definition of threshold T in (4), we can deduce that

$$\begin{aligned}
&\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq T)}{1\vee\sum_{j=1}^{p}\mathbb{1}(\widehat{W}_j\geq T)}\cdot\mathbb{1}(\mathcal{B}_1)\mathbb{1}(\mathcal{B}_{2,\epsilon})\\
&=\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq T)}{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\leq -T)}\cdot\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\leq -T)}{1\vee\sum_{j=1}^{p}\mathbb{1}(\widehat{W}_j\geq T)}\cdot\mathbb{1}(\mathcal{B}_1)\mathbb{1}(\mathcal{B}_{2,\epsilon})\\
&\leq q\cdot\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq T)}{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\leq -T)}\cdot\mathbb{1}(\mathcal{B}_1)\mathbb{1}(\mathcal{B}_{2,\epsilon}).
\end{aligned}\tag{A.4}$$

Furthermore, it is easy to see that on event $\mathcal{B}_1\cap\mathcal{B}_{2,\epsilon}$, we have

$$\begin{aligned}
\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq T)}{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\leq -T)}&\leq\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]}\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq t)}{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\leq -t)}\\
&\leq\frac{1+\epsilon}{1-\epsilon}\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]}\frac{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j\geq t)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j\leq -t)}\\
&=\frac{1+\epsilon}{1-\epsilon},
\end{aligned}$$

where the last equation above is obtained by the symmetry of the perfect knockoff statistics $\{\widetilde{W}_j\}_{j\in\mathcal{H}_0}$ that $\mathbb{P}(\widetilde{W}_j\geq t)=\mathbb{P}(\widetilde{W}_j\leq -t)$. Therefore, we can obtain that for any $\epsilon>0$,

$$\mathrm{FDR}\leq q\cdot\frac{1+\epsilon}{1-\epsilon}+o(1),\tag{A.5}$$

which yields the desired result (9). This completes the proof of Theorem 1.

A.2. Proof of Theorem 2. The main idea of the proof is to directly apply Theorem 1 by verifying Conditions 1–5 involved. We will show in the lemmas below that Conditions 1–5 are satisfied for the marginal correlation knockoff statistics under Conditions 6–10 and the setting of nonparametric regression model (14) with normal features. Proofs of Lemmas 5–8 are presented in Sections B.5–B.8.

LEMMA 5. *Assume that Condition 6 is satisfied. Then we have that*

$$\mathbb{P}\Big(\max_{1\leq j\leq p}|\widehat{W}_j-\widetilde{W}_j|\geq\Delta_n\Big)\to 0.\tag{A.6}$$

Lemma 5 above shows that Condition 1 is satisfied with sequences $b_n:=\Delta_n$. Define $w_j=(\mathbb{E}Y^2)^{-1/2}(|\mathbb{E}(X_jY)|-|\mathbb{E}(\widetilde{X}_jY)|)$ for $1\leq j\leq p$. Note that $w_j=0$ for $j\in\mathcal{H}_0$ since $(X_j,X_{\mathcal{H}_1})\overset{d}{=}(\widetilde{X}_j,X_{\mathcal{H}_1})$ for $j\in\mathcal{H}_0$ by the exchangeability between X_j and $\widetilde{X}_j$. Recall from the definition in (16) that

$$\delta_n=\sqrt{\frac{\log p}{n}}\max_{1\leq j\leq p}\Big\{\frac{16\sqrt{2}\|X_j\|_{\psi_2}\|Y\|_{\psi_2}}{(\mathbb{E}Y^2)^{1/2}}\vee\frac{8\sqrt{2}|w_j|\|Y\|_{\psi_2}^2}{\mathbb{E}Y^2}\Big\}.$$

We have the concentration inequality below for $\widetilde{W}_j$ under the sub-Gaussian assumption in Condition 7.

LEMMA 6. *Assume that Condition 7 is satisfied. When* $\log p = o(n)$*, we have that*

$$\sum_{j=1}^{p} \mathbb{P}(|\widetilde{W}_j - w_j| \geq \delta_n) \leq 6p^{-1} + p\exp\Big\{-\frac{n(\mathbb{E}Y^2)^2}{8\mathbb{E}Y^4}\Big\}. \tag{A.7}$$

Lemma 6 above indicates that Condition 2 related to the concentration rate of $\widetilde{W}_j$ is satisfied with δ_n defined in (16) and that $\Delta_n \leq \delta_n$, where Δ_n is the approximation accuracy of the approximate knockoff statistics obtained in Lemma 5. In addition, from the definition of w_j, under Condition 8 we have that the general Condition 3 on the signal strength is also satisfied. Next we will turn to the verification of Conditions 4–5.

LEMMA 7. *Assume that Condition 9 is satisfied. Then we have that for each* $t \geq 0$*,*

$$\frac{\mathrm{Var}\left(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j \geq t)\right)}{p_0 G(t)} \leq 2m_n. \tag{A.8}$$

LEMMA 8. *Assume that Conditions 9 and 10 are satisfied. Then when* $(\log p)^{1/\gamma} m_n/a_n \to 0$ *and* $\sqrt{n}\Delta_n(\log p)^{1/2+1/\gamma} \to 0$ *for some constant* $0 < \gamma < 1$*, we have that*

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p}))]} \frac{G(t-\Delta_n) - G(t+\Delta_n)}{G(t)} \to 0 \tag{A.9}$$

and

$$a_n^{-1} \sum_{j\in\mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + \Delta_n\Big) \to 0 \tag{A.10}$$

as $n \to \infty$*.*

Lemma 7 above shows that Condition 4 is satisfied, while Lemma 8 above implies that Condition 5 is satisfied. Finally, the conclusion of Theorem 2 can be obtained by directly applying the general Theorem 1. This completes the proof of Theorem 2.

A.3. Proof of Theorem 3. The main idea of the proof is to directly apply Theorem 1 by verifying Conditions 1–5 for the knockoff statistics constructed from the debiased Lasso coefficients. A key observation is that the debiased Lasso coefficients are asymptotically normal. Denote by

$$\tau_j = \|\widetilde{\mathbf{z}}_j\|_2 / |\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\mathrm{aug}}|.$$

The debiased Lasso coefficient can be written as

$$\sqrt{n}(\widetilde{\beta}_j - \beta_j^{\mathrm{aug}}) = \frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2} \cdot \sqrt{n}\tau_j + \sum_{k\neq j} \frac{\sqrt{n}\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\mathrm{aug}} (\beta_k^{\mathrm{aug}} - \widetilde{\beta}_k^{\mathrm{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\mathrm{aug}}}. \tag{A.11}$$

Observe that $\frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2} \sim N(0, \sigma^2)$, $\sqrt{n}\tau_j = O_p(1)$, and the remainder term in (A.11) above is of order $o_p(1)$. Thus, the debiased Lasso estimator is asymptotically normal in the sense that

$$\tau_j^{-1}(\widetilde{\beta}_j - \beta_j^{\mathrm{aug}}) \xrightarrow{d} N(0, \sigma^2).$$

Our proof will build mainly on such intuition. Throughout the proof below, constant C may take different values from line to line.

We first present two lemmas below about the consistency of Lasso estimators $\widetilde{\boldsymbol{\beta}}^{\mathrm{init}}$ and $\widetilde{\boldsymbol{\gamma}}_j$. We omit the proofs of Lemmas 9 and 10 here to avoid redundancy since they are well-known results for the consistency of Lasso estimators in the literature.

LEMMA 9. *Under Conditions 11–13, we have that with probability* $1 - o(p^{-3})$,

$$\|\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}\|_1 \leq Cs\sqrt{\frac{\log p}{n}}, \tag{A.12}$$

$$\|\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}\|_2 \leq C\sqrt{\frac{s\log p}{n}}, \tag{A.13}$$

$$\|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}})\|_2 \leq C\sqrt{s\log p}. \tag{A.14}$$

LEMMA 10. *Under Conditions 11–13, we have that with probability* $1 - o(p^{-3})$,

$$\max_{1\leq j\leq 2p} \|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_1 \leq Cm_n\sqrt{\frac{\log p}{n}}, \tag{A.15}$$

$$\max_{1\leq j\leq 2p} \|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_2 \leq C\sqrt{\frac{m_n\log p}{n}}, \tag{A.16}$$

$$\max_{1\leq j\leq 2p} \|\widetilde{\mathbf{X}}^{\text{aug}}_{-j}(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)\|_2 \leq C\sqrt{m_n\log p}. \tag{A.17}$$

In addition, when $\frac{m_n\log p}{n} \to 0$ *we have that with probability* $1 - o(p^{-3})$,

$$|\sqrt{n}\tau_j - (\mathbb{E}e_j^2)^{-1/2}| \leq C\sqrt{\frac{m_n\log p}{n}}, \tag{A.18}$$

$$\left|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_l - \text{Cov}(e_j, e_l)\right| \leq C\sqrt{\frac{m_n\log p}{n}}. \tag{A.19}$$

The four lemmas below outline the proof for verifying the general Conditions 1–5. Proofs of Lemma 11–14 are provided in Sections B.9–B.12, respectively.

LEMMA 11. *Assume that Conditions 6 and 11–13 are satisfied. Then as* $\Delta_n s^{1/2} \to 0$ *and* $\sqrt{\frac{s\log p}{n}} \to 0$, *we have that*

$$\mathbb{P}\left(\max_{1\leq j\leq 2p} |\widetilde{\beta}_j - \widehat{\beta}_j| \geq C\Delta_n s\sqrt{\frac{\log p}{n}}\right) \to 0. \tag{A.20}$$

Lemma 11 above indicates that Condition 1 is satisfied with sequences $b_n := C\Delta_n s\sqrt{\frac{\log p}{n}}$. Let us define $w_j = |\beta_j|$.

LEMMA 12. *Assume that Conditions 11–13 are satisfied. Then as* $s\sqrt{\frac{m_n\log p}{n}} \to 0$, *we have that for some* $C > 0$, $\sum_{j=1}^p \mathbb{P}(|\widetilde{W}_j - w_j| \geq C\sqrt{n^{-1}\log p}) \to 0$.

Lemma 12 above shows that Condition 2 related to the concentration rate of $\widetilde{W}_j$ is satisfied with $\delta_n = C\sqrt{n^{-1}\log p}$. In addition, it holds that $b_n \ll C\sqrt{n^{-1}\log p}$ due to the assumption $\Delta_n s \to 0$ in Theorem 3. In addition, in light of the definition of w_j, under Condition 14 we have that the general Condition 3 on the signal strength is also satisfied. We next turn to the verification of Conditions 4–5.

LEMMA 13. *Assume that Conditions 11–13 are satisfied. Then as* $\frac{m_n^{1/2} s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$, *we have that* $\mathrm{Var}\left(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j > t)\right) \le V_1(t) + V_2(t)$, *where for some* $0<\gamma<1$ *and* $0<c_1<1$,

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p}))} \frac{V_1(t)}{[p_0 G(t)]^2} \to 0 \tag{A.21}$$

and

$$\sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p}))} \frac{V_2(t)}{p_0 G(t)} \lesssim m_n. \tag{A.22}$$

LEMMA 14. *Assume that Conditions 6, 10, and 11 –13 are satisfied. Then when* $\frac{m_n^{1/2} s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$ *and* $\Delta_n s(\log p)^{1+1/\gamma} \to 0$, *we have that*

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p}))} \frac{G(t-b_n) - G(t+b_n)}{G(t)} \to 0 \tag{A.23}$$

and

$$a_n^{-1} \sum_{j\in\mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + b_n\Big) \to 0 \tag{A.24}$$

as $n\to\infty$.

Lemma 13 above shows that Condition 4 is satisfied, whereas Lemma 14 implies that Condition 5 is satisfied. Finally, the conclusion of Theorem 3 can be derived by directly applying the general Theorem 1. This completes the proof of Theorem 3.

A.4. Proof of Theorem 4. We first define the corresponding threshold $\widetilde{T}_v$ for the perfect knockoff statistics $\{\widetilde{W}_j\}_{j=1}^p$ in the model-X knockoffs inference for the k-FWER control as

$$\widetilde{T}_v = \sup\{t\in\widetilde{\mathcal{W}} : \#\{j : -\widetilde{W}_j \ge t\} = v\},$$

where v is defined as in (52) and $\widetilde{\mathcal{W}} = \{|\widetilde{W}_1|, \cdots, |\widetilde{W}_p|\}$. As sketched in Lemmas 15–17 below, the main idea of the proof is to show that the threshold T_v based on the approximate knockoff statistics and the threshold $\widetilde{T}_v$ based on the perfect knockoff statistics are sufficiently close under Condition 1 such that for any $\varepsilon>0$, the number of $\widetilde{W}_j$'s that lie between T_v and $\widetilde{T}_v$ is at most $v\varepsilon$ with asymptotic probability one, where v satisfies $v/k\to 1$ as $k\to\infty$. Specifically, let M_v be the integer such that

$$\widetilde{T}_{v+M_v} \ge \widetilde{T}_v - 2b_n > \widetilde{T}_{v+M_v+1}. \tag{A.25}$$

Then we can establish a bound for M_v as shown in Lemma 17 below. We first present the three lemmas below that provide an outline of the proof. The proofs of Lemmas 15–17 are provided in Sections B.13–B.15, respectively.

LEMMA 15. *Under Condition 1, we have that*

$$\mathbb{P}(|T_v - \widetilde{T}_v| \ge b_n) \to 0. \tag{A.26}$$

LEMMA 16. *Assume that* $k\to\infty$. *Then we have that*

$$\frac{v}{k} = 1 + O(k^{-1/2}). \tag{A.27}$$

LEMMA 17. *Under all the conditions of Theorem 4, we have that for each $\varepsilon > 0$,*

$$\mathbb{P}(M_v \le v\varepsilon) \to 1. \tag{A.28}$$

We are now ready to prove Theorem 4. It follows straightforwardly from Lemma 15 that

$$\begin{aligned}
\mathbb{P}(\widehat{V} \ge k(1+2\varepsilon)) &= \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widehat{W}_j \ge T_v) \ge k(1+2\varepsilon)\Big) \\
&\le \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j \ge \widetilde{T}_v - 2b_n) \ge k(1+2\varepsilon)\Big) \\
&\le \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j \ge \widetilde{T}_{v+M_v}) \ge k(1+2\varepsilon)\Big) \\
&= \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(-\widetilde{W}_j \ge \widetilde{T}_{v+M_v}) \ge k(1+2\varepsilon)\Big) \\
&\le \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(-\widetilde{W}_j \ge \widetilde{T}_v) \ge k(1+2\varepsilon) - M_v\Big),
\end{aligned}$$

where the second last step above is because of the symmetry of $\widetilde{W}_j$'s with $j \in \mathcal{H}_0$ and the last step above is due to

$$\sum_{j\in\mathcal{H}_0} \mathbb{1}(-\widetilde{W}_j \ge \widetilde{T}_{v+M_v}) - \sum_{j\in\mathcal{H}_0} \mathbb{1}(-\widetilde{W}_j \ge \widetilde{T}_v) \le M_v$$

by the definitions of $\widetilde{T}_v$ and M_v.

Moreover, Lemma 17 above shows that $M_v \le v\varepsilon$ with asymptotic probability one and Lemma 16 above proves that $v/k = 1 + o(1)$. Then it holds that $2k\varepsilon > M_v$ with asymptotic probability one. Hence, combining the above results and by the union bound, we can deduce that

$$\mathbb{P}(\widehat{V} \ge k(1+2\varepsilon)) \le \mathbb{P}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(-\widetilde{W}_j \ge \widetilde{T}_v) \ge k\Big) + o(1) = q + o(1).$$

Consequently, it follows that for each $\varepsilon > 0$,

$$\limsup_{n\to\infty} \mathbb{P}(\widehat{V} \ge k(1+2\varepsilon)) \le q.$$

This concludes the proof of Theorem 4.

A.5. Proof of Theorem 5. The proof of Theorem 5 is analogous to that of Theorem 2 in Section A.2. We omit the detailed proof here to avoid redundancy.

A.6. Proof of Theorem 6. The proof of Theorem 6 is similar to that of Theorem 3 in Section A.3. Hence we omit the detailed proof here to avoid redundancy.

A.7. Proof of Proposition 1. Let $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ be matrices generated from the conditional coupling measure η^* given $\mathbf{X}$. By Chebyshev's inequality, we have

$$\begin{aligned}
\mathbb{P}(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \ge \Delta_n) &= \mathbb{E}_{\boldsymbol{X}}[\mathbb{P}^*(\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2} \ge \Delta_n|\mathbf{X})] \le \mathbb{E}_{\boldsymbol{X}}\Big[\frac{\mathbb{E}^*[\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2}|\mathbf{X}]}{\Delta_n}\Big] \\
&\le \mathbb{E}_{\boldsymbol{X}}[C_{\boldsymbol{X}}]c_n\Delta_n^{-1} \to 0.
\end{aligned}$$

A.8. Proof of Proposition 2. From the definitions in (31) and (33), we see that

$$\widehat{\mathbf{X}} - \widetilde{\mathbf{X}} = r\mathbf{X}\mathbf{A} + \mathbf{Z}\mathbf{B} + \mathrm{diag}(1 - \frac{1}{\sqrt{Q_1/\nu}}, \cdots, 1 - \frac{1}{\sqrt{Q_n/\nu}})\mathbf{Z}\mathbf{C}, \tag{A.29}$$

where $\mathbf{A} = \mathbf{\Omega} - \widehat{\mathbf{\Theta}}$, $\mathbf{B} = (2rI_p - r^2\widehat{\mathbf{\Theta}})^{1/2} - (2rI_p - r^2\mathbf{\Omega})^{1/2}$, and $\mathbf{C} = (2rI_p - r^2\mathbf{\Omega})^{1/2}$. In view of assumption (35) and the fact that $\mathbf{\Theta} := [\mathrm{Cov}(X)]^{-1} = \frac{\nu-2}{\nu}\mathbf{\Omega}$, it follows from the triangle inequality that with probability $1 - o(1)$,

$$\begin{aligned} \|\widehat{\mathbf{\Theta}} - \mathbf{\Omega}\|_2 &\le \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 + \|\mathbf{\Theta} - \mathbf{\Omega}\|_2 = \|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}\|_2 + 2\nu^{-1}\|\mathbf{\Omega}\|_2 \\ &\le C\rho_n\sqrt{\frac{\log p}{n}} + 2\nu^{-1}C_l^{-1}. \end{aligned} \tag{A.30}$$

Now we deal with the three terms on the right-hand side of (A.29) above separately. First, for the second term above, an application of similar arguments as for (A.52) gives that with probability $1 - o(1)$,

$$\max_{1\le j\le p} n^{-1}\|(\mathbf{Z}\mathbf{B})_j\|_2^2 \le 3\|\mathbf{B}\|_2^2/2 \le C\|\widehat{\mathbf{\Theta}} - \mathbf{\Omega}\|_2^2 \le C\Big(\frac{\rho_n^2\log p}{n} + \nu^{-2}\Big). \tag{A.31}$$

Regarding the first term on the right-hand side of (A.29) above, observe that

$$(\mathbf{X}_{i,j}, \mathbf{X}_{i,l}) \overset{d}{=} (\frac{\eta_{i,j}}{\sqrt{Q_i/\nu}}, \frac{\eta_{i,l}}{\sqrt{Q_i/\nu}}),$$

where $(\eta_{i,1}, \cdots, \eta_{i,p}) \overset{d}{\sim} N(\mathbf{0}, \mathbf{\Omega}^{-1})$ and $\{Q_i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) chi-square random variables with ν degrees of freedom. It holds that for some large constant $C_1 > 0$,

$$\begin{aligned} &\mathbb{P}\Big(\|n^{-1}\mathbf{X}^T\mathbf{X} - \mathbf{\Theta}^{-1}\|_{\max} \ge C_1\sqrt{\frac{\log p}{n}} + \nu^{-1/2}\Big) \\ &= \mathbb{P}\Big(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n \frac{\eta_{i,j}\eta_{i,l}}{Q_i/\nu} - \mathbb{E}(\eta_{i,j}\eta_{i,l})\mathbb{E}(\frac{\nu}{Q_i})\Big| \ge C_1\sqrt{\frac{\log p}{n}} + \nu^{-1/2}\Big) \\ &\le \mathbb{P}\Big(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n \frac{\nu}{Q_i}(\eta_{i,j}\eta_{i,l} - \mathbb{E}(\eta_{i,j}\eta_{i,l}))\Big| \ge C_1\sqrt{\frac{\log p}{n}}\Big) \\ &\quad + \mathbb{P}\Big(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n \mathbb{E}(\eta_{i,j}\eta_{i,l})\Big(\frac{\nu}{Q_i} - \mathbb{E}(\frac{\nu}{Q_i})\Big)\Big| \ge \nu^{-1/2}\Big). \end{aligned} \tag{A.32}$$

Before showing the bounds for the two probabilities on the right-hand side of the expression above, we first present some basic results for chi-square random variables. Note that from the property of the chi-square distribution, we have through some immediate calculations that

$$\mathbb{E}\Big(\frac{\nu^2}{Q_i^2}\Big) = \frac{\nu^2}{(\nu-2)(\nu-4)}, \tag{A.33}$$

$$\mathrm{Var}\Big(\frac{\nu}{Q_i}\Big) = \frac{\nu^2}{(\nu-2)(\nu-4)} - (\frac{\nu}{\nu-2})^2 = O(\nu^{-1}), \tag{A.34}$$

$$\mathrm{Var}\Big(\frac{\nu^2}{Q_i^2}\Big) = \frac{\nu^4}{(\nu-2)(\nu-4)(\nu-6)(\nu-8)} - \Big(\frac{\nu^2}{(\nu-2)(\nu-4)}\Big)^2 = O(\nu^{-1}). \tag{A.35}$$

Thus, noting that $\mathbb{E}\Big(\frac{\nu^2}{Q_i^2}\Big) + \nu^{-1/2} = \frac{\nu^2}{(\nu-2)(\nu-4)} + \nu^{-1/2} \le 3$ and $\mathbb{E}\Big(\frac{\nu^2}{Q_i^2}\Big) - \nu^{-1/2} \ge 2/3$ when $\nu \ge 9$, an application of the Markov inequality leads to

$$\begin{aligned}
&\mathbb{P}\bigg(n^{-1}\sum_{i=1}^n \frac{\nu^2}{Q_i^2} \ge 3\bigg) + \mathbb{P}\bigg(n^{-1}\sum_{i=1}^n \frac{\nu^2}{Q_i^2} \le 2/3\bigg) \\
&\le \mathbb{P}\bigg(n^{-1}\sum_{i=1}^n \frac{\nu^2}{Q_i^2} \ge \mathbb{E}\Big(\frac{\nu^2}{Q_i^2}\Big) + \nu^{-1/2}\bigg) + \mathbb{P}\bigg(n^{-1}\sum_{i=1}^n \frac{\nu^2}{Q_i^2} \le \mathbb{E}\Big(\frac{\nu^2}{Q_i^2}\Big) - \nu^{-1/2}\bigg) \\
&\le \nu n^{-1} \operatorname{Var}\Big(\frac{\nu^2}{Q_i^2}\Big) = O(n^{-1}) \to 0.
\end{aligned} \tag{A.36}$$

In addition, noting that $e^{-x/2} \le 1$ and Stirling's formula for the gamma function $\Gamma(x) = \sqrt{2\pi/x}(x/e)^x(1+O(x^{-1}))$ for $x > 0$, we have through applying the density function of the chi-square distribution that for each constant $C > 0$,

$$\begin{aligned}
\mathbb{P}\bigg(\max_{1\le i\le n}\frac{\nu}{Q_i} \ge C\sqrt{\frac{n}{\log p}}\bigg) &\le n\int_0^{C^{-1}\nu\sqrt{\frac{\log p}{n}}} \frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}\,dx \\
&\le \frac{2n(C^{-1}\nu\sqrt{\frac{\log p}{n}})^{\nu/2}}{\nu 2^{\nu/2}\Gamma(\nu/2)} \\
&\lesssim n\Big(C^{-2}\frac{\log p}{n}\Big)^{\nu/4}\frac{\nu^{\nu/2}}{\nu 2^{\nu/2}\sqrt{4\pi/\nu}(\nu/2e)^{\nu/2}} \\
&= \Big(C^{-2}e^2\frac{\log p}{n^{1-4/\nu}}\Big)^{\nu/4}\frac{1}{\sqrt{4\pi\nu}} \to 0
\end{aligned} \tag{A.37}$$

when $\log p = o(n^{1-4/\nu})$.

Now we are ready to deal with the two probabilities on the right-hand side of (A.32) above. Let us define two events $\mathscr{D}_1 = \{\max_{1\le i\le n}\frac{\nu}{Q_i} \le C_2\sqrt{\frac{n}{\log p}}\}$ for a small constant $C_2 > 0$ and $\mathscr{D}_2 = \{2/3 \le n^{-1}\sum_{i=1}^n \frac{\nu^2}{Q_i^2} \le 3\}$. It follows from (A.36) and (A.37) that $\mathbb{P}(\mathscr{D}_1^c) \to 0$ and $\mathbb{P}(\mathscr{D}_2^c) \to 0$. For the first probability in (A.32) above, since $\eta_{i,j}\eta_{i,l}$ is a sub-exponential random variable and $Q_i \perp\!\!\!\perp \eta_{i,j}\eta_{i,l}$, we can obtain by applying the concentration inequality for the weighted sum of sub-exponential random variables (cf. Corollary 4.2 in Zhang and Chen (2021)) that when C_1 is large enough and C_2 is small enough,

$$\begin{aligned}
&\mathbb{P}\bigg(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n \frac{\nu}{Q_i}(\eta_{i,j}\eta_{i,l} - \mathbb{E}(\eta_{i,j}\eta_{i,l}))\Big| \ge C_1\sqrt{\frac{\log p}{n}}\bigg) \\
&\le \mathbb{P}\bigg(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n \frac{\nu}{Q_i}(\eta_{i,j}\eta_{i,l} - \mathbb{E}(\eta_{i,j}\eta_{i,l}))\Big| \ge C_1\sqrt{\frac{\log p}{n}}, \mathscr{D}_1\cap\mathscr{D}_2\bigg) \\
&\quad + \mathbb{P}(\mathscr{D}_1^c) + \mathbb{P}(\mathscr{D}_2^c) \\
&\le 2p^2\exp\{-3\log p\} + o(1) \to 0.
\end{aligned} \tag{A.38}$$

Regarding the second probability in (A.32), since $\max_{1\le j,l\le p}|\mathbb{E}(\eta_{i,j}\eta_{i,l})|\le \max_{1\le j\le p}\mathbb{E}(\eta_{i,j}^2)\le \max_{1\le j\le p}(\mathbf{\Omega}^{-1})_{j,j}\le C_u$, an application of the Markov inequality and (A.34) yields that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le j,l\le p}\Big|n^{-1}\sum_{i=1}^n\mathbb{E}(\eta_{i,j}\eta_{i,l})\Big(\frac{\nu}{Q_i}-\mathbb{E}(\frac{\nu}{Q_i})\Big)\Big|\ge \nu^{-1/2}\Big)\\
&\le \mathbb{P}\Big(\Big|n^{-1}\sum_i\Big(\frac{\nu}{Q_i}-\mathbb{E}(\frac{\nu}{Q_i})\Big)\Big|\ge C_u^{-1}\nu^{-1/2}\Big)\\
&\le C_u^{-2}\nu n^{-1}\operatorname{Var}(\frac{\nu}{Q_i})=O(n^{-1})\to 0.
\end{aligned}\tag{A.39}$$

By plugging (A.38) and (A.39) into (A.32), we can show that with probability $1-o(1)$,

$$\max_{\delta:\|\delta\|_0\le\rho_n}\frac{|\delta^T(n^{-1}\mathbf{X}^T\mathbf{X}-\mathbf{\Theta}^{-1})\delta|}{\|\delta\|_2^2}\le C\rho_n\Big(\sqrt{\frac{\log p}{n}}+\nu^{-1/2}\Big),$$

which along with the fact $\|\mathbf{\Theta}^{-1}\|_2=\frac{\nu}{\nu-2}\|\mathbf{\Omega}^{-1}\|_2\le\frac{\nu}{\nu-2}C_u$ entails that as $\rho_n=o(\sqrt{n/(\log p)})$ and $\rho_n=o(\sqrt{\nu})$,

$$\max_{\delta:\|\delta\|_0\le\rho_n}\frac{\delta^T\mathbf{X}^T\mathbf{X}\delta}{n\|\delta\|_2^2}\le C\tag{A.40}$$

for some constant $C>0$. Using (A.30) and the sparsity assumption that $\max_{1\le j\le p}\|\mathbf{\Omega}_j\|_0+\|\mathbf{\Omega}_n\|_0\le\rho_n$, an application of similar arguments as for (A.49) gives that with probability $1-o(1)$,

$$\begin{aligned}
\max_{1\le j\le p}n^{-1}\|\mathbf{X}\mathbf{A}_j\|_2^2&=n^{-1}\mathbf{A}_j^T\mathbf{X}^T\mathbf{X}\mathbf{A}_j\le C\max_{1\le j\le p}\|\mathbf{A}_j\|_2^2\\
&=C\|\widehat{\mathbf{\Theta}}-\mathbf{\Omega}\|_2^2\le C\Big(\frac{\rho_n^2\log p}{n}+\nu^{-2}\Big).
\end{aligned}\tag{A.41}$$

We now proceed with examining the third term on the right-hand side of (A.29) above. Observe that $\mathbf{Z}\mathbf{C}_j\overset{d}{\sim}N(\mathbf{0},\|\mathbf{C}_j\|_2^2I_n)$ and $\max_{1\le j\le p}\|\mathbf{C}_j\|_2\le\|\mathbf{C}\|_2\le 2r$. Hence, it holds for some large constant $C_3>0$ that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le j\le p}n^{-1}\Big\|\operatorname{diag}(1-\frac{1}{\sqrt{Q_1/\nu}},\cdots,1-\frac{1}{\sqrt{Q_n/\nu}})\mathbf{Z}\mathbf{C}_j\Big\|_2^2\ge C_3\nu^{-1}\Big)\\
&=\mathbb{P}\Big(\max_{1\le j\le p}n^{-1}\sum_{i=1}^n\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2\|\mathbf{C}_j\|^2Z_i^2\ge C_3\nu^{-1}\Big)\\
&\le\mathbb{P}\Big(n^{-1}\sum_{i=1}^n\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2Z_i^2\ge C_3\nu^{-1}/4r^2\Big),
\end{aligned}\tag{A.42}$$

where $\{Z_i\}_{i=1}^n$ are i.i.d. standard normal random variables that are independent of $\mathbf{C}$ and $\{Q_i\}_{i=1}^n$.

Similar to the calculations in (A.34) and (A.35), we can deduce that

$$\begin{aligned}
\mathbb{E}\Big[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2Z_i^2\Big]&=\mathbb{E}(Z_i^2)\mathbb{E}\Big[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2\Big]\\
&=1-\mathbb{E}\Big(\frac{2}{\sqrt{Q_i/\nu}}\Big)+\mathbb{E}\Big(\frac{1}{Q_i/\nu}\Big)\\
&=1-\frac{\sqrt{2\nu}\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}+\frac{\nu}{\nu-2}
\end{aligned}\tag{A.43}$$

and

$$\begin{aligned}&\mathbb{E}\bigg[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^4 Z_i^4\bigg]\\ &=3\bigg(1-\frac{2\sqrt{2}\nu\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}+\frac{6(\nu-2)}{\nu}-\frac{\sqrt{2}\nu^{3/2}\Gamma(\frac{\nu-3}{2})}{\Gamma(\frac{\nu}{2})}+\frac{\nu^2}{(\nu-2)(\nu-4)}\bigg).\end{aligned} \tag{A.44}$$

By applying the asymptotic series of the gamma function

$$\frac{\Gamma(x+1/2)}{\Gamma(x)}=\sqrt{x}\Big(1-\frac{1}{8x}+O(x^{-2})\Big),$$

we can obtain through some direct calculations that

$$\mathbb{E}\bigg[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2 Z_i^2\bigg]=O(\nu^{-1}) \text{ and } \mathbb{E}\bigg[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^4 Z_i^4\bigg]=O(\nu^{-2}). \tag{A.45}$$

Combining (A.42) and (A.45) and applying the Markov inequality, we have that for some large enough constant $C_3>0$,

$$\begin{aligned}&\mathbb{P}\bigg(\max_{1\le j\le p} n^{-1}\Big\|\mathrm{diag}(1-\frac{1}{\sqrt{Q_1/\nu}},\cdots,1-\frac{1}{\sqrt{Q_n/\nu}})\mathbf{Z}\mathbf{C}_j\Big\|_2^2\ge C_3\nu^{-1}\bigg)\\ &\le\mathbb{P}\bigg(n^{-1}\sum_{i=1}^n\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2 Z_i^2-\mathbb{E}\bigg[\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2 Z_i^2\bigg]\\ &\quad\ge C_3(\nu^{-1})/4r^2-O(\nu^{-1})\bigg)\\ &\le C\nu^{-2}n^{-1}\,\mathrm{Var}\bigg(\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^2 Z_i^2\bigg)\\ &\le C\nu^{-2}n^{-1}\mathbb{E}\bigg(\Big(\Big(1-\frac{1}{\sqrt{Q_i/\nu}}\Big)^4 Z_i^4\Big)\bigg)\\ &=O(n^{-1})\to 0.\end{aligned} \tag{A.46}$$

Therefore, a combination of (A.29), (A.31), (A.41), and (A.46) yields the desired conclusion in (36). This concludes the proof of Proposition 2.

A.9. Proof of Proposition 3. It follows from (37) and (38) that

$$\widehat{\mathbf{X}}-\widetilde{\mathbf{X}}=r\mathbf{X}\mathbf{A}+\mathbf{Z}\mathbf{B}, \tag{A.47}$$

where $\mathbf{A}=\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}$ and $\mathbf{B}=(2rI_p-r^2\widehat{\boldsymbol{\Omega}})^{1/2}-(2rI_p-r^2\boldsymbol{\Omega})^{1/2}$. By the Gaussianity of X, we see that X_jX_l is a sub-exponential random variable and thus for $0<u<C$,

$$\mathbb{P}(|n^{-1}\mathbf{X}_j^T\mathbf{X}_l-\mathbb{E}(X_jX_l)|\ge u)\le 2\exp\{-Cnu^2\}.$$

Then we can obtain that

$$\mathbb{P}\bigg(\max_{1\le j\le p,1\le l\le p}|n^{-1}\mathbf{X}_j\mathbf{X}_l-\mathbb{E}(X_jX_l)|\ge C\sqrt{\frac{\log p}{n}}\bigg)=o(1).$$

Consequently, with probability $1-o(1)$ it holds that

$$\max_{\delta:\|\delta\|_0\le\rho_n}\frac{|\delta^T(n^{-1}\mathbf{X}^T\mathbf{X}-\boldsymbol{\Omega}^{-1})\delta|}{\|\delta\|_2^2}\le C\rho_n\sqrt{\frac{\log p}{n}},$$

which combined with the assumption that $\|\mathbf{\Omega}^{-1}\|_2 \le C_u$ leads to

$$\max_{\delta:\|\delta\|_0\le\rho_n} \frac{\delta^T\mathbf{X}^T\mathbf{X}\delta}{n\|\delta\|_2^2} \le C_u + C\rho_n\sqrt{\frac{\log p}{n}} \le \widetilde{C} \tag{A.48}$$

for some constant $\widetilde{C} > 0$. Since $\|\mathbf{A}_j\|_0 = \|(\mathbf{\Omega} - \widehat{\mathbf{\Omega}})_j\|_0 \le C\rho_n$ because of the sparsity of $\mathbf{\Omega}$ and $\widehat{\mathbf{\Omega}}$, it follows from (A.48) that with probability $1 - o(1)$,

$$\begin{aligned}\max_{1\le j\le p} n^{-1}\|(\mathbf{XA})_j\|_2^2 &= \max_{1\le j\le p} n^{-1}\|\mathbf{XA}_j\|_2^2 \le \max_{1\le j\le p}\widetilde{C}\|\mathbf{A}_j\|_2^2 \\ &= \max_{1\le j\le p}\widetilde{C}\|(\widehat{\mathbf{\Omega}} - \mathbf{\Omega})_j\|_2^2 \le \max_{1\le j\le p}\widetilde{C}\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2^2 \\ &\le \widetilde{C}\frac{\rho_n^2\log p}{n},\end{aligned} \tag{A.49}$$

where we have used the accuracy assumption in (39).

Next we proceed with analyzing the term $\mathbf{ZB}$. Observe that given $\mathbf{B}$, $\mathbf{Z}$ has i.i.d. standard normal components and is independent of $\mathbf{B}$, and hence

$$\mathbf{ZB}_j|\mathbf{B}_j \overset{d}{\sim} N(\mathbf{0}, \|\mathbf{B}_j\|_2^2 I_n).$$

It holds that $\mathbf{ZB}_j|\mathbf{B}_j \overset{d}{=} (Z_1\|\mathbf{B}_j\|_2, \cdots, Z_n\|\mathbf{B}_j\|_2)$ with $\{Z_i\}_{i=1}^n$ i.i.d. standard normal random variables. Then we can deduce that

$$\begin{aligned}&\mathbb{P}(\max_{1\le j\le p} n^{-1}\|(\mathbf{ZB})_j\|_2^2 \ge 3\|\mathbf{B}\|_2^2/2\big|\mathbf{B}) \\ &= \mathbb{P}(\max_{1\le j\le} n^{-1}\|\mathbf{ZB}_j\|_2^2 \ge 3\|\mathbf{B}\|_2^2/2\big|\mathbf{B}) \\ &= \mathbb{P}\Big(\max_{1\le j\le p} n^{-1}\sum_{i=1}^n Z_i^2\|\mathbf{B}_j\|_2^2 \ge 3\|\mathbf{B}\|_2^2/2\big|\mathbf{B}\Big) \\ &\le \mathbb{P}\Big(n^{-1}\sum_{i=1}^n Z_i^2\|\mathbf{B}\|_2^2 \ge 2\|\mathbf{B}\|_2^2\big|\mathbf{B}\Big) \\ &= \mathbb{P}\Big(n^{-1}\sum_{i=1}^n Z_i^2 \ge 3/2\Big) \le e^{-n/32} \to 0\end{aligned} \tag{A.50}$$

as $n \to \infty$, where we have used the fact that $\max_{1\le j\le p}\|\mathbf{B}_j\|_2 \le \|\mathbf{B}\|_2$ and the concentration inequality for chi-square random variables that for $0 < t < 1$,

$$\mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n Z_i^2 - 1\Big| \ge t\Big) \le 2e^{-nt^2/8}.$$

Now we aim to bound $\|\mathbf{B}\|_2$. For two square matrices A and B, it holds that

$$\begin{aligned}\|A^{1/2} - B^{1/2}\|_2 &= \|A^{1/2}(B - A)B^{-1} + (A^{3/2} - B^{3/2})B^{-1}\|_2 \\ &\le \|A^{1/2}(B - A)B^{-1}\|_2 + 3\max\{\|A\|_2^{1/2}, \|B\|_2^{1/2}\}\|A - B\|_2\|B^{-1}\|_2.\end{aligned}$$

Applying the above inequality to $\mathbf{B}$ leads to

$$\begin{aligned}\|\mathbf{B}\|_2 &\le \|2rI_p - r^2\widehat{\mathbf{\Omega}}\|_2^{1/2}\cdot r^2\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2\cdot\|2rI_p - r^2\mathbf{\Omega}\|^{-1} \\ &\quad + 3\max\{\|2rI_p - r^2\widehat{\mathbf{\Omega}}\|_2^{1/2}, \|2rI_p - r^2\mathbf{\Omega}\|_2^{1/2}\}\cdot r^2\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2\cdot\|2rI_p - r^2\mathbf{\Omega}\|^{-1} \\ &\le C\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2.\end{aligned} \tag{A.51}$$

Thus, from (A.50) and assumption (39), we can obtain that with probability $1-o(1)$,

$$\max_{1\le j\le p} n^{-1}\|(\mathbf{ZB})_j\|_2^2 \le 3\|\mathbf{B}\|_2^2/2 \le C\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_2^2 \le C\frac{\rho_n^2\log p}{n}. \tag{A.52}$$

Note that

$$\|\widehat{\mathbf{X}}_j-\widetilde{\mathbf{X}}_j\|_2 \le r\|\mathbf{X}\mathbf{A}_j\|_2+\|\mathbf{Z}\mathbf{B}_j\|_2.$$

Therefore, in view of (A.49) and (A.52) we can show that for some constant $C>0$,

$$\mathbb{P}\left(n^{-1/2}\|\widehat{\mathbf{X}}_j-\widetilde{\mathbf{X}}_j\|_2 \le C\rho_n\sqrt{\frac{\log p}{n}}\right)\to 1. \tag{A.53}$$

This completes the proof of Proposition 3.

A.10. Proof of Proposition 4. In light of the definitions of $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$, we can obtain through the triangle inequality that

$$\begin{aligned} &n^{-1/2}\max_{1\le j\le p}\|\widehat{\mathbf{X}}_j-\widetilde{\mathbf{X}}_j\|_2 \\ &\le \max_{1\le j\le p} n^{-1/2}\left(\sum_{i=1}^n\left[\widehat{F}_j^{-1}(\Phi(\widehat{\mathbf{U}}_{i,j}))-\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\right]^2\right)^{1/2} \\ &\quad+\max_{1\le j\le p} n^{-1/2}\left(\sum_{i=1}^n\left[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))-F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\right]^2\right)^{1/2}. \end{aligned} \tag{A.54}$$

We claim that

$$\mathbb{P}\left(\max_{1\le j\le p} n^{-1}\sum_{i=1}^n\left[\widehat{F}_j^{-1}(\Phi(\widehat{\mathbf{U}}_{i,j}))-\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\right]^2 \ge \widetilde{C}\left(\frac{\rho_n^2\log p}{n}+\frac{p\rho_n(\log n)^3}{n}\right)\right)$$

$$\to 0, \tag{A.55}$$

$$\mathbb{P}\left(\max_{1\le j\le p} n^{-1}\sum_{i=1}^n\left[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))-F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\right]^2 \ge \frac{2Mp(\log n)^2}{n}\right)\to 0, \tag{A.56}$$

which together with (A.54) yield the desired conclusion of Proposition 4. It remains to establish (A.55) and (A.56). We will begin with the proof of (A.55).

Proof of (A.55). From assumption (48) and the observation that $\frac{\log n}{n^2} \ll \frac{p\rho_n(\log n)^3}{n}$, it holds that for some large constant $C > 0$,

(A.57)
$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widehat{\mathbf{U}}_{i,j})) - \widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \ge C\Big(\frac{\rho_n^2 \log p}{n} + \frac{p\rho_n(\log n)^3}{n}\Big)\Big) \\
&\le \mathbb{P}\Big(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \Big[|\Phi(\widehat{\mathbf{U}}_{i,j}) - \Phi(\widetilde{\mathbf{U}}_{i,j})|^2 + (\log n)^2 n^{-2} \\
&\qquad + n^{-1}(\log n)|\Phi(\widehat{\mathbf{U}}_{i,j}) - \Phi(\widetilde{\mathbf{U}}_{i,j})|\Big] \ge C\Big(\frac{\rho_n^2 \log p}{n} + \frac{p\rho_n(\log n)^3}{n}\Big)\Big) \\
&\quad + \mathbb{P}\Big(\max_{1\le j\le p} \sup_{x,y\in(0,1)} \frac{\big|\widehat{F}_j^{-1}(x) - \widehat{F}_j^{-1}(y)\big|}{|x-y| + (n^{-1}(\log n)|x-y|)^{1/2} + n^{-1}\log n} \ge M\Big) \\
&\le \mathbb{P}\Big(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \Big[|\Phi(\widehat{\mathbf{U}}_{i,j}) - \Phi(\widetilde{\mathbf{U}}_{i,j})|^2 \\
&\qquad + n^{-1}(\log n)|\Phi(\widehat{\mathbf{U}}_{i,j}) - \Phi(\widetilde{\mathbf{U}}_{i,j})|\Big] \ge C\Big(\frac{\rho_n^2 \log p}{n} + \frac{p\rho_n(\log n)^3}{n}\Big)\Big) + o(1) \\
&:= P_1 + o(1).
\end{aligned}$$

We next bound term P_1 above. Using the fact that $|\Phi(x) - \Phi(y)| \le \frac{1}{\sqrt{2\pi}}|x-y|$ and the basic inequality $\sum_{i=1}^n |a_n| \le \sqrt{n}(\sum_{i=1}^n a_n{}^2)^{1/2}$, we have that

(A.58)
$$\begin{aligned}
P_1 \le \mathbb{P}\Big(&\max_{1\le j\le p} \Big(n^{-1}\|\widehat{\mathbf{U}}_j - \widetilde{\mathbf{U}}_j\|_2^2 + (\log n) n^{-3/2}\|\widehat{\mathbf{U}}_j - \widetilde{\mathbf{U}}_j\|_2\Big) \\
&\ge C\Big(\frac{\rho_n^2 \log p}{n} + \frac{p\rho_n(\log n)^3}{n}\Big)\Big).
\end{aligned}$$

It suffices to consider the bound of $\max_{1\le j\le p} n^{-1}\|\widehat{\mathbf{U}}_j - \widetilde{\mathbf{U}}_j\|_2^2$. With the aid of the triangle inequality and the definitions of $\widehat{\mathbf{U}}$ and $\widetilde{\mathbf{U}}$, it follows that

(A.59)
$$\begin{aligned}
\max_{1\le j\le p} n^{-1}\|\widehat{\mathbf{U}}_j - \widetilde{\mathbf{U}}_j\|_2^2 \le\ & 3 \max_{1\le j\le p} n^{-1}\|(\widehat{\mathbf{V}} - \widetilde{\mathbf{V}})(I_p - r\widehat{\mathbf{\Omega}})_j\|_2^2 \\
&+ 3r^2 \max_{1\le j\le p} n^{-1}\|\widetilde{\mathbf{V}}(\widehat{\mathbf{\Omega}}_j - \mathbf{\Omega}_j)\|_2^2 \\
&+ 3 \max_{1\le j\le p} n^{-1}\|\mathbf{Z}[(2rI_p - r^2\widehat{\mathbf{\Omega}})^{1/2} - (2rI_p - r^2\mathbf{\Omega})^{1/2}]\|_2^2.
\end{aligned}$$

We will investigate the three terms in the upper bound above separately. Regarding the third term above, under the assumption in (39) it has been shown in (A.52) that with probability $1 - o(1)$,

(A.60)
$$\max_{1\le j\le p} n^{-1}\|\mathbf{Z}[(2rI_p - r^2\widehat{\mathbf{\Omega}})^{1/2} - (2rI_p - r^2\mathbf{\Omega})^{1/2}]\|_2^2 \le C\frac{\rho_n^2 \log p}{n}.$$

As for the second term in the upper bound in (A.59), noting that the rows of $\widetilde{\mathbf{V}}$ are i.i.d. and follow the Gaussian distribution $N(\mathbf{0}, \mathbf{\Omega}^{-1})$, an application of similar arguments as for

(A.49) gives that with probability $1-o(1)$,

$$\max_{1\le j\le p} n^{-1}\|\widetilde{\mathbf{V}}(\widehat{\mathbf{\Omega}}_j-\mathbf{\Omega}_j)\|_2^2 \le C\frac{\rho_n^2\log p}{n}. \tag{A.61}$$

For the first term in the upper bound in (A.59) above, noting that $\|I_p - r\widehat{\mathbf{\Omega}})_j\| \le \rho_n+1$ by the sparsity assumption that $\|\widehat{\mathbf{\Omega}}_j\|\le\rho_n$, we have that

$$\begin{aligned}\max_{1\le j\le p} n^{-1}\|(\widehat{\mathbf{V}}-\widetilde{\mathbf{V}})(I_p-r\widehat{\mathbf{\Omega}})_j\|_2^2 \le & \max_{J:|J|\le\rho_n+1}\|n^{-1}(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)^T(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)\|_2 \\ & \times \max_{1\le j\le p}\|(I_p-r\widehat{\mathbf{\Omega}})_j\|_2^2.\end{aligned} \tag{A.62}$$

For the second term in the bound above, from the triangle inequality and inequality $\|\mathbf{A}_j\|_2\le\|\mathbf{A}\|_2$ for each matrix $\mathbf{A}$, it is easy to see that

$$\max_{1\le j\le p}\|(I_p-r\widehat{\mathbf{\Omega}})_j\|_2 \le \|I_p-r\widehat{\mathbf{\Omega}}\|_2 \le \|I_p-r\mathbf{\Omega}\|_2 + r\|\widehat{\mathbf{\Omega}}-\mathbf{\Omega}\|_2.$$

Thus it follows from assumption (39) that for a constant $C>0$, with probability $1-o(1)$ we have

$$\max_{1\le j\le p}\|(I_p-r\widehat{\mathbf{\Omega}})_j\|_2\le C. \tag{A.63}$$

Regarding the first term on the right-hand side of (A.62) above, using the definitions of $\widehat{\mathbf{V}}$ and $\widetilde{\mathbf{V}}$, and inequality $\|\mathbf{A}\|_2\le d\|\mathbf{A}\|_{\max}$ for each square matrix $\mathbf{A}\in\mathbb{R}^{d\times d}$, we can deduce that

$$\begin{aligned}&\max_{J:|J|\le\rho_n+1}\|n^{-1}(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)^T(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)\|_2 \\ &\le(\rho_n+1)\|n^{-1}(\widehat{\mathbf{V}}-\widetilde{\mathbf{V}})^T(\widehat{\mathbf{V}}-\widetilde{\mathbf{V}})\|_{\max} \\ &\le(\rho_n+1)\max_{1\le j\le p} n^{-1}\sum_{i=1}^n|\widehat{\mathbf{V}}_{i,j}-\widetilde{\mathbf{V}}_{i,j}|^2 \\ &=(\rho_n+1)\max_{1\le j\le p} n^{-1}\sum_{i=1}^n|\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))-\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2.\end{aligned} \tag{A.64}$$

Denote by $H_{j,n}=[F_j^{-1}(2Mn^{-1}\log n), F_j^{-1}(1-2Mn^{-1}\log n)]$ with constant M as given in assumption (47). We can write that

$$\begin{aligned}&\max_{1\le j\le p} n^{-1}\sum_{i=1}^n|\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))-\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2 \\ &=\max_{1\le j\le p} n^{-1}\sum_{i=1}^n|\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))-\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2\mathbb{1}(\mathbf{X}_{i,j}\in H_{j,n}) \\ &\quad+\max_{1\le j\le p} n^{-1}\sum_{i=1}^n|\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))-\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2\mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}) \\ &:=E_1+E_2.\end{aligned} \tag{A.65}$$

Let us first consider term E_2 above. Observe that

$$\begin{aligned} E_2 \leq & \max_{1\leq j\leq p} n^{-1}\sum_{i=1}^{n} |\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))|^2 \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}) \\ & + \max_{1\leq j\leq p} n^{-1}\sum_{i=1}^{n} |\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2 \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}). \end{aligned} \tag{A.66}$$

For the first term in the bound above, notice that

$$|\Phi^{-1}(\widehat{F}_j(\mathbf{X}_{i,j}))| = O(\sqrt{\log n})$$

due to the assumption that $\frac{1}{2n} \leq F_j(x) \leq 1-\frac{1}{2n}$ for each $x\in \mathrm{supp}(X_j)$. Then it follows from the union bound, the Markov inequality, and the definition of $H_{j,n}$ that

$$\begin{aligned} & \mathbb{P}\bigg(\max_{1\leq j\leq p} n^{-1}\sum_{i=1}^{n} |\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2 \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}) \geq \frac{p(\log n)^3}{n}\bigg) \\ & \leq \sum_{j=1}^{p} \mathbb{P}\bigg(n^{-1}\log n \sum_{i=1}^{n} \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}) \geq \frac{p(\log n)^3}{n}\bigg) \\ & \leq \frac{n}{p(\log n)^2}\sum_{j=1}^{p} \mathbb{P}(\mathbf{X}_{i,j}\notin H_{j,n}) \\ & = \frac{pn}{p(\log n)^2}\cdot 4Mn^{-1}\log n \\ & = 4M(\log n)^{-1} \to 0. \end{aligned} \tag{A.67}$$

As for the second term in the upper bound in (A.66) above, an application of the Markov inequality and the fact that $F_j(\mathbf{X}_{i,j})$ follows the standard uniform distribution gives that

$$\begin{aligned} & \mathbb{P}\bigg(\max_{1\leq j\leq p} n^{-1}\sum_{i=1}^{n} |\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2 \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n}) \geq \frac{p(\log n)^3}{n}\bigg) \\ & \leq \frac{n}{p(\log n)^3}\sum_{j=1}^{p} \mathbb{E}\Big(|\Phi^{-1}(F_j(\mathbf{X}_{i,j}))|^2 \mathbb{1}(\mathbf{X}_{i,j}\notin H_{j,n})\Big) \\ & = \frac{2n}{(\log n)^3}\int_{-\infty}^{\Phi^{-1}(\frac{2M\log n}{n})} \frac{1}{\sqrt{2\pi}} u^2 e^{-u^2/2} du \\ & \leq \frac{2n}{(\log n)^3 |\Phi^{-1}(\frac{2M\log n}{n})|}\int_{-\infty}^{\Phi^{-1}(\frac{2M\log n}{n})} \frac{1}{\sqrt{2\pi}} |u|^3 e^{-u^2/2} du \\ & \leq C\frac{n}{(\log n)^3 |\Phi^{-1}(\frac{2M\log n}{n})|}\cdot \Big|\Phi^{-1}(\frac{2M\log n}{n})\Big|^3 \cdot \Phi(\Phi^{-1}(\frac{2M\log n}{n})) \\ & \leq C(\log n)^{-1} \to 0, \end{aligned} \tag{A.68}$$

where in the last step above, we have used the facts that $|\Phi^{-1}(\frac{M\log n}{n})| \leq C\sqrt{\log n}$, $\int u^3 e^{-u^2/2} du = -(u^2+2)e^{-u^2/2}$, and $e^{-x^2/2}/\Phi(x) = O(|x|)$ for $x<-2$. Combining (A.66), (A.67), and (A.68) yields that with probability $1-o(1)$,

$$E_2 \leq \frac{p(\log n)^3}{n}. \tag{A.69}$$

Next we proceed with studying term E_1. First, note that when $|\Phi^{-1}(y)| > 2$, it holds that

$$[\Phi^{-1}(y)]' = \frac{1}{\Phi'(\Phi^{-1}(y))} \leq C \frac{1}{(y \wedge (1-y))|\Phi^{-1}(y)|}$$

due to the fact that $\Phi'(x)/(1-\Phi(x)) \geq Cx$ for $x > 2$ and $\Phi'(x)/\Phi(x) \geq C|x|$ for $x < -2$. When $|\Phi^{-1}(y)| \leq 2$, it is easy to see that

$$[\Phi^{-1}(y)]' = \frac{1}{\Phi'(\Phi^{-1}(y))} \leq C.$$

Thus, combining the previous two results shows that for $y \in \mathbb{R}$,

$$[\Phi^{-1}(y)]' \leq \frac{C}{(y \wedge (1-y))|\Phi^{-1}(y)|} \leq \frac{C}{(y \wedge (1-y))}. \tag{A.70}$$

Let us define an interval

$$\delta_j(x) = \left[F_j(x) - \sqrt{\frac{M[F_j(x) \wedge (1-F_j(x))] \log n}{n}}, F_j(x) + \sqrt{\frac{M[F_j(x) \wedge (1-F_j(x))] \log n}{n}} \right].$$

Observe that under assumption (47), we have that

$$\begin{aligned} &\mathbb{P}(E_1 \geq x) \\ &\leq \mathbb{P}\bigg(\max_{1 \leq j \leq p} n^{-1} \Big(\frac{M \log n}{n}\Big) \sum_{i=1}^{n} \Big(\sup_{y \in \delta_j(\mathbf{X}_{i,j})} [\Phi^{-1}(y)]' \Big)^2 F_j(\mathbf{X}_{i,j})(1 - F_j(\mathbf{X}_{i,j}) \\ &\quad \cdot \mathbb{1}(\mathbf{X}_{i,j} \in H_{j,n}) \geq x \bigg) + o(1). \end{aligned} \tag{A.71}$$

When $\mathbf{X}_{i,j} \in H_{j,n}$, it holds that $F_j(\mathbf{X}_{i,j}) \in [2Mn^{-1} \log n, 1 - 2Mn^{-1} \log n]$ and hence

$$\sup_{y \in \delta(\mathbf{X}_{i,j})} \left| \frac{y}{F(\mathbf{X}_{i,j})} - 1 \right| \leq \sqrt{\frac{M \log n}{n F_j(\mathbf{X}_{i,j})}} \leq 1/\sqrt{2}.$$

Similarly, we have that

$$\sup_{y \in \delta(\mathbf{X}_{i,j})} \left| \frac{1-y}{1-F(\mathbf{X}_{i,j})} - 1 \right| \leq 1/\sqrt{2}.$$

The above two bounds combined with (A.70) yields that for $\mathbf{X}_{i,j} \in H_{j,n}$,

$$\sup_{y \in \delta_j(\mathbf{X}_{i,j})} [\Phi^{-1}(y)]' \leq \sup_{y \in \delta_j(\mathbf{X}_{i,j})} \frac{C}{y \wedge (1-y)} \leq \frac{C}{F_j(\mathbf{X}_{i,j}) \wedge (1 - F_j(\mathbf{X}_{i,j}))}.$$

In view of the above bound, (A.71), and the fact that $F_j(\mathbf{X}_{i,j})$ follows the standard uniform distribution, we can deduce that

$$\begin{aligned}
&\mathbb{P}\Big(E_1 \geq \frac{p(\log n)^3}{n}\Big) \\
&\leq \mathbb{P}\bigg(\max_{1\leq j\leq p} n^{-1}\big(\frac{M\log n}{n}\big)\sum_{i=1}^{n}\frac{C}{F_j(\mathbf{X}_{i,j})\wedge(1-F_j(\mathbf{X}_{i,j}))}\mathbb{1}(\mathbf{X}_{i,j}\in H_{j,n}) \\
&\qquad \geq \frac{p(\log n)^3}{n}\bigg) + o(1) \\
&\leq \frac{CM}{p(\log n)^2}\sum_{j=1}^{p}\mathbb{E}\Big(\frac{1}{F_j(\mathbf{X}_{i,j})\wedge(1-F_j(\mathbf{X}_{i,j}))}\mathbb{1}(\mathbf{X}_{i,j}\in H_{j,n})\Big) \qquad \text{(A.72)} \\
&= \frac{CM}{(\log n)^2}\int_{2Mn^{-1}\log n}^{1-2Mn^{-1}\log n}\frac{1}{u\wedge(1-u)}du \\
&\leq \frac{CM}{(\log n)^2}\cdot C\log n \\
&\leq \frac{CM}{\log n}\to 0.
\end{aligned}$$

A combination of (A.64), (A.65), (A.69), and (A.72) shows that with probability $1-o(1)$,

$$\max_{J:|J|\leq\rho_n+1}\|n^{-1}(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)^T(\widehat{\mathbf{V}}_J-\widetilde{\mathbf{V}}_J)\|_2 \leq \frac{Cp\rho_n(\log n)^3}{n}, \qquad \text{(A.73)}$$

which together with (A.59)–(A.63) entails that with probability $1-o(1)$,

$$n^{-1}\max_{1\leq j\leq p}\|\widehat{\mathbf{U}}_j-\widetilde{\mathbf{U}}_j\|_2^2 \leq C\Big(\frac{\rho_n^2\log p}{n}+\frac{p\rho_n(\log n)^3}{n}\Big) \qquad \text{(A.74)}$$

and

$$(\log n)n^{-3/2}\max_{1\leq j\leq p}\|\widehat{\mathbf{U}}_j-\widetilde{\mathbf{U}}_j\|_2 \leq C(\log n)n^{-1}\Big(\rho_n\frac{\log p}{n}+\sqrt{\frac{p\rho_n(\log n)^3}{n}}\Big). \qquad \text{(A.75)}$$

Plugging (A.74) into (A.58), it follows that

$$P_1\to 0. \qquad \text{(A.76)}$$

Therefore, substituting (A.76) into (A.77) derives the desired result (A.55). It remains to establish (A.56).

Proof of (A.56). Let us define $I_n = [2Mn^{-1}\log n, 1 - 2Mn^{-1}\log n]$. It holds that

$$
\begin{aligned}
&\mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})) - F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \ge \frac{2Mp(\log n)^2}{n} \bigg) \\
&= \mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})) - F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \mathbb{1}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \in I_n) \\
&\qquad \ge \frac{Mp(\log n)^2}{n} \bigg) \\
&\quad + \mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})) - F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \mathbb{1}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \notin I_n) \\
&\qquad \ge \frac{Mp(\log n)^2}{n} \bigg).
\end{aligned} \tag{A.77}
$$

For the first term on the right-hand side of (A.77) above, under assumption (46) we have that

$$
\begin{aligned}
&\mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})) - F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \mathbb{1}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \in I_n) \\
&\qquad \ge \frac{Mp(\log n)^2}{n} \bigg) \\
&\le \mathbb{P}\bigg(\frac{M\log n}{n} \ge \frac{Mp(\log n)^2}{n} \bigg) + o(1) \\
&= 0 + o(1) \to 0.
\end{aligned} \tag{A.78}
$$

Regarding the second term on the right-hand side of (A.77) above, observe that $|F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))| \le b$ and $|\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))| \le b$ by the assumption $\mathrm{supp}(X_j) \in [-b, b]$. In addition, $\Phi(\widetilde{\mathbf{U}}_{i,j})$ follows the standard uniform distribution and thus $\mathbb{P}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \notin I_n) = 4Mn^{-1}\log n$. Then we can deduce that

$$
\begin{aligned}
&\mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \big[\widehat{F}_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j})) - F_j^{-1}(\Phi(\widetilde{\mathbf{U}}_{i,j}))\big]^2 \mathbb{1}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \notin I_{1,n}) \\
&\qquad \ge \frac{Mp(\log n)^2}{n} \bigg) \\
&\le \mathbb{P}\bigg(\max_{1\le j\le p} n^{-1} \sum_{i=1}^{n} \mathbb{1}(\Phi(\widetilde{\mathbf{U}}_{i,j}) \notin I_n) \ge \frac{Mp(\log n)^2}{4nb^2} \bigg) \\
&\le \frac{4nb^2}{Mp(\log n)^2} \cdot p\mathbb{P}(\Phi(\widetilde{\mathbf{U}}_{i,j} \notin I_n)) \\
&= \frac{16b^2}{\log n} \to 0.
\end{aligned} \tag{A.79}
$$

Finally, combining (A.77)–(A.79) leads to the desired result (A.56). This concludes the proof of Proposition 4.

APPENDIX B: PROOFS OF SOME KEY LEMMAS

B.1. Proof of Lemma 1. We claim the following upper bound for $\mathbb{E}[\|\widehat{\mathbf{X}}-\widetilde{\mathbf{X}}\|_{1,2}^2|\mathbf{X}]$ as presented in (A.80) and lower bound for $\mathbb{W}_{1,2}(\widehat{\mu}^n,\widetilde{\mu}^n)$ as shown in (A.81)

$$\mathbb{E}[\|\widehat{\mathbf{X}}-\widetilde{\mathbf{X}}\|_{1,2}^2|\mathbf{X}] \le 2\big(1+\sqrt{2n^{-1}}\big)(r^2\vee 1)\max_{1\le j\le p}\big(n^{-1}\mathbf{A}_j^T\mathbf{X}^T\mathbf{X}\mathbf{A}_j+\|\mathbf{B}_j\|_2^2\big), \tag{A.80}$$

$$\mathbb{W}_{1,2}(\widehat{\mu}^n,\widetilde{\mu}^n) \ge \max_{1\le j\le p}\Big(n^{-1}r^2\mathbf{A}_j^T\mathbf{X}^T\mathbf{X}\mathbf{A}_j+\big((2r-r^2\widehat{\mathbf{\Omega}}_{j,j})^{1/2}-(2r-r^2\mathbf{\Omega}_{j,j})^{1/2}\big)^2\Big), \tag{A.81}$$

where $\mathbf{A}=\widehat{\mathbf{\Omega}}-\mathbf{\Omega}$ and $\mathbf{B}=\widehat{\mathbf{D}}-\mathbf{D}$, and $\mathbf{A}_j$ and $\mathbf{B}_j$ stand for the jth columns of $\mathbf{A}$ and $\mathbf{B}$, respectively. Their proofs are postponed to the end of the proof. In what follows, we will use subscript j to denote the jth column of a generic matrix.

Next we will show that the upper bound in (A.80) can be bounded from above by the lower bound in (A.81) up to a multiplicative constant. Define the eigen-decompositions $2rI_p-r^2\widehat{\mathbf{\Omega}}=\widehat{P}^T\widehat{\Lambda}\widehat{P}$ and $2rI_p-r^2\mathbf{\Omega}=P^T\Lambda P$, where $\widehat{\Lambda}$ and Λ are diagonal matrices with positive eigenvalues, and $\widehat{P}$ and P are the corresponding eigenvector matrices. By definition, we have $\widehat{\mathbf{D}}=\widehat{P}^T\widehat{\Lambda}^{1/2}\widehat{P}$ and $\mathbf{D}=P^T\Lambda^{1/2}P$. For the second term in the upper bound in (A.80), it holds that

$$\begin{aligned}\|\mathbf{B}_j\|_2^2&=\|\widehat{\mathbf{D}}_j-\mathbf{D}_j\|_2^2\\&=\widehat{P}_j^T\widehat{\Lambda}\widehat{P}_j+P_j^T\Lambda P_j-2\widehat{P}_j^T\widehat{\Lambda}^{1/2}\widehat{P}P^T\Lambda^{1/2}P_j\\&=(2r-r^2\widehat{\mathbf{\Omega}}_{j,j})+(2r-r^2\mathbf{\Omega}_{j,j})-2\widehat{\mathbf{D}}_j^T\mathbf{D}_j.\end{aligned} \tag{A.82}$$

Moreover, the second term in the lower bound presented in (A.81) can be written as

$$\begin{aligned}&\big((2r-r^2\widehat{\mathbf{\Omega}}_{j,j})^{1/2}-(2r-r^2\mathbf{\Omega}_{j,j})^{1/2}\big)^2\\&=\big((\widehat{P}_j^T\widehat{\Lambda}\widehat{P}_j)^{1/2}-(P_j^T\Lambda P_j)^{1/2}\big)^2\\&=(2r-r^2\widehat{\mathbf{\Omega}}_{j,j})+(2r-r^2\mathbf{\Omega}_{j,j})-2(\widehat{P}_j^T\widehat{\Lambda}\widehat{P}_jP_j^T\Lambda P_j)^{1/2}\\&=(2r-r^2\widehat{\mathbf{\Omega}}_{j,j})+(2r-r^2\mathbf{\Omega}_{j,j})-2\|\widehat{\mathbf{D}}_j\|_2\|\mathbf{D}_j\|_2.\end{aligned} \tag{A.83}$$

Therefore, under the assumption that $\|\widehat{\mathbf{D}}_j\|_2\|\mathbf{D}_j\|_2-\widehat{\mathbf{D}}_j^T\mathbf{D}_j\le C\|\widehat{\mathbf{D}}_j-\mathbf{D}_j\|_2^2$ for a constant $C<1/2$, we have

$$\|\mathbf{B}_j\|_2^2\le\big((2r-r^2\widehat{\mathbf{\Omega}}_{j,j})^{1/2}-(2r-r^2\mathbf{\Omega}_{j,j})^{1/2}\big)^2+2C\|\mathbf{B}_j\|_2^2$$

and hence

$$\|\mathbf{B}_j\|_2^2\le\frac{1}{1-2C}\big((2r-r^2\widehat{\mathbf{\Omega}}_{j,j})^{1/2}-(2r-r^2\mathbf{\Omega}_{j,j})^{1/2}\big)^2.$$

This combined with (A.80) and (A.81) proves the desired result in the lemma

$$\mathbb{E}[\|\widehat{\mathbf{X}}-\widetilde{\mathbf{X}}\|_{1,2}^2|\mathbf{X}]\le\frac{2}{1-2C}\big(1+\sqrt{2n^{-1}}\big)(r^2\vee 1)\mathbb{W}_{1,2}(\widehat{\mu}^n,\widetilde{\mu}^n). \tag{A.84}$$

It remains to prove (A.80) and (A.81). We first prove (A.80). Recall our construction of coupling in (37) and (38) that

$$\widehat{\mathbf{X}}=\mathbf{X}(I_p-r\widehat{\mathbf{\Omega}})+\mathbf{Z}\widehat{\mathbf{D}},$$

$$\widetilde{\mathbf{X}}=\mathbf{X}(I_p-r\mathbf{\Omega})+\mathbf{Z}\mathbf{D},$$

where $\mathbf{Z} = (Z_{i,j}) \in \mathbb{R}^{n\times p}$ is independent of $(\mathbf{X}, \mathbf{y})$ and consists of i.i.d. standard normal entries $Z_{i,j} \overset{d}{\sim} N(0,1)$. It immediately follows that

$$\widehat{\mathbf{X}} - \widetilde{\mathbf{X}} = -r\mathbf{X}\mathbf{A} + \mathbf{Z}\mathbf{B}.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\|_{1,2}^2|\mathbf{X}] &= \mathbb{E}\Big[\max_{1\le j\le p} n^{-1}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2^2\Big|\mathbf{X}\Big] \\
&= \mathbb{E}\Big[\max_{1\le j\le p} n^{-1}\|r\mathbf{X}\mathbf{A}_j + \mathbf{Z}\mathbf{B}_j\|_2^2\Big|\mathbf{X}\Big] \\
&\le 2\mathbb{E}\Big[\max_{1\le j\le p}\big(r^2 n^{-1}\mathbf{A}_j^T\mathbf{X}^T\mathbf{X}\mathbf{A}_j + n^{-1}\mathbf{B}_j^T\mathbf{Z}^T\mathbf{Z}\mathbf{B}_j\big)\Big|\mathbf{X}\Big] \\
&\le 2\max_{1\le j\le p}(r^2 n^{-1}\mathbf{A}_j^T\mathbf{X}^T\mathbf{X}\mathbf{A}_j + \|\mathbf{B}_j\|_2^2) \\
&\quad + 2\mathbb{E}\Big[\max_{1\le j\le p}\big|n^{-1}\mathbf{B}_j^T\mathbf{Z}^T\mathbf{Z}\mathbf{B}_j - \|\mathbf{B}_j\|_2^2\big|\Big|\mathbf{X}\Big],
\end{aligned} \tag{A.85}$$

where the second last inequality follows from the Cauchy–Schwarz inequality. To deal with the second term $\mathbb{E}\Big[\max_{1\le j\le p}\big|n^{-1}\mathbf{B}_j^T\mathbf{Z}^T\mathbf{Z}\mathbf{B}_j - \|\mathbf{B}_j\|_2^2\big|\Big|\mathbf{X}\Big]$ in the above upper bound, a key observation is that $\mathbf{Z}\mathbf{B}_j \overset{d}{=} (\widetilde{Z}_1\|\mathbf{B}_j\|_2, \cdots, \widetilde{Z}_n\|\mathbf{B}_j\|_2)$, where $\{\widetilde{Z}_i\}$ are i.i.d. standard normal random variables and are independent of all other variables. Hence, it can be obtained that

$$\begin{aligned}
\mathbb{E}\Big[\max_{1\le j\le p}\big|n^{-1}\mathbf{B}_j^T\mathbf{Z}^T\mathbf{Z}\mathbf{B}_j - \|\mathbf{B}_j\|_2^2\big|\Big|\mathbf{X}\Big] &= \mathbb{E}\Big[\max_{1\le j\le p}\|\mathbf{B}_j\|_2^2\Big|n^{-1}\sum_{i=1}^n(\widetilde{Z}_i^2 - 1)\Big|\Big|\mathbf{X}\Big] \\
&= \max_{1\le j\le p}\|\mathbf{B}_j\|_2^2\,\mathbb{E}\Big[\Big|n^{-1}\sum_{i=1}^n(\widetilde{Z}_i^2 - 1)\Big|\Big] \\
&\le \max_{1\le j\le p}\|\mathbf{B}_j\|_2^2\left(\mathbb{E}\Big[\Big|n^{-1}\sum_{i=1}^n(\widetilde{Z}_i^2 - 1)\Big|^2\Big]\right)^{1/2} \\
&= \sqrt{\frac{2}{n}}\max_{1\le j\le p}\|\mathbf{B}_j\|_2^2,
\end{aligned} \tag{A.86}$$

where we have used the fact that $\mathbb{E}[(\widetilde{Z}_i^2 - 1)^2] = 2$. Combining (A.85) and (A.86) yields the desired result (A.80).

Now we proceed to prove the lower bound in (A.81). Note that by Jensen's inequality,

$$\begin{aligned}
\mathbb{W}_{1,2}^2(\widehat{\mu}^n, \widetilde{\mu}^n) &= \inf_{\gamma\in\Gamma(\widehat{\mu}^n,\widetilde{\mu}^n)}\mathbb{E}_{(\mathrm{vec}(\widehat{\mathbf{X}}),\mathrm{vec}(\widetilde{\mathbf{X}}))\overset{d}{\sim}\gamma}\Big(\max_{1\le j\le p} n^{-1}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2^2\Big) \\
&\ge \inf_{\gamma\in\Gamma(\widehat{\mu}^n,\widetilde{\mu}^n)}\max_{1\le j\le p}\mathbb{E}_{(\mathrm{vec}(\widehat{\mathbf{X}}),\mathrm{vec}(\widetilde{\mathbf{X}}))\overset{d}{\sim}\gamma}\big(n^{-1}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2^2\big).
\end{aligned}$$

Observe that given $\mathbf{X}$, we have $\widehat{\mathbf{X}}_j \overset{d}{\sim} \widehat{\nu}_j^n$ and $\widetilde{\mathbf{X}}_j \overset{d}{\sim} \widetilde{\nu}_j^n$, where $\widehat{\nu}_j^n$ is the Gaussian distribution $N(\mathbf{X}(I_p - r\widehat{\mathbf{\Omega}})_j, (2r - r^2\widehat{\mathbf{\Omega}}_{j,j})I_n)$ and $\widetilde{\nu}_j^n$ is the Gaussian distribution $N(\mathbf{X}(I_p - r\mathbf{\Omega})_j, (2r - r^2\mathbf{\Omega}_{j,j})I_n)$. Given $\mathbf{X}$, let $\Gamma(\widehat{\nu}_j^n, \widetilde{\nu}_j^n)$ be the set of all couplings of $\widehat{\nu}_j^n$ and $\widetilde{\nu}_j^n$. Note that if $(\mathrm{vec}(\widehat{\mathbf{X}}), \mathrm{vec}(\widetilde{\mathbf{X}})) \overset{d}{\sim} \gamma$ for some $\gamma \in \Gamma(\widehat{\mu}^n, \widetilde{\mu}^n)$, then it must hold that $(\widehat{\mathbf{X}}_j, \widetilde{\mathbf{X}}_j) \overset{d}{\sim} \gamma_j$ for some

$\gamma_j \in \Gamma(\widehat{\nu}_j^n, \widetilde{\nu}_j^n)$. Therefore, we can obtain that

$$\begin{aligned}\mathbb{W}_{1,2}^2(\widehat{\mu}^n, \widetilde{\mu}^n) &\geq \inf_{\gamma \in \Gamma(\widehat{\mu}^n, \widetilde{\mu}^n)} \max_{1\leq j\leq p} \inf_{\gamma_j \in \Gamma(\widehat{\nu}_j^n, \widetilde{\nu}_j^n)} \mathbb{E}_{(\widehat{\mathbf{X}}_j, \widetilde{\mathbf{X}}_j) \overset{d}{\sim} \gamma_j} \big(n^{-1}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2^2\big) \\ &= \max_{1\leq j\leq p} \inf_{\gamma_j \in \Gamma(\widehat{\nu}_j^n, \widetilde{\nu}_j^n)} \mathbb{E}_{(\widehat{\mathbf{X}}_j, \widetilde{\mathbf{X}}_j) \overset{d}{\sim} \gamma_j} \big(n^{-1}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2^2\big) \\ &= \max_{1\leq j\leq p} n^{-1}\mathbb{W}_2^2(\widehat{\nu}_j^n, \widetilde{\nu}_j^n),\end{aligned} \tag{A.87}$$

where $\mathbb{W}_2^2(\widehat{\nu}_j^n, \widetilde{\nu}_j^n)$ is the squared 2-Wasserstein distance between $\widehat{\nu}_j^n$ and $\widetilde{\nu}_j^n$. By the well-known result for the 2-Wasserstein distance for Gaussian measures (Givens and Shortt (1984)), we have

$$n^{-1}\mathbb{W}_2^2(\widehat{\nu}_j^n, \widetilde{\nu}_j^n) = n^{-1}\|r\mathbf{X}\mathbf{A}_j\|_2^2 + \big((2r - r^2\widehat{\mathbf{\Omega}}_{j,j})^{1/2} - (2r - r^2\mathbf{\Omega}_{j,j})^{1/2}\big)^2. \tag{A.88}$$

Plugging (A.88) into (A.87) derives (A.81). This completes the proof of Lemma 1.

B.2. Proof of Lemma 2. Let $g_j(\cdot|\mathbf{x}_{-j})$ be the conditional density function of $X_j|X_{-j} = \mathbf{x}_{-j}$ for $X = (X_1, \cdots, X_p)^T \overset{d}{\sim} t_\nu(\mathbf{0}, I_p)$ and $h_j(\cdot|\mathbf{x}_{-j})$ the conditional density function of $\widehat{X}_j|\widehat{X}_{-j} = \mathbf{x}_{-j}$ for $\widehat{X} = (\widehat{X}_1, \cdots, \widehat{X}_p)^T \overset{d}{\sim} N(\mathbf{0}, \frac{\nu}{\nu-2} I_p)$. Following the definition in Barber, Candès and Samworth (2020), we define

$$\widehat{KL}_j := \sum_{i=1}^n \log\left(\frac{g_j(\mathbf{X}_{i,j}|\mathbf{X}_{i,-j}) h_j(\widehat{\mathbf{X}}_{i,j}|\mathbf{X}_{i,j})}{h_j(\mathbf{X}_{i,j}|\mathbf{X}_{i,-j}) g_j(\widehat{\mathbf{X}}_{i,j}|\mathbf{X}_{i,-j})}\right), \tag{A.89}$$

where $\mathbf{X} = (\mathbf{X}_{i,j}) \in \mathbb{R}^{n\times p}$ consists of i.i.d. rows sampled from $t_\nu(\mathbf{0}, I_p)$ and $\widehat{\mathbf{X}} = (\widehat{\mathbf{X}}_{i,j}) \in \mathbb{R}^{n\times p}$ consists of i.i.d. rows sampled from $N(\mathbf{0}, I_p)$. Note that Theorem 1 in Barber, Candès and Samworth (2020) states that

$$\mathrm{FDR} \leq \min_{\varepsilon \geq 0}\left\{qe^{\varepsilon} + \mathbb{P}\left(\max_{j\in\mathcal{H}_0} \widehat{KL}_j > \varepsilon\right)\right\}. \tag{A.90}$$

We claim that if $\frac{np}{\nu(\nu+p)} \geq C$ for some constant $C > 0$, there exists some positive constant α such that

$$\mathbb{P}\left(\widehat{KL}_j \geq C/4\right) \geq \alpha. \tag{A.91}$$

Then it holds that for $0 < \varepsilon < C/4$,

$$\mathbb{P}\big(\max_{1\leq j\leq p} \widehat{KL}_j \geq \varepsilon\big) \geq \alpha,$$

and thus we cannot obtain the desired asymptotic FDR control $\limsup_{(n,p)} \mathrm{FDR} \leq q$ via applying Theorem 1 in Barber, Candès and Samworth (2020). By contradiction, to allow $\mathbb{P}\big(\max_{1\leq j\leq p} \widehat{KL}_j \geq \varepsilon\big) \to 0$, we must have that $\frac{np}{\nu(\nu+p)} \to 0$, which is equivalent to $\nu^2 \gg n\min(n,p)$. Hence, Lemma 2 is proved. Now it remains to establish (A.91).

Proof of (A.91). Note that Ding (2016) showed that the conditional density $g_j(\mathbf{x}_j|\mathbf{x}_{-j})$ of the multivariate t-distribution satisfies that

$$g_j(\mathbf{X}_{i,j}|\mathbf{X}_{i,-j}) \propto \left(1 + \frac{\mathbf{X}_{i,j}^2}{\nu + \|\mathbf{X}_{i,-j}\|_2^2}\right)^{-(\nu+p)/2}.$$

It is easy to see that the conditional density $h_j(\mathbf{X}_{i,j}|\mathbf{X}_{i,-j})$ of the standard normal distribution satisfies that

$$h_j(\mathbf{X}_{i,j}|\mathbf{X}_{i,-j}) \propto \exp\{-\mathbf{X}_{i,j}^2(\nu-2)/2\nu\}.$$

Plugging the two expressions above into (A.89) yields that

$$\begin{aligned}\widehat{KL}_j = \sum_{i=1}^{n}\Bigg[&\frac{\mathbf{X}_{i,j}^2(\nu-2)}{2\nu} - \frac{\nu+p}{2}\log\left(1+\frac{\mathbf{X}_{i,j}^2}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}\right)\\ &-\left(\frac{\widehat{\mathbf{X}}_{i,j}^2(\nu-2)}{2\nu} - \frac{\nu+p}{2}\log\left(1+\frac{\widehat{\mathbf{X}}_{i,j}^2}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}\right)\right)\Bigg].\end{aligned}$$

Applying the basic inequality that $|\log(1+x)-(x-x^2/2)| \le x^3$ for each $x>0$, we can obtain that

$$\widehat{KL}_j = R_{1,j} + R_{2,j} + O(R_{3,j}), \tag{A.92}$$

where

$$R_{1,j} = \sum_{i=1}^{n}\left[\frac{\mathbf{X}_{i,j}^2(\nu+p)}{2(\nu+\|\mathbf{X}_{i,-j}\|_2^2)}\left(\frac{\nu+\|\mathbf{X}_{i,-j}\|_2^2}{\nu+p}\cdot\frac{\nu-2}{\nu}-1\right) - \frac{\widehat{\mathbf{X}}_{i,j}^2(\nu-2)}{2\nu}\left(1-\frac{\nu+p}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}\right)\right], \tag{A.93}$$

$$R_{2,j} = \sum_{i=1}^{n}\frac{\nu+p}{4}\left(\frac{\widehat{\mathbf{X}}_{i,j}^4}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^2} - \frac{\mathbf{X}_{i,j}^4}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^2}\right), \tag{A.94}$$

$$R_{3,j} = \sum_{i=1}^{n}\frac{\nu+p}{2}\left(\frac{\widehat{\mathbf{X}}_{i,j}^6}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^3} + \frac{\mathbf{X}_{i,j}^6}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^3}\right). \tag{A.95}$$

We now calculate the mean and variance of $\widehat{KL}_j$ separately. Observe that $\sqrt{\frac{\nu-2}{\nu}}\widehat{\mathbf{X}}_{i,j} \overset{d}{\sim} N(0,1)$, $(p-1)^{-1}\|\mathbf{X}_{i,-j}\|_2^2 \overset{d}{\sim} F_{p-1,\nu}$, $\mathbf{X}_{i,-j} \perp\!\!\!\perp \sqrt{\frac{\nu+p}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}}\mathbf{X}_{i,j}$, and

$$\sqrt{\frac{\nu+p-1}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}}\mathbf{X}_{i,j} \overset{d}{\sim} t_{\nu+p-1}$$

as shown in Ding (2016). Using the properties of the multivariate t-distribution and F-distribution, some straightforward calculations show that

$$\begin{aligned}\mathbb{E}(R_{1,j}) &= \frac{n}{2}\left[\frac{\nu+p}{\nu+p-3}\left(\frac{\nu(\nu+p-3)}{(\nu-2)(\nu+p)}\cdot\frac{\nu-2}{\nu}-1\right) - \left(1-\frac{(\nu+2)(\nu+p)}{\nu(\nu+p-1)}\right)\right]\\ &= n\left(\frac{2p}{\nu(\nu+p)}+O(\nu^{-2})\right),\end{aligned} \tag{A.96}$$

$$\begin{aligned}\mathbb{E}(R_{2,j}) &= \frac{3n(\nu+p)}{4}\left[\frac{1}{(\nu+p-3)(\nu+p-5)} - \frac{\nu+2}{\nu(\nu+p-1)(\nu+p+1)}\right]\\ &= O(\frac{n}{\nu(\nu+p)}),\end{aligned} \tag{A.97}$$

and

$$\mathbb{E}(R_{3,j}) \le Cn(\nu+p)^{-2}. \tag{A.98}$$

Combining (A.96)–(A.98) yields that when ν and p are large,

$$\mathbb{E}(\widehat{KL}_j) = \frac{np}{\nu(\nu+p)} + O(n\nu^{-2}) \ge \frac{np}{2\nu(\nu+p)}. \tag{A.99}$$

Next we analyze the variance of $\widehat{KL}_j$. Notice that

$$\begin{aligned}
\mathrm{Var}(\widehat{KL}_j) &= \mathbb{E}\big((\widehat{KL}_j - \mathbb{E}\widehat{KL}_j)^2\big) \\
&\le C\sum_{i=1}^{n}\mathbb{E}\Bigg\{\Bigg[\frac{\mathbf{X}_{i,j}^2(\nu+p)}{2(\nu+\|\mathbf{X}_{i,-j}\|_2^2)}\Bigg(\frac{\nu+\|\mathbf{X}_{i,-j}\|_2^2}{\nu+p}\cdot\frac{\nu-2}{\nu}-1\Bigg) \\
&\qquad - \frac{\widehat{\mathbf{X}}_{i,j}^2(\nu-2)}{2\nu}\Bigg(1-\frac{\nu+p}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}\Bigg)\Bigg]^2\Bigg\} \\
&\quad + C\sum_{i=1}^{n}\mathbb{E}\Bigg[\frac{(\nu+p)^2}{16}\Bigg(\frac{\widehat{\mathbf{X}}_{i,j}^4}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^2}-\frac{\mathbf{X}_{i,j}^4}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^2}\Bigg)^2\Bigg] \\
&\le \frac{Cnp}{\nu(\nu+p)},
\end{aligned} \tag{A.100}$$

where in the last step above, we have used the facts that

$$\begin{aligned}
&\mathbb{E}\Bigg(\frac{\mathbf{X}_{i,j}^4(\nu+p)^2}{(\nu+\|\mathbf{X}_{i,-j}\|_2^2)^2}\Bigg) \le C, \\
&\mathbb{E}\Bigg[\Bigg(\frac{\nu+\|\mathbf{X}_{i,-j}\|_2^2}{\nu+p}\cdot\frac{\nu-2}{\nu}-1\Bigg)^2\Bigg] = \frac{2p}{\nu(\nu+p)}+O(\nu^{-2}), \\
&\mathbb{E}\Bigg[\Bigg(1-\frac{\nu+p}{\nu+\|\mathbf{X}_{i,-j}\|_2^2}\Bigg)^2\Bigg] = \frac{2p}{\nu(\nu+p)}+O(\nu^{-2}).
\end{aligned}$$

In view of the results on the mean and variance of $\widehat{KL}_j$ shown in (A.98) and (A.99) above, we see that if $\frac{np}{\nu(\nu+p)} \ge C$ for some constant $C>0$,

$$\mathbb{E}(\widehat{KL}_j) \ge \frac{np}{2\nu(\nu+p)} \ge C/2.$$

Therefore, we can obtain through the one-sided Markov inequality that for a small constant $\alpha>0$ (noting that $\mathbb{E}(\widehat{KL}_j) > 2\alpha\sqrt{\mathrm{Var}(\widehat{KL}_j)}$ if α is small),

$$\begin{aligned}
\mathbb{P}(\widehat{KL}_j \ge C/4) &\ge \mathbb{P}(\widehat{KL}_j \ge \mathbb{E}(\widehat{KL}_j)/2) \\
&\ge \mathbb{P}\Big(\widehat{KL}_j \ge \mathbb{E}(\widehat{KL}_j) - \alpha\sqrt{\mathrm{Var}(\widehat{KL}_j)}\Big) \\
&\ge 1 - \frac{\mathrm{Var}(\widehat{KL}_j)}{\mathrm{Var}(\widehat{KL}_j)+\alpha^2\,\mathrm{Var}(\widehat{KL}_j)} \\
&= \frac{\alpha^2}{1+\alpha^2},
\end{aligned} \tag{A.101}$$

which establishes (A.91). This completes the proof of Lemma 2.

B.3. Proof of Lemma 3. Recall that $G(t) = p_0^{-1}\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$ and $G(t)$ is a decreasing, continuous function. The main idea of the proof is to divide the continuous interval $(0, G^{-1}(\frac{c_1 q a_n}{p})]$ into a diverging number of smaller intervals with end points $\{t_i\}_{i=0}^{l_n}$ such that $t_0 \geq t_1 \geq \cdots \geq t_{l_n}$ and

$$|G(t_i)/G(t_{i+1}) - 1| \to 0$$

uniformly for $0 \leq i \leq l_n$ as $l_n \to \infty$. Then the supreme over the continuous interval $(0, G^{-1}(\frac{c_1 q a_n}{p})]$ can be reduced to the supreme over the set of discrete points $\{t_i\}_{i=0}^{l_n}$ and hence, we can apply the union bound to establish the desired result. Similar arguments have also been used in Liu (2013), Cai and Liu (2016), and Guo et al. (2022). We detail only the proof of (A.1) here since (A.2) can be shown in a similar fashion.

We start with defining a sequence $0 \leq z_0 < z_1 < \cdots < z_{l_n} = 1$ and

$$t_i = G^{-1}(z_i),$$

where $z_0 = \frac{c_1 q a_n}{p}$, $z_i = \frac{c_1 q a_n}{p} + \frac{h_n e^{i^\gamma}}{p}$, and $l_n = [\log((p - c_1 q a_n)/h_n)]^{1/\gamma}$ with $0 < \gamma < 1$ and sequence $h_n \to \infty$ satisfying that $h_n/a_n \to 0$. As long as $m_n/a_n = o(1)$, we can choose

$$h_n = \frac{a_n}{(a_n/m_n)^\eta}$$

for some $\eta \in (0, 1)$. Then an application of similar technical analysis as in Guo et al. (2022) shows that as $a_n \to \infty$,

$$\sup_{0\leq i\leq l_n} |G(t_i)/G(t_{i+1}) - 1| \to 0. \tag{A.102}$$

For $t \in (0, G(\frac{c_1 q a_n}{p})]$, there exists some $0 \leq i \leq l_n - 1$ such that $t \in [t_{i+1}, t_i]$. It follows from the monotonicity of $\mathbb{P}(\widetilde{W}_j \geq t)$ and $\mathbb{1}(\widehat{W}_j \geq t)$ that

$$\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t)}{p_0 G(t)} - 1\right| \leq \max\left\{\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_{i+1})}{p_0 G(t_i)} - 1\right|, \left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_i)}{p_0 G(t_{i+1})} - 1\right|\right\}.$$

The two terms within the brackets on the right-hand side of the expression above can be bounded similarly and we will provide only the details on how to bound the first term for simplicity.

With the aid of the fact that $|xy - 1| \leq |x-1||y-1| + |x-1| + |y-1|$ for all $x, y \in \mathbb{R}$, we can deduce that

$$\begin{aligned}
\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_{i+1})}{p_0 G(t_i)} - 1\right| &\leq \left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_{i+1})}{p_0 G(t_{i+1})} - 1\right| \cdot \sup_{0\leq i\leq l_n}\left|\frac{G(t_i)}{G(t_{i+1})} - 1\right| \\
&\quad + \left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_{i+1})}{p_0 G(t_{i+1})} - 1\right| + \sup_{0\leq i\leq l_n}\left|\frac{G(t_i)}{G(t_{i+1})} - 1\right| \\
&\leq \left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j \geq t_{i+1})}{p_0 G(t_{i+1})} - 1\right| \cdot (1 + o(1)) + \sup_{0\leq i\leq l_n}\left|\frac{G(t_i)}{G(t_{i+1})} - 1\right|,
\end{aligned}$$

where the last step above is because of (A.102) and the $o(1)$ term is uniformly over all i. Combining the above two results and applying (A.102) again lead to

$$\begin{aligned}&\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq t)}{p_0G(t)}-1\right|\\&\leq\max\left\{\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq t_{i+1})}{p_0G(t_{i+1})}-1\right|,\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq t_i)}{p_0G(t_i)}-1\right|\right\}\\&\quad\times\big(1+o(1)\big)+o(1).\end{aligned}\tag{A.103}$$

Thus, to prove the desired result, it is sufficient to show that

$$D_n:=\sup_{0\leq i\leq l_n}\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widehat{W}_j\geq t_i)}{p_0G(t_i)}-1\right|=o_p(1).\tag{A.104}$$

We now proceed with establishing (A.104). Let us define an event

$$\mathcal{B}_3=\{\max_{1\leq j\leq p}|\widehat{W}_j-\widetilde{W}_j|\leq b_n\}.$$

From Condition 1, it holds that $\mathbb{P}(\mathcal{B}_3^c)\to 0$. Note that for any two events A and B, we have that $\mathbb{P}(A)\leq\mathbb{P}(A\cap B)+P(B^c)$. Repeatedly using such inequality, the union bound, and the property that $\mathbb{P}(\mathcal{B}_3^c)\to 0$, we can deduce that for each $\epsilon>0$,

$$\begin{aligned}\mathbb{P}(D_n\geq\epsilon)&\leq\sum_{i=0}^{l_n}\mathbb{P}\Big(\Big|\frac{\sum_{j\in\mathcal{H}_0}\{\mathbb{1}(\widehat{W}_j\geq t_i)-\mathbb{P}(\widetilde{W}_i\geq t_i)\}}{p_0G(t_i)}\Big|\geq\epsilon,\mathcal{B}_3\Big)+\mathbb{P}(\mathcal{B}_3^c)\\&\leq\sum_{i=0}^{l_n}\mathbb{P}\Big(\Big|\frac{\sum_{j\in\mathcal{H}_0}\{\mathbb{1}(\widetilde{W}_j\geq t_i)-\mathbb{P}(\widetilde{W}_i\geq t_i)\}}{p_0G(t_i)}\Big|\geq\epsilon/2\Big)\\&\quad+\sum_{i=0}^{l_n}\mathbb{P}\Big(\Big|\frac{\sum_{j\in\mathcal{H}_0}[\mathbb{1}(\widehat{W}_j\geq t_i)-\mathbb{1}(\widetilde{W}_i\geq t_i)]}{p_0G(t_i)}\Big|\geq\epsilon/2,\mathcal{B}_3\Big)+o(1)\\&\leq\sum_{i=0}^{l_n}\frac{4\mathbb{E}\big[\{\sum_{j\in\mathcal{H}_0}[\mathbb{1}(\widetilde{W}_j\geq t_i)-\mathbb{P}(\widetilde{W}_i\geq t_i)]\}^2\big]}{\epsilon^2p_0^2G^2(t_i)}\\&\quad+\sum_{i=0}^{l_n}\frac{2\sum_{j\in\mathcal{H}_0}\mathbb{P}\big(t_i-b_n\leq\widetilde{W}_j\leq t_i+b_n\big)}{\epsilon p_0G(t_i)}+o(1),\end{aligned}\tag{A.105}$$

where the last step above is due to the Markov inequality and the fact that $|\mathbb{1}(\widehat{W}_j\geq t_i)-\mathbb{1}(\widetilde{W}_i\geq t_i)|\leq\mathbb{1}(t_i-b_n\leq\widetilde{W}_j\leq t_i+b_n)$ on event $\mathcal{B}_3$.

We next bound the first two terms on the very right-hand side of (A.105) above. For the first term, under Condition 4 for the weak dependence between $\{W_j\}$, we have that

$$\begin{aligned}&\sum_{i=0}^{l_n}\frac{4\mathbb{E}\big[\{\sum_{j\in\mathcal{H}_0}[\mathbb{1}(\widetilde{W}_j\geq t_i)-\mathbb{P}(\widetilde{W}_i\geq t_i)]\}^2\big]}{\epsilon^2p_0^2G^2(t_i)}\\&\leq C\sum_{i=0}^{l_n}\frac{m_np_0G(t_i)+o\big((\log p)^{-1/\gamma}[p_0G(t_i)]^2\big)}{\epsilon^2p_0^2G^2(t_i)}\\&=C\epsilon^{-2}m_n\sum_{i=0}^{l_n}\frac{1}{p_0G(t_i)}+C\epsilon^{-2}o\big(l_n(\log p)^{-1/\gamma}\big).\end{aligned}\tag{A.106}$$

Moreover, it holds that

$$\sum_{i=0}^{l_n} \frac{1}{p_0 G(t_i)} = p_0^{-1} \sum_{i=0}^{l_n} \frac{1}{z_i} = \frac{p}{p_0} \sum_{i=0}^{l_n} \frac{1}{c_1 q a_n + h_n e^{i\gamma}} \leq C h_n^{-1}, \tag{A.107}$$

where the last inequality above is related to the proof of Theorem 3 in Guo et al. (2022).

In light of the definition of h_n and the assumption of $m_n/a_n \to 0$, we have that

$$m_n/h_n = (m_n/a_n)^{1-\eta} \to 0.$$

Therefore, combining (A.106)–(A.107) and the fact that

$$l_n = [\log((p - c_1 q a_n)/h_n)]^{1/\gamma} \leq (\log p)^{1/\gamma}$$

shows that the first term for the bound in (A.105) tends to zero as $n \to \infty$. Moreover, since $l_n \leq (\log p)^{1/\gamma}$, the second term on the very right-hand side of (A.105) above is bounded by

$$\frac{2}{\epsilon}(\log p)^{1/\gamma} \sup_{t \in (0, G^{-1}(\frac{c_1 q a_n}{p})]} \frac{G(t - b_n) - G(t + b_n)}{G(t)},$$

which converges to zero as $n \to \infty$ under Condition 5. Finally, we can obtain that for each $\epsilon > 0$,

$$\mathbb{P}(D_n > \epsilon) \to 0, \tag{A.108}$$

which establishes the desired result in (A.1). This concludes the proof of Lemma 3.

B.4. Proof of Lemma 4. We will show that with asymptotic probability one, it holds that for some $0 < c_1 < 1$,

$$1 + \sum_{j=1}^{p} \mathbb{1}\big(\widehat{W}_j < -G^{-1}(\frac{c_1 q a_n}{p})\big) \leq q a_n \leq q \sum_{j=1}^{p} \mathbb{1}\big(\widehat{W}_j \geq G^{-1}(\frac{c_1 q a_n}{p})\big). \tag{A.109}$$

Then from the definition of T, we can obtain the desired result of the lemma. We aim to establish (A.109). The main idea of the proof is to prove that the population counterpart of (A.109) holds. Then with an application of Lemma 3 to both left- and right-hand sides of (A.109), we can connect it to the population counterpart and thus prove that (A.109) holds with asymptotic probability one.

First, it follows from the union bound and the fact that $\mathbb{P}(A) \leq \mathbb{P}(A \cap B) + \mathbb{P}(B^c)$ for any two events A and B that under Conditions 1–3,

$$\begin{aligned}
&\mathbb{P}(\widehat{W}_j < 3\delta_n \text{ for some } j \in \mathscr{A}_n) \\
&\leq \mathbb{P}(\widehat{W}_j < 3\delta_n \text{ for some } j \in \mathscr{A}_n, \max_{1 \leq j \leq p} |\widehat{W}_j - \widetilde{W}_j| < b_n) + \mathbb{P}(\max_{1 \leq j \leq p} |\widehat{W}_j - \widetilde{W}_j| \geq b_n) \\
&\leq \mathbb{P}(\widetilde{W}_j < 3\delta_n + b_n \text{ for some } j \in \mathscr{A}_n) + \mathbb{P}(\max_{1 \leq j \leq p} |\widehat{W}_j - \widetilde{W}_j| \geq b_n) \\
&\leq \sum_{j \in \mathscr{A}_n} \mathbb{P}(\widetilde{W}_j - w_j < 3\delta_n + b_n - w_j) + o(1) \\
&\leq \sum_{j \in \mathscr{A}_n} \mathbb{P}(|\widetilde{W}_j - w_j| > \delta_n) + o(1) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(|\widetilde{W}_j - w_j| > \delta_n) + o(1) \to 0.
\end{aligned}$$

Then we have

$$\mathbb{P}(\cap_{j\in\mathscr{A}_n}\{\widehat{W}_j \geq 3\delta_n\}) \to 1$$

and thus with asymptotic probability one,

$$\sum_{j=1}^{p} \mathbb{1}(\widehat{W}_j \geq 3\delta_n) \geq a_n, \tag{A.110}$$

where $a_n = |\mathscr{A}_n|$.

In addition, since $w_j > -\delta_n$ for $1 \leq j \leq p$ by assumption, we can deduce that

$$\begin{aligned}
\sum_{j=1}^{p} \mathbb{P}(\widehat{W}_j < -3\delta_n) &\leq \sum_{j=1}^{p} \mathbb{P}(\widehat{W}_j < -3\delta_n, \max_{1\leq j\leq p} |\widehat{W}_j - \widetilde{W}_j| < b_n) \\
&\quad + \mathbb{P}(\max_{1\leq j\leq p} |\widehat{W}_j - \widetilde{W}_j| \geq b_n) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(\widetilde{W}_j < -3\delta_n + b_n) + o(1) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(\widetilde{W}_j - w_j \leq -3\delta_n + b_n - w_j) + o(1) \\
&\leq \sum_{j=1}^{p} \mathbb{P}(|\widetilde{W}_j - w_j| > \delta_n) + o(1) \to 0,
\end{aligned} \tag{A.111}$$

which yields $\sum_{j=1}^{p} \mathbb{P}(\widehat{W}_j < -3\delta_n) \to 0$. Using similar arguments as for (A.111), it holds that

$$\sum_{j=1}^{p} \mathbb{P}(\widetilde{W}_j \leq -3\delta_n) \to 0.$$

Then we can obtain that

$$\begin{aligned}
G(3\delta_n) &= p_0^{-1} \sum_{j\in\mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \leq -3\delta_n) \leq p_0^{-1} \sum_{j=1}^{p} \mathbb{P}(\widetilde{W}_j \leq -3\delta_n) \\
&= o(p_0^{-1}).
\end{aligned}$$

Since $a_n \to \infty$, $p_0/p \to 1$, and $G(t)$ is a nonincreasing, continuous function, it follows that $G(3\delta_n) \leq \frac{c_1 q a_n}{p}$ and thus

$$G^{-1}(\frac{c_1 q a_n}{p}) \leq 3\delta_n$$

for some constant $0 < c_1 < 1$ when n is sufficiently large. This together with (A.110) entails that with asymptotic probability one,

$$\sum_{j=1}^{p} \mathbb{1}(\widehat{W}_j \geq G^{-1}(\frac{c_1 q a_n}{p})) \geq a_n.$$

This completes the proof of the second inequality in (A.109).

It remains to establish the first inequality in (A.109). From the definition of $G(t)$ and Lemma 3, it holds that

$$\begin{aligned}\frac{c_1 q a_n}{p} &= p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \le -G^{-1}(\frac{c_1 q a_n}{p})) \\ &= (1 + o_p(1)) \cdot p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{1}\big(\widehat{W}_j < -G^{-1}(\frac{c_1 q a_n}{p})\big).\end{aligned} \tag{A.112}$$

Then for some constant c_2 satisfying $0 < c_1 < c_2 < 1$, we can obtain that with asymptotic probability one,

$$1 + \sum_{j \in \mathcal{H}_0} \mathbb{1}\big(\widehat{W}_j < -G^{-1}(\frac{c_1 q a_n}{p})\big) \le \frac{c_1 q a_n p_0}{p}(1 + o_p(1)) \le c_2 q a_n, \tag{A.113}$$

where we have used the assumption of $p_0/p \to 1$. Further, under (8) in Condition 5, an application of the union bound yields that

(A.114)

$$\begin{aligned}&\mathbb{P}\Big(\sum_{j \in \mathcal{H}_1} \mathbb{1}\big(\widehat{W}_j < -G^{-1}(\frac{c_1 q a_n}{p})\big) \ge (1 - c_2) q a_n \Big) \\ &\le \mathbb{P}\Big(\sum_{j \in \mathcal{H}_1} \mathbb{1}\big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + b_n\big) \ge (1 - c_2) q a_n, \max_{1 \le j \le p} |\widehat{W}_j - \widetilde{W}_j| < b_n \Big) \\ &\quad + o(1) \\ &\le \frac{1}{(1 - c_2) q a_n} \sum_{j \in \mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + b_n\Big) + o(1) \to 0,\end{aligned}$$

which together with (A.113) implies that

$$1 + \sum_{j=1}^{p} \mathbb{1}\big(\widehat{W}_j < -G^{-1}(\frac{c_1 q a_n}{p})\big) \le q a_n \tag{A.115}$$

with asymptotic probability one. This proves the first inequality in (A.109), which completes the proof of Lemma 4.

B.5. Proof of Lemma 5. Recall that the perfect and approximate knockoff statistics based on the marginal correlation are defined as

$$\widetilde{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T \mathbf{y}| - |\widetilde{\mathbf{X}}_j^T \mathbf{y}|) \ \text{ and } \ \widehat{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T \mathbf{y}| - |\widehat{\mathbf{X}}_j^T \mathbf{y}|),$$

respectively. By the triangle inequality, it is easy to see that

$$\max_{1 \le j \le p} |\widehat{W}_j - \widetilde{W}_j| \le \max_{1 \le j \le p} (\sqrt{n}\|\mathbf{y}\|_2)^{-1} |(\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j)^T \mathbf{y}|.$$

Then an application of the Cauchy–Schwarz inequality gives that

$$\max_{1 \le j \le p} |\widehat{W}_j - \widetilde{W}_j| \le (\sqrt{n})^{-1} \max_{1 \le j \le p} \|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2.$$

Thus, the conclusion of Lemma 5 can be derived under Condition 6. This completes the proof of Lemma 5.

B.6. Proof of Lemma 6. From the definitions of $\widetilde{W}_j$ and w_j and the triangle inequality, it holds that

$$\begin{aligned}
&\mathbb{P}(|\widetilde{W}_j - w_j| \geq \delta_n) \\
&\leq \mathbb{P}\bigg((n^{-1}\|\mathbf{y}\|_2^2)^{-1/2}\Big|n^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widetilde{\mathbf{X}}_j^T\mathbf{y}|) - (|\mathbb{E}(X_jY)| - |\mathbb{E}(\widetilde{X}_jY)|)\Big| \geq \delta_n/2\bigg) \\
&\quad + \mathbb{P}\bigg(\Big|(n^{-1}\|\mathbf{y}\|_2^2)^{-1/2} - (\mathbb{E}Y^2)^{-1/2}\Big| \cdot \Big||\mathbb{E}(X_jY)| - |\mathbb{E}(\widetilde{X}_jY)|\Big| \geq \delta_n/2\bigg) \\
&:= P_1 + P_2.
\end{aligned}$$

We will aim to show that for $\delta_n \to 0$,

$$P_1 \leq 4\exp\Big\{-\frac{n\delta_n^2\mathbb{E}Y^2}{256\|X_j\|_{\psi_2}^2\|Y\|_{\psi_2}^2}\Big\} + \exp\Big\{-\frac{n(\mathbb{E}Y^2)^2}{8\mathbb{E}Y^4}\Big\} \tag{A.116}$$

and

$$P_2 \leq 2\exp\Big\{-\frac{n\delta_n^2(\mathbb{E}Y^2)^2}{64|w_j|^2\|Y\|_{\psi_2}^4}\Big\} + \exp\Big\{-\frac{n(\mathbb{E}Y^2)^2}{8\mathbb{E}Y^4}\Big\}. \tag{A.117}$$

Then setting $\delta_n = \sqrt{\frac{\log p}{n}}\max_{1\leq j\leq p}\Big\{\frac{16\sqrt{2}\|X_j\|_{\psi_2}\|Y\|_{\psi_2}}{(\mathbb{E}Y^2)^{1/2}} \vee \frac{8\sqrt{2}|w_j|\|Y\|_{\psi_2}^2}{\mathbb{E}Y^2}\Big\}$, a combination of the above results leads to the desired conclusion of this lemma.

We proceed with proving (A.116). Since $\|\mathbf{y}\|_2^2 = \sum_{i=1}^n y_i^2$ is the sum of i.i.d. random variables, an application of Bernstein's inequality yields that

$$\mathbb{P}(n^{-1}\|\mathbf{y}\|_2^2 \leq \mathbb{E}[Y^2]/2) \leq \exp\Big\{-\frac{n(\mathbb{E}Y^2)^2}{8\mathbb{E}Y^4}\Big\}. \tag{A.118}$$

It follows from the triangle inequality and (A.118) that

$$\begin{aligned}
P_1 &\leq \mathbb{P}\bigg(\Big|n^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widetilde{\mathbf{X}}_j^T\mathbf{y}|) - (|\mathbb{E}(X_jY)| - |\mathbb{E}(\widetilde{X}_jY)|)\Big| \geq \frac{\delta_n(\mathbb{E}Y^2)^{1/2}}{2\sqrt{2}}\bigg) \\
&\quad + \mathbb{P}(n^{1/2}(\|\mathbf{y}\|_2)^{-1} \geq \sqrt{2}(\mathbb{E}[Y^2])^{-1/2}) \\
&\leq \mathbb{P}\bigg(\frac{1}{n}\Big|\sum_{i=1}^n[\mathbf{X}_{i,j}y_i - \mathbb{E}(X_jY)]\Big| \geq \frac{\delta_n(\mathbb{E}Y^2)^{1/2}}{4\sqrt{2}}\bigg) \\
&\quad + \mathbb{P}\bigg(\frac{1}{n}\Big|\sum_{i=1}^n[\widetilde{\mathbf{X}}_{i,j}y_i - \mathbb{E}(\widetilde{X}_jY)]\Big| \geq \frac{\delta_n(\mathbb{E}Y^2)^{1/2}}{4\sqrt{2}}\bigg) \\
&\quad + \exp\Big\{-\frac{n(\mathbb{E}Y^2)^2}{8\mathbb{E}Y^4}\Big\}.
\end{aligned}$$

We next bound the first two terms on the right-hand side of the expression above. Under Condition 7, we see that $\mathbf{X}_{i,j}y_i$ and $\widetilde{\mathbf{X}}_{i,j}y_i$ are both sub-exponential random variables, with sub-exponential norms $\|X_j\|_{\psi_2}\|Y\|_{\psi_2}$ and $\|X_j\|_{\psi_2}\|Y\|_{\psi_2}$, respectively. Then we can obtain through applying Bernstein's inequality for sub-exponential random variables (see, e.g., Corollary 2.8.3 in Vershynin (2018)) that when $\delta_n = o(1)$,

$$\mathbb{P}\bigg(\frac{1}{n}\Big|\sum_{i=1}^n[\mathbf{X}_{i,j}y_i - \mathbb{E}(X_jY)]\Big| \geq \frac{\delta_n(\mathbb{E}Y^2)^{1/2}}{4\sqrt{2}}\bigg) \leq 2\exp\Big\{-\frac{n\delta_n^2\mathbb{E}Y^2}{256\|X_j\|_{\psi_2}^2\|Y\|_{\psi_2}^2}\Big\}$$

and

$$\mathbb{P}\bigg(\frac{1}{n}\Big|\sum_{i=1}^{n}[\widetilde{\mathbf{X}}_{i,j}y_i-\mathbb{E}(\widetilde{X}_jY)]\Big|\geq\frac{\delta_n(\mathbb{E}Y^2)^{1/2}}{4\sqrt{2}}\bigg)\leq 2\exp\Big\{-\frac{n\delta_n^2\mathbb{E}Y^2}{256\|X_j\|_{\psi_2}^2\|Y\|_{\psi_2}^2}\Big\}.$$

Thus, combining the above three inequalities establishes (A.116).

As for term P_2, noting that $w_j=(\mathbb{E}Y^2)^{-1/2}(|\mathbb{E}(X_jY)|-|\mathbb{E}(\widetilde{X}_jY)|)$ and

$$\Big|(n^{-1}\|\mathbf{y}\|_2^2)^{-1/2}-(\mathbb{E}Y^2)^{-1/2}\Big|=\frac{|n^{-1}\|\mathbf{y}\|_2^2-\mathbb{E}Y^2|}{n^{-1/2}\|\mathbf{y}\|_2(\mathbb{E}Y^2)^{1/2}((\mathbb{E}Y^2)^{1/2}+n^{-1/2}\|\mathbf{y}\|_2)},$$

we can deduce that

$$\begin{aligned}P_2&=\mathbb{P}\bigg(|w_j|\frac{|n^{-1}\|\mathbf{y}\|_2^2-\mathbb{E}Y^2|}{n^{-1/2}\|\mathbf{y}\|_2((\mathbb{E}Y^2)^{1/2}+n^{-1/2}\|\mathbf{y}\|_2)}\geq\delta_n/2\bigg)\\&\leq\mathbb{P}\bigg(|w_j|\frac{|n^{-1}\|\mathbf{y}\|_2^2-\mathbb{E}Y^2|}{n^{-1/2}\|\mathbf{y}\|_2(\mathbb{E}Y^2)^{1/2}}\geq\delta_n/2\bigg)\\&=\mathbb{P}\bigg(|n^{-1}\|\mathbf{y}\|_2^2-\mathbb{E}Y^2|\geq\frac{\delta_n\mathbb{E}Y^2}{2\sqrt{2}|w_j|}\bigg)+\mathbb{P}(n^{-1}\|\mathbf{y}\|_2^2\leq\mathbb{E}Y^2/2).\end{aligned}\tag{A.119}$$

The very last term above can be bounded by applying (A.118).

Again we can see that under Condition 7, y_i^2 is a sub-exponential random variable with sub-exponential norm $\|Y\|_{\psi_2}^2$. With the aid of Bernstein's inequality for sub-exponential random variables (Corollary 2.8.3 in Vershynin (2018)), we can obtain that for $\delta_n=o(1)$,

$$\mathbb{P}\bigg(\frac{1}{n}\Big|\sum_{i=1}^{n}[y_i^2-\mathbb{E}(Y^2)]\Big|\geq\frac{\delta_n\mathbb{E}Y^2}{2\sqrt{2}|w_j|}\bigg)\leq 2\exp\Big\{-\frac{n\delta_n^2(\mathbb{E}Y^2)^2}{64|w_j|^2\|Y\|_{\psi_2}^4}\Big\}.$$

Therefore, the bound for term P_2 in (A.117) can be shown. This concludes the proof of Lemma 6.

B.7. Proof of Lemma 7. The main idea of the proof is to apply the law of total variance and decompose the total into two terms by conditioning on $(\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})$, where $\mathbf{X}_{\mathcal{H}_1}=(\mathbf{X}_j)_{j\in\mathcal{H}_1}$ and $\boldsymbol{\varepsilon}=(\varepsilon_1,\cdots,\varepsilon_n)^T$. Specifically, it holds that

$$\begin{aligned}\mathrm{Var}\bigg(\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j\geq t)\bigg)&=\mathbb{E}\bigg\{\mathbb{E}\bigg[\bigg(\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j\geq t)-\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j\geq t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\bigg)^2\bigg|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon}\bigg]\bigg\}\\&\quad+\mathbb{E}\bigg\{\bigg(\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j\geq t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})-\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j\geq t)\bigg)^2\bigg\}\\&:=V_1+V_2.\end{aligned}\tag{A.120}$$

We will bound terms V_1 and V_2 above separately.

Let us begin with the first term V_1. We can expand the square and obtain that

$$\begin{aligned}V_1=\sum_{j\in\mathcal{H}_0}\sum_{\ell\in\mathcal{H}_0}\mathbb{E}\bigg\{\mathbb{E}\bigg[&\Big(\mathbb{1}(\widetilde{W}_j\geq t)-\mathbb{P}(\widetilde{W}_j\geq t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\Big)\\&\times\Big(\mathbb{1}(\widetilde{W}_\ell\geq t)-\mathbb{P}(\widetilde{W}_\ell\geq t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\Big)\bigg|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon}\bigg]\bigg\}.\end{aligned}\tag{A.121}$$

Observe that conditional on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$, it follows from model (14) that $\mathbf{y}$ is deterministic. In addition, $\widetilde{W}_j$ depends only on $\mathbf{X}_j$ and $\widetilde{\mathbf{X}}_j$ besides $\mathbf{y}$. Thus, we need only to consider the conditional distribution of $(\mathbf{X}_j, \widetilde{\mathbf{X}}_j, \mathbf{X}_k, \widetilde{\mathbf{X}}_k)|(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$. We will aim to show that each $\widetilde{W}_j$ depends on at most m_n random variables in $\{\widetilde{W}_k : k \in \mathcal{H}_0\}$. Indeed, it suffices to show that conditional on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$, the number of $(\mathbf{X}_k, \widetilde{\mathbf{X}}_k)$'s that are dependent on $(\mathbf{X}_j, \widetilde{\mathbf{X}}_j)$ is at most m_n. Since the rows of $(\mathbf{X}, \widetilde{\mathbf{X}})$ are i.i.d. and are independent of $\boldsymbol{\varepsilon}$, we need only to consider the distribution of a single row; that is, $(X_j, \widetilde{X}_j, X_k, \widetilde{X}_k)|(X_{\mathcal{H}_1}, \varepsilon) \stackrel{d}{=} (X_j, \widetilde{X}_j, X_k, \widetilde{X}_k)|X_{\mathcal{H}_1}$.

In view of the multinormal distribution in (15), it follows that the conditional distribution $(X_j, \widetilde{X}_j, X_k, \widetilde{X}_k)|X_{\mathcal{H}_1}$ is still normal. We can obtain from the conditional distribution that

$$\begin{aligned}
&\mathrm{Cov}\left\{ \left(\begin{pmatrix} X_j \\ \widetilde{X}_j \end{pmatrix}, \begin{pmatrix} X_k \\ \widetilde{X}_k \end{pmatrix} \right) \middle| X_{\mathcal{H}_1} \right\} \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} & \boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} \\ \boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} & \boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} \end{pmatrix}.
\end{aligned}$$

In particular, $(X_j, \widetilde{X}_j)$ and $(X_k, \widetilde{X}_k)$ are independent conditional on $X_{\mathcal{H}_1}$ if and only if

$$\boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} = 0.$$

Thus, to count the number of dependent pairs of $(X_j, \widetilde{X}_j)$ and $(X_k, \widetilde{X}_k)$ for $j, k \in \mathcal{H}_0$, we need only to count the number of nonzero $(\boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k})$'s. Without loss of generality, let us assume that $X = (X_{\mathcal{H}_1}, X_{\mathcal{H}_0})$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1} & \boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0} \\ \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1} & \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_0} \end{pmatrix}.$$

Using the formula for the block matrix inverse, it holds that

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}^{-1})_{11} & (\boldsymbol{\Sigma}^{-1})_{12} \\ (\boldsymbol{\Sigma}^{-1})_{21} & \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_0} - \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0} \end{pmatrix},$$

where

$$(\boldsymbol{\Sigma}^{-1})_{11} = \boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1} + \boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0}(\boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_0} - \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0})^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1},$$

$$(\boldsymbol{\Sigma}^{-1})_{12} = -\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0}(\boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_0} - \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0})^{-1},$$

and $(\boldsymbol{\Sigma}^{-1})_{21} = (\boldsymbol{\Sigma}^{-1})_{12}^T$. In addition, Condition 9 assumes that $\max_{1\le j\le p} \|(\boldsymbol{\Sigma}^{-1})_j\|_0 \le m_n$, which indicates that

$$\max_{j\in\mathcal{H}_0} \|(\boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_0} - \boldsymbol{\Sigma}_{\mathcal{H}_0,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_0})_j\|_0 \le m_n$$

since it is a submatrix of $\boldsymbol{\Sigma}^{-1}$. Hence, we can obtain that for a given $j \in \mathcal{H}_0$,

$$\sum_{k\in\mathcal{H}_0} \mathbb{1}\Big(\boldsymbol{\Sigma}_{j,k} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,k} = 0\Big) \le m_n.$$

Consequently, we see that conditional on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$, the number of $k \in \mathcal{H}_0$ such that $(\mathbf{X}_k, \widetilde{\mathbf{X}}_k)$ is dependent on $(\mathbf{X}_j, \widetilde{\mathbf{X}}_j)$ is at most m_n. For $j \in \mathcal{H}_0$, let us define

$$N(j) := \{k \in \mathcal{H}_0 : \widetilde{W}_k \not\perp\!\!\!\perp \widetilde{W}_j | (\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})\}.$$

Then it holds that $|N(j)| \leq m_n$. From (A.121) and the fact that the indicator function takes values between 0 and 1, we can deduce that

$$\begin{aligned} V_1 &= \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N(j)} \mathbb{E}\Big\{\mathbb{E}\Big[\mathbb{1}(\widetilde{W}_j \geq t)\cdot \mathbb{1}(\widetilde{W}_\ell \geq t)\Big|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}\Big]\Big\} \\ &\quad - \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N(j)} \mathbb{E}\Big\{\mathbb{E}\Big[\mathbb{P}(\widetilde{W}_j \geq t|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})\mathbb{P}(\widetilde{W}_\ell \geq t\Big|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})\Big]\Big\} \\ &\leq \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N(j)} \mathbb{E}\Big\{\mathbb{E}\Big[\mathbb{1}(\widetilde{W}_j \geq t)\cdot \mathbb{1}(\widetilde{W}_\ell \geq t)\Big|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}\Big]\Big\} \\ &\leq m_n \sum_{j\in\mathcal{H}_0} \mathbb{E}\Big\{\mathbb{E}\Big[\mathbb{1}(\widetilde{W}_j \geq t)\Big|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}\Big]\Big\} \\ &= m_n \sum_{j\in\mathcal{H}_0} \mathbb{P}(\widetilde{W}_j \geq t) = m_n p_0 G(t). \end{aligned} \tag{A.122}$$

We next proceed with showing the bound for term V_2. We can expand V_2 as

$$\begin{aligned} V_2 = \sum_{j\in\mathcal{H}_0}\sum_{\ell\in\mathcal{H}_0} \mathbb{E}\Big\{&\Big(\mathbb{P}(\widetilde{W}_j \geq t|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}) - \mathbb{P}(\widetilde{W}_j \geq t)\Big) \\ &\times \Big(\mathbb{P}(\widetilde{W}_\ell \geq t|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}) - \mathbb{P}(\widetilde{W}_\ell \geq t)\Big)\Big\}. \end{aligned} \tag{A.123}$$

The key idea of the proof is to examine the conditional distribution $\mathbb{P}(\widetilde{W}_j \geq t|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$ and show that given $j \in \mathcal{H}_0$, the number of dependent $\mathbb{P}(\widetilde{W}_\ell \geq t|\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$ is at most m_n. Since $(X, \widetilde{X})$ is multinormal, it holds that

$$(X_j, \widetilde{X}_j)\big|(X_{\mathcal{H}_1}, \varepsilon) \overset{d}{\sim} N\left(\begin{pmatrix} \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}X_{\mathcal{H}_1} \\ \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}X_{\mathcal{H}_1} \end{pmatrix}, \mathrm{Cov}_{cond}\right),$$

where

$$\mathrm{Cov}_{cond} = \begin{pmatrix} \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,j} & \boldsymbol{\Sigma}_{j,j} - r - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,j} \\ \boldsymbol{\Sigma}_{j,j} - r - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,j} & \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,j} \end{pmatrix}.$$

Since the rows of the augmented data matrix $(\mathbf{X}, \widetilde{\mathbf{X}})$ are i.i.d. and $\mathbf{y}$ is deterministic given $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$, we can obtain that

$$\begin{aligned} &\left(\frac{\mathbf{X}_j^T\mathbf{y}}{\sqrt{n}\|\mathbf{y}\|_2}, \frac{\widetilde{\mathbf{X}}_j^T\mathbf{y}}{\sqrt{n}\|\mathbf{y}\|_2}\right)\Big|(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}) \\ &\overset{d}{\sim} N\left((\sqrt{n}\|\mathbf{y}\|_2)^{-1}\begin{pmatrix} \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\mathbf{X}_{\mathcal{H}_1}^T\mathbf{y} \\ \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}^{-1}_{\mathcal{H}_1,\mathcal{H}_1}\mathbf{X}_{\mathcal{H}_1}^T\mathbf{y} \end{pmatrix}, n^{-1}\,\mathrm{Cov}_{cond}\right). \end{aligned} \tag{A.124}$$

Note that when $\boldsymbol{\Sigma}_{\mathcal{H}_1,j} = \mathbf{0}$, the conditional distribution above does not depend on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$ and hence any term involving such $j \in \mathcal{H}_0$ in the expansion of V_2 will disappear. Denote by

$$N_{dep} = \{j \in \mathcal{H}_0 : \boldsymbol{\Sigma}_{\mathcal{H}_1,j} \neq \mathbf{0}\}.$$

It follows from Condition 9 that $|N_{dep}| \le m_n$. Then we have that

$$
\begin{aligned}
V_2 = \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N_{dep}} \mathbb{E}\Big\{ &\Big(\mathbb{P}(\widetilde{W}_j \ge t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon}) - \mathbb{P}(\widetilde{W}_j \ge t)\Big) \\
&\times \Big(\mathbb{P}(\widetilde{W}_\ell \ge t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon}) - \mathbb{P}(\widetilde{W}_\ell \ge t)\Big)\Big\} \\
\le \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N_{dep}} &\mathbb{E}\Big\{\mathbb{P}(\widetilde{W}_j \ge t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\mathbb{P}(\widetilde{W}_\ell \ge t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\Big\} \\
\le \sum_{j\in\mathcal{H}_0}\sum_{\ell\in N_{dep}} &\mathbb{E}\Big\{\mathbb{P}(\widetilde{W}_j \ge t|\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon})\Big\} \le m_n p_0 G(t).
\end{aligned} \tag{A.125}
$$

Therefore, substituting (A.122) and (A.125) into (A.120) yields (A.8). This completes the proof of Lemma 7.

B.8. Proof of Lemma 8. Proof of (A.9). In the proof of Lemma 7 in Section B.7 (cf. (A.124)), we have shown that

$$
\left(\frac{\mathbf{X}_j^T\mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\widetilde{\mathbf{X}}_j^T\mathbf{y}}{\|\mathbf{y}\|_2}\right)\Big|(\mathbf{X}_{\mathcal{H}_1},\boldsymbol{\varepsilon}) \overset{d}{\sim} N\left(\begin{pmatrix}\mu_j\\ \mu_j\end{pmatrix}, \sigma_j^2\begin{pmatrix}1 & \rho_j\\ \rho_j & 1\end{pmatrix}\right), \tag{A.126}
$$

where

$$
\mu_j = \|\mathbf{y}\|_2^{-1}\boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\mathbf{X}_{\mathcal{H}_1}^T\mathbf{y},
$$

$$
\sigma_j^2 = \boldsymbol{\Sigma}_{j,j} - \boldsymbol{\Sigma}_{j,\mathcal{H}_1}\boldsymbol{\Sigma}_{\mathcal{H}_1,\mathcal{H}_1}^{-1}\boldsymbol{\Sigma}_{\mathcal{H}_1,j},\ \ \rho_j = 1 - r/\sigma_j^2,
$$

and r is as given in (15). Recall the definition $N_{dep} = \{j \in \mathcal{H}_0 : \boldsymbol{\Sigma}_{\mathcal{H}_1,j} \ne \mathbf{0}\}$ in the proof of Lemma 7. It holds that $|N_{dep}| \le m_n$ in view of Condition 9. Furthermore, note that

$$
G(t) \ge c_1 q a_n / p
$$

for $t \in (0,\, G^{-1}(\frac{c_1 q a_n}{p})]$. Let us define

$$
R_n := \sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0\cap N_{dep}^c}\mathbb{P}(t-\Delta_n \le \widetilde{W}_j < t+\Delta_n)}{\sum_{j\in\mathcal{H}_0\cap N_{dep}^c}\mathbb{P}(\widetilde{W}_j \ge t)}. \tag{A.127}
$$

Then we can write

$$
\begin{aligned}
&\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]} \frac{G(t-\Delta_n) - G(t+\Delta_n)}{G(t)} \\
&= \sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0\cap N_{dep}}\mathbb{P}(t-\Delta_n \le \widetilde{W}_j < t+\Delta_n)}{p_0 G(t)} + R_n \\
&\le \frac{m_n p}{c_1 q a_n p_0} + R_n.
\end{aligned} \tag{A.128}
$$

From the assumptions that $(\log p)^{1/\gamma} m_n/a_n \to 0$ and $p_0/p \to 1$, we have that

$$
(\log p)^{1/\gamma}\frac{m_n p}{c_1 q a_n p_0} \to 0.
$$

It remains to establish $(\log p)^{1/\gamma} R_n \to 0$. A key observation is that when $j \in \mathcal{H}_0 \cap N_{dep}^c$, it follows that the conditional distribution

$$\left(\frac{\mathbf{X}_j^T \mathbf{y}}{\|\mathbf{y}\|_2}, \frac{\widetilde{\mathbf{X}}_j^T \mathbf{y}}{\|\mathbf{y}\|_2}\right) \Big| (\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon}) \overset{d}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{j,j}^2 & \boldsymbol{\Sigma}_{j,j}^2 - r \\ \boldsymbol{\Sigma}_{j,j}^2 - r & \boldsymbol{\Sigma}_{j,j}^2 \end{pmatrix} \right), \tag{A.129}$$

which does not depend on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$. Then we see that the distribution of $\widetilde{W}_j$ does not depend on $(\mathbf{X}_{\mathcal{H}_1}, \boldsymbol{\varepsilon})$ and satisfies that

$$\mathbb{P}(\sqrt{n}\widetilde{W}_j \geq t) = \mathbb{P}(|Z_1| - |Z_2| \geq t), \tag{A.130}$$

where $(Z_1, Z_2)^T$ is a two-dimensional multinormal random variable with mean $(0,0)^T$ and covariance matrix

$$\begin{pmatrix} \boldsymbol{\Sigma}_{j,j}^2 & \boldsymbol{\Sigma}_{j,j}^2 - r \\ \boldsymbol{\Sigma}_{j,j}^2 - r & \boldsymbol{\Sigma}_{j,j}^2 \end{pmatrix}.$$

For $j \in \mathcal{H}_0 \cap N_{dep}^c$ and $t > 0$, the density function of $\sqrt{n}\widetilde{W}_j$ is given by

$$\begin{aligned} f_{\sqrt{n}\widetilde{W}_j}(t) = & \frac{\sqrt{2}}{\sqrt{\pi} c_{2,j}} \Big[1 - \Phi\Big(\frac{t}{c_{1,j}}\Big)\Big] \exp\Big\{ - \frac{t^2}{2c_{2,j}^2} \Big\} \\ & + \frac{\sqrt{2}}{\sqrt{\pi} c_{1,j}} \Big[1 - \Phi\Big(\frac{t}{c_{2,j}}\Big)\Big] \exp\Big\{ - \frac{t^2}{2c_{1,j}^2} \Big\}, \end{aligned} \tag{A.131}$$

where $c_{1,j} = \sqrt{4\boldsymbol{\Sigma}_{j,j}^2 - 2r}$ and $c_{2,j} = \sqrt{2r}$. Based on the density function of $\sqrt{n}\widetilde{W}_t$ above and the basic inequality that $1 - \Phi(x) \leq e^{-x^2/2}$ for $x \geq 0$, it is easy to see that

$$\begin{aligned} \mathbb{P}(\widetilde{W}_j \geq t) &= \mathbb{P}(\sqrt{n}\widetilde{W}_j \geq \sqrt{n}t) \\ &\leq \int_{\sqrt{n}t}^{\infty} \frac{\sqrt{2}}{\sqrt{\pi} c_{2,j}} \exp\Big\{ - \frac{x^2}{2c_{2,j}^2} \Big\} dx + \int_{\sqrt{n}t}^{\infty} \frac{\sqrt{2}}{\sqrt{\pi} c_{1,j}} \Phi\Big(\frac{-x}{c_{2,j}}\Big) dx \\ &\leq \Big(2 + \frac{2c_{2,j}}{c_{1,j}}\Big) \Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{2,j}}\Big)\Big]. \end{aligned} \tag{A.132}$$

Then we can obtain that

$$G(t) \leq \max_{j \in \mathcal{H}_0} \Big(2 + \frac{2c_{2,j}}{c_{1,j}}\Big) \Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{2,j}}\Big)\Big].$$

Setting $t = G^{-1}(\frac{c_1 q a_n}{p})$ in the inequality above yields that

$$G^{-1}(\frac{c_1 q a_n}{p}) = O(\sqrt{\frac{\log p}{n}})$$

when $C_1 < r < \boldsymbol{\Sigma}_{j,j}^2 < C_2$ with some absolute constants $C_1 > 0$ and $C_2 > 0$ for each $j \in \mathcal{H}_0$.

We will bound the ratio in R_n by considering two ranges of $t \in (0, 4n^{-1/2} \max_{j \in \mathcal{H}_0} c_{1,j} \vee c_{2,j})$ and $t \in [4n^{-1/2} \max_{j \in \mathcal{H}_0} c_{1,j} \vee c_{2,j}, G^{-1}(c_1 q a_n / p)]$ separately. When t falls into the first range, in view of (A.131) the denominator $G(t)$ in the ratio in R_n is of a constant order, while the numerator is uniformly bounded from above by $O(\sqrt{n}\Delta_n)$ over all t in this range because the density $f_{\sqrt{n}\widetilde{W}_j}(t)$ is bounded from above by a constant.

We now consider the ratio in R_n in the second range of $t \in [4n^{-1/2} \max_{j \in \mathcal{H}_0} c_{1,j} \vee c_{2,j}, G^{-1}(c_1 q a_n / p)]$. We will bound the numerator and denominator in (7) separately in

this range. It follows from (A.131) and the mean value theorem that there exists some $\xi \in (\sqrt{n}t - \sqrt{n}\Delta_n, \sqrt{n}t + \sqrt{n}\Delta_n)$ such that

$$\begin{aligned}
&\mathbb{P}(\sqrt{n}t - \sqrt{n}\Delta_n \le \sqrt{n}\widetilde{W}_j \le \sqrt{n}t + \sqrt{n}\Delta_n) \\
&= 2\sqrt{n}\Delta_n \Big\{ \frac{\sqrt{2}}{\sqrt{\pi}c_{2,j}} \Big[1 - \Phi\Big(\frac{\xi}{c_{1,j}}\Big)\Big] \exp\Big\{ - \frac{\xi^2}{2c_{2,j}^2} \Big\} \\
&\quad + \frac{\sqrt{2}}{\sqrt{\pi}c_{1,j}} \exp\Big\{ - \frac{\xi^2}{2c_{1,j}^2} \Big\} \Big[1 - \Phi\Big(\frac{\xi}{c_{2,j}}\Big)\Big] \Big\}.
\end{aligned}$$

Moreover, since $\sqrt{n}t \le \sqrt{n}G^{-1}(\frac{c_1 q a_n}{p}) = O(\sqrt{\log p})$ and $\Delta_n\sqrt{n\log p} \to 0$, we can obtain through some direct calculations that

$$\left| \frac{1 - \Phi\Big(\frac{\xi}{c_{1,j}}\Big)}{1 - \Phi\Big(\frac{\sqrt{n}t}{c_{1,j}}\Big)} - 1 \right| \le C\sqrt{n}t \cdot \sqrt{n}\Delta_n = O(\Delta_n\sqrt{n\log p}).$$

Similarly, it holds that

$$\left| \frac{\exp\Big\{ - \frac{\xi^2}{2c_{1,j}^2} \Big\}}{\exp\Big\{ - \frac{(\sqrt{n}t)^2}{2c_{1,j}^2} \Big\}} - 1 \right| \le C\sqrt{n}t \cdot \sqrt{n}\Delta_n = O(\Delta_n\sqrt{n\log p}).$$

Combining the above three inequalities yields that when $\Delta_n\sqrt{n\log p} \to 0$,

$$\begin{aligned}
&\mathbb{P}(t - \Delta_n \le \widetilde{W}_j < t + \Delta_n) \\
&= \mathbb{P}(\sqrt{n}t - \sqrt{n}\Delta_n \le \sqrt{n}\widetilde{W}_j \le \sqrt{n}t + \sqrt{n}\Delta_n) \\
&\le C\sqrt{n}\Delta_n[1 + O(\sqrt{n}\Delta_n \log p)] \Big\{ \frac{\sqrt{2}}{\sqrt{\pi}c_{2,j}} \Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{1,j}}\Big)\Big] \exp\Big\{ - \frac{(\sqrt{n}t)^2}{2c_{2,j}^2} \Big\} \\
&\quad + \frac{\sqrt{2}}{\sqrt{\pi}c_{1,j}} \Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{2,j}}\Big)\Big] \exp\Big\{ - \frac{(\sqrt{n}t)^2}{2c_{1,j}^2} \Big\} \Big\}.
\end{aligned} \tag{A.133}$$

Next we need to deal with the denominator $\mathbb{P}(\sqrt{n}\widetilde{W}_j \ge t)$. Via integration by parts, we can deduce that for $t \in [4n^{-1/2}\max_{j\in\mathcal{H}_0} c_{1,j} \vee c_{2,j}, G^{-1}(c_1 q a_n/p)]$,

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\widetilde{W}_j \ge \sqrt{n}t) &= 2\Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{1,j}}\Big)\Big]\Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{2,j}}\Big)\Big] \\
&\ge C\Big\{ (\sqrt{n}t)^{-1}\Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{1,j}}\Big)\Big] \exp\Big\{ - \frac{(\sqrt{n}t)^2}{2c_{2,j}^2} \Big\} \\
&\quad + (\sqrt{n}t)^{-1}\Big[1 - \Phi\Big(\frac{\sqrt{n}t}{c_{2,j}}\Big)\Big] \exp\Big\{ - \frac{(\sqrt{n}t)^2}{2c_{1,j}^2} \Big\} \Big\} \\
&\ge \widetilde{C}(\sqrt{n}t)^{-1} f_{\sqrt{n}\widetilde{W}_j}(\sqrt{n}t),
\end{aligned} \tag{A.134}$$

where we have used the definition of the density in (A.131) and the fact that

$$1 - \Phi(x) \ge 0.75x^{-1}e^{-x^2/2}$$

for $x \ge 4$, and $\widetilde{C}$ is some constant depending on $c_{1,j}$ and $c_{2,j}$.

Combining (A.133) and (A.134) and using some direct calculations, we can obtain the bound for the ratio in R_n in the second range

$$\begin{aligned}(\text{A.135}) \quad &\sup_{t\in[4n^{-1/2}\max_{j\in\mathcal{H}_0}c_{1,j}\vee c_{2,j},G^{-1}(c_1qa_n/p))} \frac{\sum_{j\in\mathcal{H}_0\cap N_{dep}^c}\mathbb{P}(t-\Delta_n\le\widetilde{W}_j<t+\Delta_n)}{\sum_{j\in\mathcal{H}_0\cap N_{dep}^c}\mathbb{P}(\widetilde{W}_j\ge t)} \\ &\le \widetilde{C}\sqrt{n}\Delta_n\cdot\sqrt{n}G^{-1}(\frac{c_1qa_n}{p}) = O(\sqrt{n}\Delta_n\sqrt{\log p}).\end{aligned}$$

This together with the result for the first range proven previously leads to

$$(\text{A.136}) \qquad R_n = O(\sqrt{n}\Delta_n\sqrt{\log p}).$$

Finally, plugging (A.136) into (A.128) yields (A.9) because $(\log p)^{1/\gamma}m_n/a_n\to 0$ and

$$\sqrt{n}\Delta_n(\log p)^{1/2+1/\gamma}\to 0.$$

Proof of (A.10). Recall from Condition 10 that

$$p_1^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}(\widetilde{W}_j<-t)\le G(t)$$

for $t\in(0,C\sqrt{n^{-1}\log p})$ with C some large constant. Also, note that

$$\Delta_n = o(G^{-1}(\frac{c_1qa_n}{p}))$$

since $\sqrt{n}\Delta_n\to 0$ by assumption and $G^{-1}(\frac{c_1qa_n}{p}) = O(\sqrt{n^{-1}\log p})$ as shown in the proof of (A.9). It follows from some direct calculations that

$$\begin{aligned}(\text{A.137}) \quad &a_n^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}\Big(\widetilde{W}_j<-G^{-1}(\frac{c_1qa_n}{p})+\Delta_n\Big) \\ &\le a_n^{-1}(p-p_0)G\Big(G^{-1}(\frac{c_1qa_n}{p})-\Delta_n\Big) \\ &= \frac{c_1q(p-p_0)}{p}+a_n^{-1}(p-p_0)|G'(\xi)|\Delta_n,\end{aligned}$$

where ξ is some number lying between $G^{-1}(\frac{c_1qa_n}{p})$ and $G^{-1}(\frac{c_1qa_n}{p})-\Delta_n$. From (A.131) and $f_{\sqrt{n}\widetilde{W}_j}(\sqrt{n}\xi)\le C$ with $C>0$ some constant, we can deduce that

$$\begin{aligned}|G'(\xi)| &= \sum_{j\in\mathcal{H}_0}p_0^{-1}\sqrt{n}f_{\sqrt{n}\widetilde{W}_j}(\sqrt{n}\xi) \\ &\le \frac{C\sqrt{n}m_n}{p_0}+p_0^{-1}\sum_{j\in\mathcal{H}_0\cap N_{dep}^c}\sqrt{n}f_{\sqrt{n}\widetilde{W}_j}(\sqrt{n}\xi) \\ &\le \frac{C\sqrt{n}m_n}{p_0}+Cp_0^{-1}\sqrt{n}\cdot\sqrt{n}G(\frac{c_1qa_n}{p})\sum_{H_0\cap N_{dep}^c}\mathbb{P}\Big(\widetilde{W}_j\ge G(\frac{c_1qa_n}{p})\Big) \\ &\le \frac{C\sqrt{n}m_n}{p_0}+Cp_0^{-1}\sqrt{n\log p}\,p_0\frac{c_1qa_n}{p},\end{aligned}$$

where the second last step above is due to (A.134).

Therefore, substituting the bound above into (A.137) gives that

$$\begin{aligned} & a_n^{-1} \sum_{j \in \mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p} + \Delta_n\Big) \\ & \leq \frac{c_1 q(p-p_0)}{p} + \frac{C\Delta_n \sqrt{n} m_n (p-p_0)}{a_n p_0} \\ & \quad + \frac{C\Delta_n \sqrt{n \log p}\, q(p-p_0)}{p} \\ & \to 0, \end{aligned}$$

where we have used the assumption that $p_0/p \to 1$, $\Delta_n \sqrt{n \log p} \to 0$, and $m_n/a_n \to 0$. This derives (A.10), which concludes the proof of Lemma 8.

B.9. Proof of Lemma 11. The main intuition of the proof is that when the approximate augmented data matrix $\widehat{\mathbf{X}}^{\text{aug}}$ is close to its perfect counterpart $\widetilde{\mathbf{X}}^{\text{aug}}$, the corresponding Lasso estimators would be close as well. From the definitions of $\widetilde{\boldsymbol{\beta}}_j$ in (22) and $\widehat{\boldsymbol{\beta}}_j$ in (19), it holds that

$$\begin{aligned} \max_{1 \leq j \leq 2p} |\widetilde{\beta}_j - \widehat{\beta}_j| & \leq \max_{1 \leq j \leq 2p} |\widetilde{\beta}_j^{\text{init}} - \widehat{\beta}_j^{\text{init}}| \\ & \quad + \max_{1 \leq j \leq 2p} \left| \frac{\widetilde{\mathbf{z}}_j^T \big(\mathbf{y} - \widetilde{\mathbf{X}}^{\text{aug}} \widetilde{\boldsymbol{\beta}}^{\text{init}}\big)}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} - \frac{\widehat{\mathbf{z}}_j^T \big(\mathbf{y} - \widehat{\mathbf{X}}^{\text{aug}} \widehat{\boldsymbol{\beta}}^{\text{init}}\big)}{\widehat{\mathbf{z}}_j^T \widehat{\mathbf{X}}_j^{\text{aug}}} \right|. \end{aligned} \tag{A.138}$$

We will aim to prove that for some large enough constant C,

$$\mathbb{P}\left(\|\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}}\|_2 \leq C\Delta_n s \sqrt{\frac{\log p}{n}} \right) \to 1, \tag{A.139}$$

$$\mathbb{P}\left(\max_{1 \leq j \leq 2p} \left| \frac{\widetilde{\mathbf{z}}_j^T \big(\mathbf{y} - \widetilde{\mathbf{X}}^{\text{aug}} \widetilde{\boldsymbol{\beta}}^{\text{init}}\big)}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} - \frac{\widehat{\mathbf{z}}_j^T \big(\mathbf{y} - \widehat{\mathbf{X}}^{\text{aug}} \widehat{\boldsymbol{\beta}}^{\text{init}}\big)}{\widehat{\mathbf{z}}_j^T \widehat{\mathbf{X}}_j^{\text{aug}}} \right| \leq C\Delta_n s \sqrt{\frac{\log p}{n}} \right) \to 1. \tag{A.140}$$

Then combining the two results above can establish the desired conclusion of Lemma 11. We next proceed with proving (A.139) and (A.140).

Proof of (A.139). It follows from the Karush–Kuhn–Tucker (KKT) condition that

$$n^{-1} [\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}) = n^{-1} [\widetilde{\mathbf{X}}^{\text{aug}}]^T \boldsymbol{\varepsilon} - \lambda \widetilde{\boldsymbol{\zeta}}, \tag{A.141}$$

$$n^{-1} [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} (\widehat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}) = n^{-1} [\widehat{\mathbf{X}}^{\text{aug}}]^T \boldsymbol{\varepsilon} - \lambda \widehat{\boldsymbol{\zeta}}, \tag{A.142}$$

where $\widetilde{\boldsymbol{\zeta}} = (\widetilde{\zeta}_1, \cdots, \widetilde{\zeta}_{2p})$ and $\widehat{\boldsymbol{\zeta}} = (\widehat{\zeta}_1, \cdots, \widehat{\zeta}_{2p})$ with

$$\widetilde{\zeta}_j = \begin{cases} \operatorname{sgn}(\widetilde{\beta}_j^{\text{init}}) & \text{if } \widetilde{\beta}_j^{\text{init}} \neq 0, \\ \in [-1,1] & \text{if } \widetilde{\beta}_j^{\text{init}} = 0, \end{cases} \quad \text{and} \quad \widehat{\zeta}_j = \begin{cases} \operatorname{sgn}(\widehat{\beta}_j^{\text{init}}) & \text{if } \widehat{\beta}_j^{\text{init}} \neq 0, \\ \in [-1,1] & \text{if } \widehat{\beta}_j^{\text{init}} = 0. \end{cases}$$

Taking the difference between (A.141) and (A.142) above leads to

$$\begin{aligned} & n^{-1} [\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}}) + n^{-1} \Big([\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big) (\widehat{\boldsymbol{\beta}}^{\text{init}} - \widetilde{\boldsymbol{\beta}}^{\text{init}}) \\ & = -n^{-1} \Big([\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big) (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}) \\ & \quad + n^{-1} \Big(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}} \Big)^T \boldsymbol{\varepsilon} - \lambda (\widetilde{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\zeta}}). \end{aligned}$$

Furthermore, multiplying both sides of the equation above by $(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T$ yields that

$$
\begin{aligned}
& n^{-1}\|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})\|_2^2 \\
& = n^{-1}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T \Big([\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big) (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}}) \\
& \quad - n^{-1}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T \Big([\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big) (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}}) \\
& \quad + n^{-1}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T \Big(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}} \Big)^T \boldsymbol{\varepsilon} - \lambda (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T (\widetilde{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\zeta}}).
\end{aligned} \tag{A.143}
$$

We claim that the last term on the right-hand side of the expression above satisfies that

$$
(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T (\widetilde{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\zeta}}) \geq 0.
$$

To understand this, observe that when both $\widetilde{\beta}_j^{\text{init}}$ and $\widehat{\beta}_j^{\text{init}}$ are nonzero or zero, it is easy to see that

$$
(\widetilde{\beta}_j^{\text{init}} - \widehat{\beta}_j^{\text{init}})(\widetilde{\zeta}_j - \widehat{\zeta}_j) \geq 0.
$$

When either of $\widetilde{\beta}_j^{\text{init}}$ and $\widehat{\beta}_j^{\text{init}}$ is zero, without loss of generality let us assume that $\widetilde{\beta}_j^{\text{init}} = 0$ and $\widehat{\beta}_j^{\text{init}} \neq 0$. When $\widetilde{\beta}_j^{\text{init}} = 0$ and $\widehat{\beta}_j^{\text{init}} > 0$, it follows that $\widetilde{\zeta}_j \leq 1 = \widehat{\zeta}_j$ and hence

$$
(\widetilde{\beta}_j^{\text{init}} - \widehat{\beta}_j^{\text{init}})(\widetilde{\zeta}_j - \widehat{\zeta}_j) = -\widehat{\beta}_j^{\text{init}}((\widetilde{\zeta}_j - \widehat{\zeta}_j)) \geq 0.
$$

Similarly, we can show that

$$
(\widetilde{\beta}_j^{\text{init}} - \widehat{\beta}_j^{\text{init}})(\widetilde{\zeta}_j - \widehat{\zeta}_j) \geq 0
$$

when $\widetilde{\beta}_j^{\text{init}} = 0$ and $\widehat{\beta}_j^{\text{init}} < 0$. Thus, the last term on the right-hand side of (A.143) above satisfies that

$$
-(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T (\widetilde{\boldsymbol{\zeta}} - \widehat{\boldsymbol{\zeta}}) \leq 0.
$$

We next examine the three terms on the right-hand side of the earlier expression above separately.

First, we observe that

$$
\begin{aligned}
& \Big\| n^{-1}[\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big\|_{\max} \\
& \leq \Big\| n^{-1}[\widetilde{\mathbf{X}}^{\text{aug}}]^T (\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}}) \Big\|_{\max} + \|n^{-1}(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})]^T \widehat{\mathbf{X}}^{\text{aug}}\|_{\max} \\
& \leq \max_j \|n^{-1/2} \widetilde{\mathbf{X}}_j^{\text{aug}}\|_2 \max_j \|n^{-1/2}(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_j^{\text{aug}})\|_2 \\
& \quad + \max_j \|n^{-1/2} \widehat{\mathbf{X}}_j^{\text{aug}}\|_2 \max_j \|n^{-1/2}(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_j^{\text{aug}})\|_2.
\end{aligned}
$$

Under Condition 6 and the sub-Gaussian assumption for $\mathbf{X}$, it can be shown that

$$
\mathbb{P}\left(\Big\| n^{-1}[\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big\|_{\max} \geq C\Delta_n \right) \to 0 \tag{A.144}
$$

for some constant $C > 0$. From the sparsity of $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ in Condition 11, we have that with probability $1 - o(1)$, the first term on the right-hand side of (A.143) can be bounded as

$$
\begin{aligned}
& n^{-1} \Big| (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})^T \Big([\widetilde{\mathbf{X}}^{\text{aug}}]^T \widetilde{\mathbf{X}}^{\text{aug}} - [\widehat{\mathbf{X}}^{\text{aug}}]^T \widehat{\mathbf{X}}^{\text{aug}} \Big) (\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}}) \Big| \\
& \leq C\Delta_n s \|\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}}\|_2^2.
\end{aligned} \tag{A.145}
$$

By the Cauchy–Schwarz inequality, we can bound the second term on the right-hand side of (A.143) as

$$\begin{aligned}&\left|n^{-1}(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\widehat{\boldsymbol{\beta}}^{\text{init}})^T\left([\widetilde{\mathbf{X}}^{\text{aug}}]^T\widetilde{\mathbf{X}}^{\text{aug}}-[\widehat{\mathbf{X}}^{\text{aug}}]^T\widehat{\mathbf{X}}^{\text{aug}}\right)(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\boldsymbol{\beta}^{\text{aug}})\right|\\&\le\|\widetilde{\boldsymbol{\beta}}^{\text{init}}-\widehat{\boldsymbol{\beta}}^{\text{init}}\|_2\left\|n^{-1}\left([\widetilde{\mathbf{X}}^{\text{aug}}]^T\widetilde{\mathbf{X}}^{\text{aug}}-[\widehat{\mathbf{X}}^{\text{aug}}]^T\widehat{\mathbf{X}}^{\text{aug}}\right)(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\boldsymbol{\beta}^{\text{aug}})\right\|_2.\end{aligned}$$

Finally, with the aid of Condition 11 on sparsity and Condition 12 on the restrictive eigenvalues, the left-hand side of (A.143) can be lower bounded by $c_1\|\widetilde{\boldsymbol{\beta}}^{\text{init}}-\widehat{\boldsymbol{\beta}}^{\text{init}}\|_2^2$. Combining all the results above and applying the Cauchy–Schwarz inequality to the second and third terms on the right-hand side of (A.143), we can deduce that as $\Delta_n s\to 0$, the representation in (A.143) entails that with probability $1-o(1)$,

$$\begin{aligned}\|\widetilde{\boldsymbol{\beta}}^{\text{init}}-\widehat{\boldsymbol{\beta}}^{\text{init}}\|_2&\lesssim\left\|n^{-1}\left([\widetilde{\mathbf{X}}^{\text{aug}}]^T\widetilde{\mathbf{X}}^{\text{aug}}-[\widehat{\mathbf{X}}^{\text{aug}}]^T\widehat{\mathbf{X}}^{\text{aug}}\right)(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\boldsymbol{\beta}^{\text{aug}})\right\|_2\\&\quad+\max_{J:|J|\le Cs}\left\|n^{-1}\left(\widetilde{\mathbf{X}}_J^{\text{aug}}-\widehat{\mathbf{X}}_J^{\text{aug}}\right)^T\boldsymbol{\varepsilon}\right\|_2:=I_1+I_2.\end{aligned}\tag{A.146}$$

We will bound the two terms I_1 and I_2 above separately. It follows from (A.144), the sparsity of $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{\text{aug}}$, and Lemma 9 that with probability $1-o(1)$,

$$I_1\le C\Delta_n s^{1/2}\|\widetilde{\boldsymbol{\beta}}^{\text{init}}-\boldsymbol{\beta}^{\text{aug}}\|_2\le C\Delta_n s\sqrt{\frac{\log p}{n}}.\tag{A.147}$$

As for term I_2, conditional on $(\widetilde{\mathbf{X}}^{\text{aug}},\widehat{\mathbf{X}}^{\text{aug}})$ we have that for each $1\le j\le 2p$,

$$n^{-1/2}\left(\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\right)^T\boldsymbol{\varepsilon}\overset{d}{\sim}N\left(0,n^{-1}\left\|\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\right\|_2^2\right).$$

Thus, it holds that

$$\begin{aligned}&\mathbb{P}\left(I_2\ge C\sigma\Delta_n\sqrt{\frac{s\log n}{n}}\right)\\&\le\mathbb{P}\left(s\max_{1\le j\le 2p}\left(n^{-1/2}\left(\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\right)^T\boldsymbol{\varepsilon}\right)^2\ge C^2\sigma^2\Delta_n^2 s\log n\right)\\&=\mathbb{P}\left(\max_{1\le j\le 2p}n^{-1/2}\left\|\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\right\|_2|Z|\ge C\sigma\Delta_n\sqrt{\log n}\right),\end{aligned}$$

where $Z\overset{d}{\sim}N(0,\sigma^2)$ is independent of $\widetilde{\mathbf{X}}^{\text{aug}}$ and $\widehat{\mathbf{X}}^{\text{aug}}$.

Moreover, Condition 6 implies that

$$\max_{1\le j\le 2p}n^{-1/2}\|\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\|_2\le\Delta_n$$

with probability $1-o(1)$. Then using the union bound, we can obtain that for some constant $C>\sqrt{2}$,

$$\begin{aligned}\mathbb{P}\left(I_2\ge C\sigma\Delta_n\sqrt{\frac{s\log n}{n}}\right)&\le\mathbb{P}\left(|Z|>C\sigma\sqrt{\log n}\right)\\&\quad+\mathbb{P}(\max_{1\le j\le 2p}\|\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}}\|_2\ge\Delta_n)\to 0.\end{aligned}\tag{A.148}$$

Consequently, substituting (A.147) and (A.148) into (A.146) leads to (A.139). Further, applying (A.143) again with the bounds in (A.146), (A.147), (A.148), and (A.139) yields that

$$\mathbb{P}\left(n^{-1/2}\|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\widehat{\boldsymbol{\beta}}^{\text{init}})\|_2 \le C\Delta_n s\sqrt{\frac{\log p}{n}}\right)\to 1. \tag{A.149}$$

Proof of (A.140)**.** Let us first state three results (A.150), (A.151), and (A.152) below that will be used repeatedly in our proof. With similar arguments as for (A.139) and (A.149) and the union bound, we can deduce that under Conditions 11–13,

$$\mathbb{P}\left(\max_{1\le j\le 2p}\|\widetilde{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_j\|_2 \le C\left(m_n^{1/2}\Delta_n+\Delta_n m_n\sqrt{\frac{\log p}{n}}\right)\le Cm_n^{1/2}\Delta_n\right)\to 1, \tag{A.150}$$

$$\mathbb{P}\left(n^{-1/2}\max_j\|\widetilde{\mathbf{X}}^{\text{aug}}_{-j}(\widetilde{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_j)\|_2 \le Cm_n^{1/2}\Delta_n\right)\to 1, \tag{A.151}$$

where we have used $\sqrt{\frac{m_n\log p}{n}}\to 0$ for showing (A.150). Observe that for $1\le j\le 2p$,

$$\begin{aligned}\|n^{-1/2}(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)\|_2 \le{}& \|n^{-1/2}(\widetilde{\mathbf{X}}^{\text{aug}}_j-\widehat{\mathbf{X}}^{\text{aug}}_j)\|_2+\|n^{-1/2}\widetilde{\mathbf{X}}^{\text{aug}}_{-j}(\widetilde{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_j)\|_2\\ &+\|n^{-1/2}(\widetilde{\mathbf{X}}^{\text{aug}}_{-j}-\widehat{\mathbf{X}}^{\text{aug}}_{-j})\boldsymbol{\gamma}_j\|_2\\ &+\|n^{-1/2}(\widetilde{\mathbf{X}}^{\text{aug}}_{-j}-\widehat{\mathbf{X}}^{\text{aug}}_{-j})(\widehat{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j)\|_2.\end{aligned}$$

Then it follows from the sparsity of $\mathcal{S}_j=\text{supp}(\boldsymbol{\gamma}_j)\cup\text{supp}(\widetilde{\boldsymbol{\gamma}}_j)\cup\text{supp}(\widehat{\boldsymbol{\gamma}}_j)$, the sub-Gaussianity of X_j, and the bound in (A.150) that with probability $1-o(p^{-1})$,

$$\begin{aligned}&\max_{1\le j\le 2p}\|n^{-1/2}(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)\|_2\\ &\le C\left(\Delta_n+\Delta_n m_n^{1/2}+\Delta_n m_n^{1/2}\max_{1\le j\le 2p}\|\boldsymbol{\gamma}_j\|_2+m_n\Delta_n\sqrt{\frac{\log p}{n}}\right)\\ &\le C\Delta_n m_n^{1/2}.\end{aligned} \tag{A.152}$$

We are now ready to establish (A.140). In particular, we have the decomposition for the main term in (A.140)

$$\begin{aligned}&\max_{1\le j\le 2p}\left|\frac{\widetilde{\mathbf{z}}_j^T(\mathbf{y}-\widetilde{\mathbf{X}}^{\text{aug}}\widetilde{\boldsymbol{\beta}}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}-\frac{\widehat{\mathbf{z}}_j^T(\mathbf{y}-\widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}})}{\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}^{\text{aug}}_j}\right|\\ &\le \max_{1\le j\le 2p}\left|\frac{(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)^T(\mathbf{y}-\widetilde{\mathbf{X}}^{\text{aug}}\widetilde{\boldsymbol{\beta}}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}\right|+\max_{1\le j\le 2p}\left|\frac{\widehat{\mathbf{z}}_j^T(\widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}}-\widetilde{\mathbf{X}}^{\text{aug}}\widetilde{\boldsymbol{\beta}}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}\right|\\ &\quad+\max_{1\le j\le 2p}\left|\widehat{\mathbf{z}}_j^T(\mathbf{y}-\widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}})\left(\frac{1}{\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}^{\text{aug}}_j}-\frac{1}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}\right)\right|:=P_1+P_2+P_3.\end{aligned} \tag{A.153}$$

We will investigate the three terms P_1, P_2, and P_3 above separately. Let us first deal with term P_1. Note that

$$P_1\le\max_{1\le j\le 2p}\left|\frac{(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)^T\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}}-\boldsymbol{\beta}^{\text{aug}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}\right|+\max_{1\le j\le 2p}\left|\frac{(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)^T\boldsymbol{\varepsilon}}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}_j}\right|. \tag{A.154}$$

Since $\varepsilon \overset{d}{\sim} N(\mathbf{0}, I_n)$ and is independent of design matrix $\mathbf{X}$, it holds that conditional on design matrix $\mathbf{X}$,

$$\frac{(\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j)^T \varepsilon}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \overset{d}{\sim} N\bigg(0, \frac{\|\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j\|_2^2}{[\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}]^2}\bigg).$$

This together with the bounds in (A.152) and (A.186) leads to

$$\begin{aligned}
&\mathbb{P}\bigg(\max_{1\le j\le 2p}\bigg|\frac{(\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j)^T \varepsilon}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| > Cm_n^{1/2}\Delta_n\sqrt{\frac{\log p}{n}}\bigg) \\
&= \sum_{j=1}^{2p}\mathbb{P}\bigg(\frac{\|\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j\|_2}{|\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}|}\cdot|Z| > Cm_n^{1/2}\Delta_n\sqrt{\frac{\log p}{n}}\bigg) \\
&\le \sum_{j=1}^{2p}\mathbb{P}\bigg(\frac{\|\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j\|_2}{n}\cdot|Z| > Cm_n^{1/2}\Delta_n\sqrt{\frac{\log p}{n}}\bigg) + o(1) \\
&\le \sum_{j=1}^{2p}\mathbb{P}(|Z| > C\sqrt{\log p}) + o(1) = o(1),
\end{aligned} \tag{A.155}$$

where $Z \overset{d}{\sim} N(0, \sigma^2)$ is independent of $\widetilde{\mathbf{X}}^{\text{aug}}$ and $\widehat{\mathbf{X}}^{\text{aug}}$, and C is some large constant that may take different value at each appearance.

In addition, from (A.186), the Cauchy–Schwarz inequality, Lemma 9, and (A.152), we can deduce that with probability $1 - o(1)$,

$$\begin{aligned}
\max_{1\le j\le 2p}\bigg|\frac{(\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j)^T \widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| &\le \max_{1\le j\le 2p}\frac{\|\widetilde{\mathbf{z}}_j - \widehat{\mathbf{z}}_j\|_2\|\widetilde{\mathbf{X}}^{\text{aug}}(\boldsymbol{\beta}^{\text{aug}} - \widetilde{\boldsymbol{\beta}}^{\text{init}})\|_2}{|\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}|} \\
&\le C\Delta_n m_n^{1/2}\sqrt{\frac{s\log p}{n}}.
\end{aligned} \tag{A.156}$$

Substituting (A.155) and (A.156) into (A.154) yields that with probability $1 - o(1)$,

$$P_1 \le C\Delta_n m_n^{1/2}\sqrt{\frac{s\log p}{n}}. \tag{A.157}$$

We next turn to the bound for term P_2. It is easy to see that

$$\begin{aligned}
P_2 \le &\max_{1\le j\le 2p}\bigg|\frac{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| + \max_{1\le j\le 2p}\bigg|\frac{\widetilde{\mathbf{z}}_j^T (\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})\widehat{\boldsymbol{\beta}}^{\text{init}}}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| \\
&+ \max_{1\le j\le 2p}\bigg|\frac{(\widehat{\mathbf{z}}_j - \widetilde{\mathbf{z}}_j)^T \widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| \\
&+ \max_{1\le j\le 2p}\bigg|\frac{(\widehat{\mathbf{z}}_j - \widetilde{\mathbf{z}}_j)^T (\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})\widehat{\boldsymbol{\beta}}^{\text{init}}}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\bigg| \\
:= &P_{21} + P_{22} + P_{23} + P_{24}.
\end{aligned} \tag{A.158}$$

Regarding term P_{21}, in view of (A.139) and the definition of $\widetilde{\mathbf{z}}_j$, we have that with probability $1-o(1)$,

$$\begin{aligned} P_{21} &\le \max_{1\le j\le 2p} |\widetilde{\beta}_j^{\text{init}} - \widehat{\beta}_j^{\text{init}}| + \max_{1\le j\le 2p} \left| \frac{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right| \\ &\le C\Delta_n s\sqrt{\frac{\log p}{n}} + \max_{1\le j\le 2p} \left| \frac{(\mathbf{e}_j + \widetilde{\mathbf{X}}_{-j}^{\text{aug}}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j))^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right| \\ &\le C\Delta_n s\sqrt{\frac{\log p}{n}} + \max_{1\le j\le 2p} \left| \frac{\mathbf{e}_j^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right| \\ &\quad + \max_{1\le j\le 2p} \left| \frac{[\widetilde{\mathbf{X}}_{-j}^{\text{aug}}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)]^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right|. \end{aligned} \tag{A.159}$$

We will bound the last two terms on the very right-hand side of the expression above separately.

Since for $\ell \neq j$, $n^{-1}\mathbb{E}[\mathbf{e}_j^T \widetilde{\mathbf{X}}_\ell^{\text{aug}}] = 0$ due to zero correlation between $\mathbf{e}_j$ and $\mathbf{X}_{-j}^{\text{aug}}$, and $\mathbf{e}_j$ and $\widetilde{\mathbf{X}}_\ell^{\text{aug}}$ both have i.i.d. sub-Gaussian entries, we can show that for $\ell \neq j$,

$$\mathbb{P}\left(\max_{1\le j\le 2p} \max_{\ell\neq j} n^{-1} |\mathbf{e}_j^T \mathbf{X}_\ell^{\text{aug}}| \ge C\sqrt{\frac{\log p}{n}} \right) \le Cp^{-1} \to 0. \tag{A.160}$$

This combined with (A.186), the sparsity assumption that $|J| = |\text{supp}(\boldsymbol{\beta}) \cup \text{supp}(\widetilde{\boldsymbol{\beta}}) \cup \text{supp}(\widehat{\boldsymbol{\beta}})| \lesssim s$, and the result in (A.139) yields that with probability $1-o(1)$, the second term on the very right-hand side of (A.159) above can be bounded as

$$\begin{aligned} &\max_{1\le j\le 2p} \left| \frac{\mathbf{e}_j^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right| \\ &\le Cn^{-1} \max_{1\le j\le 2p} \max_{J': |J'|\lesssim s} \|\mathbf{e}_j^T \widetilde{\mathbf{X}}_{J'\setminus\{j\}}^{\text{aug}}\|_2 \cdot \|\widetilde{\boldsymbol{\beta}}_{J'\setminus\{j\}}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{J'\setminus\{j\}}^{\text{init}}\|_2 \\ &\le C\sqrt{\frac{s\log p}{n}} \cdot \Delta_n s\sqrt{\frac{\log p}{n}} \le C\Delta_n s\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{A.161}$$

where the last inequality above holds due to the assumption that $\sqrt{\frac{s\log p}{n}} \to 0$. By the Cauchy–Schwarz inequality, we can deduce that with probability $1-o(1)$, the third term on the very right-hand side of (A.159) above can be bounded as

$$\begin{aligned} &\max_{1\le j\le 2p} \left| \frac{[\widetilde{\mathbf{X}}^{\text{aug}}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)]^T \widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}} \right| \\ &\le Cn^{-1} \max_{1\le j\le 2p} \|\widetilde{\mathbf{X}}^{\text{aug}}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)\|_2 \cdot \|\widetilde{\mathbf{X}}_{-j}^{\text{aug}} (\widetilde{\boldsymbol{\beta}}_{-j}^{\text{init}} - \widehat{\boldsymbol{\beta}}_{-j}^{\text{init}})\|_2. \end{aligned} \tag{A.162}$$

An application of Lemma 10, (A.139), and the sub-Gaussian assumption of X_j gives that with probability $1-o(1)$, the second term on the right-hand side above can be bounded as

$$\begin{aligned}
&\max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\big(\widetilde{\boldsymbol{\beta}}^{\mathrm{init}}_{-j}-\widehat{\boldsymbol{\beta}}^{\mathrm{init}}_{-j}\big)\|_2\\
&\le n^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}\big(\widetilde{\boldsymbol{\beta}}^{\mathrm{init}}-\widehat{\boldsymbol{\beta}}^{\mathrm{init}}\big)\|_2\\
&\quad+\max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}_{j}\|_2|\widetilde{\beta}_j-\widehat{\beta}_j|\\
&\le C\Delta_n s\sqrt{\frac{\log p}{n}}.
\end{aligned}\tag{A.163}$$

Then plugging (A.163) into (A.162) yields that

$$\begin{aligned}
&\max_{1\le j\le 2p}\left|\frac{[\widetilde{\mathbf{X}}^{\mathrm{aug}}(\boldsymbol{\gamma}_j-\widetilde{\boldsymbol{\gamma}}_j)]^T\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\big(\widetilde{\boldsymbol{\beta}}^{\mathrm{init}}_{-j}-\widehat{\boldsymbol{\beta}}^{\mathrm{init}}_{-j}\big)}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}_j}\right|\\
&\le C\sqrt{\frac{m_n\log p}{n}}\cdot C\Delta_n s\sqrt{\frac{\log p}{n}}\le C\Delta_n s\sqrt{\frac{\log p}{n}},
\end{aligned}\tag{A.164}$$

where the last inequality above is due to the assumption that $\sqrt{\frac{s\log p}{n}}\to 0$ and $m_n\lesssim s$. Hence, it follows from substituting (A.161) and (A.164) into (A.159) that with probability $1-o(1)$,

$$P_{21}\le C\Delta_n s\sqrt{\frac{\log p}{n}}.\tag{A.165}$$

We next proceed with considering term P_{22} introduced in (A.158). Observe that

$$\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}}=[\mathbf{0},\widetilde{\mathbf{X}}-\widehat{\mathbf{X}}]$$

and $\boldsymbol{\beta}^{\mathrm{aug}}=(\boldsymbol{\beta}^T,\mathbf{0}^T)^T$. Then it holds that

$$(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})\boldsymbol{\beta}^{\mathrm{aug}}=\mathbf{0}.$$

From (A.186) and the Cauchy–Schwarz inequality, we can deduce that

$$\begin{aligned}
P_{22}&\le\max_{1\le j\le 2p}\left|\frac{\widetilde{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})\boldsymbol{\beta}^{\mathrm{aug}}}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}_j}\right|+\max_{1\le j\le 2p}\left|\frac{\widetilde{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})(\widehat{\boldsymbol{\beta}}^{\mathrm{init}}-\boldsymbol{\beta}^{\mathrm{aug}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}_j}\right|\\
&\le Cn^{-1}\max_{1\le j\le 2p}\|\widetilde{\mathbf{z}}_j\|_2\cdot\|(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})(\widehat{\boldsymbol{\beta}}^{\mathrm{init}}-\boldsymbol{\beta}^{\mathrm{aug}})\|_2.
\end{aligned}$$

Moreover, we have $\widetilde{\mathbf{z}}_j=\mathbf{e}_j+\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}(\boldsymbol{\gamma}_j-\widetilde{\boldsymbol{\gamma}}_j)$. Since the components of $\mathbf{e}_j$ are i.i.d. sub-Gaussian random variables, it is easy to see that

$$\mathbb{P}(\max_{1\le j\le 2p}\|n^{-1/2}\mathbf{e}_j\|_2\ge C)\to 0$$

for some large enough constant $C>0$. Further, it follows from the sub-Gaussianity of X_j and the sparsity of $\boldsymbol{\gamma}_j$ and $\widetilde{\boldsymbol{\gamma}}_j$ that

$$\begin{aligned}
\max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}(\boldsymbol{\gamma}_j-\widetilde{\boldsymbol{\gamma}}_j)\|_2&\le Cm_n^{1/2}\sqrt{\frac{m_n\log p}{n}}\\
&\le Cm_n\sqrt{\frac{\log p}{n}}\to 0.
\end{aligned}$$

Thus, when $m_n\sqrt{\frac{\log p}{n}} \to 0$ we have

$$\mathbb{P}(n^{-1/2} \max_{1\le j\le 2p} \|\widetilde{\mathbf{z}}_j\|_2 \ge C) \to 0. \tag{A.166}$$

Similarly, based on Lemma 9 and the sparsity of $\widehat{\boldsymbol{\beta}}^{\text{init}}$ and $\boldsymbol{\beta}^{\text{aug}}$, it holds that with probability $1-o(1)$,

$$\begin{aligned}
&n^{-1/2}\|(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})(\widehat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}})\|_2 \\
&\le \max_{J':|J'|\lesssim s} \Big(\sum_{j\in J'} n^{-1}\|\widetilde{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_j^{\text{aug}}\|_2^2\Big)^{1/2} \cdot \|\widehat{\boldsymbol{\beta}}_{J'}^{\text{init}} - \boldsymbol{\beta}_{J'}^{\text{aug}}\|_2 \\
&\le Cs^{1/2}\Delta_n \cdot \Big(\sqrt{\frac{s\log p}{n}} + \Delta_n s\sqrt{\frac{\log p}{n}}\Big) \\
&\le C\Delta_n s\sqrt{\frac{\log p}{n}},
\end{aligned} \tag{A.167}$$

where the last inequality above holds due to $\Delta_n s^{1/2} \to 0$. Consequently, combining the above three inequalities shows that with probability $1-o(1)$,

$$P_{22} \le C\Delta_n s\sqrt{\frac{\log p}{n}}. \tag{A.168}$$

We now deal with term P_{23} in (A.158). In view of the Cauchy–Schwarz inequality and $\Delta_n m_n^{1/2} \to 0$, (A.186), (A.152), and (A.149), we can obtain that with probability $1-o(1)$,

$$\begin{aligned}
P_{23} &\le \max_{1\le j\le 2p} \frac{\|\widehat{\mathbf{z}}_j - \widetilde{\mathbf{z}}_j\|_2}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}} \cdot \|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}^{\text{init}} - \widehat{\boldsymbol{\beta}}^{\text{init}})\|_2 \\
&\le C\Delta_n m_n^{1/2} \cdot \Delta_n s\sqrt{\frac{\log p}{n}} \\
&\le C\Delta_n s\sqrt{\frac{\log p}{n}}.
\end{aligned} \tag{A.169}$$

As for term P_{24}, since $(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})\boldsymbol{\beta} = \mathbf{0}$ it follows that with probability $1-o(1)$,

$$\begin{aligned}
P_{24} &= \max_{1\le j\le 2p} \Big|\frac{(\widehat{\mathbf{z}}_j - \widetilde{\mathbf{z}}_j)^T(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})(\widehat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}}\Big| \\
&\le \max_{1\le j\le 2p} \frac{\|\widehat{\mathbf{z}}_j - \widetilde{\mathbf{z}}_j\|_2}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}} \cdot \|(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})(\widehat{\boldsymbol{\beta}}^{\text{init}} - \boldsymbol{\beta}^{\text{aug}})\|_2 \\
&\le C\Delta_n m_n^{1/2} \cdot \Delta_n s\sqrt{\frac{\log p}{n}} \\
&\le C\Delta_n s\sqrt{\frac{\log p}{n}},
\end{aligned} \tag{A.170}$$

where we have applied the bounds in (A.152), (A.167), and (A.186). Consequently, plugging (A.165), (A.168), (A.169), and (A.170) into (A.158) yields that with probability $1-o(1)$,

$$P_2 \le C\Delta_n s\sqrt{\frac{\log p}{n}}. \tag{A.171}$$

Now we proceed with dealing with term P_3. Note that

$$P_3 \le \max_{1\le j\le 2p} |\widehat{\mathbf{z}}_j^T(\mathbf{y}-\widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}})| \cdot \frac{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|}{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}\big| \cdot \big|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|}. \tag{A.172}$$

From (A.152) and (A.166), we can see that with probability $1-o(1)$,

$$\begin{aligned}\max_{1\le j\le 2p} n^{-1/2}\|\widehat{\mathbf{z}}_j\|_2 &\le \max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{z}}_j\|_2 + \max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j\|_2 \\ &\le C + Cm_n^{1/2}\Delta_n \le C.\end{aligned} \tag{A.173}$$

It follows from (A.152), Condition 6, and the sub-Gaussian distribution of $\widetilde{\mathbf{X}}_j^{\text{aug}}$ that with probability $1-o(1)$,

$$n^{-1}|(\widehat{\mathbf{z}}_j-\widetilde{\mathbf{z}}_j)^T\widetilde{\mathbf{X}}_j^{\text{aug}}| \le C\Delta_n m_n^{1/2}, \tag{A.174}$$

$$n^{-1}|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}})| \le C\Delta_n. \tag{A.175}$$

Then with the aid of (A.186), we can show that with probability $1-o(1)$,

$$\begin{aligned}&\min_{1\le j\le 2p} n^{-1}|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}| \\ &\ge \min_{1\le j\le 2p} n^{-1}|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}| - \max_{1\le j\le 2p}\Big(n^{-1}|(\widehat{\mathbf{z}}_j-\widetilde{\mathbf{z}}_j)^T\widetilde{\mathbf{X}}_j^{\text{aug}}| - n^{-1}|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}})|\Big) \\ &\ge C - Cm_n\Delta_n - C\Delta_n \\ &\ge C\end{aligned} \tag{A.176}$$

as $m_n\Delta_n \to 0$.

As for the second component on the right-hand side of (A.172) above, combining the results in (A.174), (A.175), and (A.176) gives that with probability $1-o(1)$,

$$\begin{aligned}\max_{1\le j\le 2p} \frac{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|}{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}\big| \cdot \big|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|} &\le \max_{1\le j\le 2p} \frac{\big|(\widetilde{\mathbf{z}}_j-\widehat{\mathbf{z}}_j)^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|}{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}\big| \cdot \big|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|} \\ &\quad + \max_{1\le j\le 2p} \frac{\big|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{X}}_j^{\text{aug}})\big|}{\big|\widehat{\mathbf{z}}_j^T\widehat{\mathbf{X}}_j^{\text{aug}}\big| \cdot \big|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}\big|} \\ &\le Cn^{-1}(m_n^{1/2}\Delta_n + \Delta_n).\end{aligned} \tag{A.177}$$

Regarding the first component on the right-hand side in (A.172), from $(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})\boldsymbol{\beta} = \mathbf{0}$ we can deduce that

$$\begin{aligned}\max_{1\le j\le 2p} n^{-1}\big|\widehat{\mathbf{z}}_j^T(\mathbf{y}-\widehat{\mathbf{X}}^{\text{aug}}\widehat{\boldsymbol{\beta}}^{\text{init}})\big| &\le \max_{1\le j\le 2p} n^{-1}\big|\widehat{\mathbf{z}}_j^T\boldsymbol{\varepsilon}\big| + \max_{1\le j\le 2p} n^{-1}\big|\widehat{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\text{aug}}(\boldsymbol{\beta}^{\text{aug}}-\widehat{\boldsymbol{\beta}}^{\text{init}})\big| \\ &\quad + \max_{1\le j\le 2p} n^{-1}\big|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}^{\text{aug}}-\widehat{\mathbf{X}}^{\text{aug}})\widehat{\boldsymbol{\beta}}^{\text{init}}\big|.\end{aligned}$$

Since $\boldsymbol{\varepsilon} \overset{d}{\sim} N(\mathbf{0}, \sigma^2 I_n)$, it is easy to see that for the standard normal random variable Z,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le j\le 2p} n^{-1}\big|\widehat{\mathbf{z}}_j^T\boldsymbol{\varepsilon}\big| > C\sqrt{\frac{\log p}{n}}\Big) &= \mathbb{P}\Big(\max_{1\le j\le 2p} n^{-1}\|\widehat{\mathbf{z}}_j\|_2 \cdot |Z| > C\sqrt{\frac{\log p}{n}}\Big) \\ &\le \mathbb{P}(|Z| > C\sqrt{\log p}) \to 0.\end{aligned}$$

Further, by Lemma 9, the sub-Gaussianity of X_j, and the sparsity of $\boldsymbol{\beta}^{\mathrm{aug}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{init}}$, we can obtain that with probability $1-o(1)$,

$$\begin{aligned} n^{-1}|\widehat{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}(\boldsymbol{\beta}^{\mathrm{aug}}-\widehat{\boldsymbol{\beta}}^{\mathrm{init}})| &\le n^{-1}|\widehat{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}(\boldsymbol{\beta}^{\mathrm{aug}}-\widetilde{\boldsymbol{\beta}}^{\mathrm{init}})| + n^{-1}|\widehat{\mathbf{z}}_j^T\widetilde{\mathbf{X}}^{\mathrm{aug}}(\widetilde{\boldsymbol{\beta}}^{\mathrm{init}}-\widehat{\boldsymbol{\beta}}^{\mathrm{init}})| \\ &\le C\left(\sqrt{\frac{s\log p}{n}}+\Delta_n s\sqrt{\frac{\log p}{n}}\right) \\ &\le C\sqrt{\frac{s\log p}{n}}. \end{aligned}$$

Similarly, since $(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})\boldsymbol{\beta}=\mathbf{0}$, it holds that with probability $1-o(1)$,

$$\begin{aligned} n^{-1}|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})\widehat{\boldsymbol{\beta}}^{\mathrm{init}}| &= n^{-1}|\widehat{\mathbf{z}}_j^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}-\widehat{\mathbf{X}}^{\mathrm{aug}})(\widehat{\boldsymbol{\beta}}^{\mathrm{init}}-\boldsymbol{\beta}^{\mathrm{aug}})| \\ &\le C\Delta_n s^{1/2}\cdot\sqrt{\frac{s\log p}{n}} \\ &\le C\sqrt{\frac{s\log p}{n}}. \end{aligned}$$

Consequently, by $m_n \lesssim s$ in Condition 11 we have that with probability $1-o(1)$,

$$P_3 \le Cm_n^{1/2}\Delta_n\cdot\sqrt{\frac{s\log p}{n}} \le C\Delta_n s\sqrt{\frac{\log p}{n}}. \tag{A.178}$$

Finally, a combination of (A.153), (A.157), (A.171), and (A.178) establishes (A.140). This completes the proof of Lemma 11.

B.10. Proof of Lemma 12. Using the definitions of $\widetilde{W}_j$ and w_j and the triangle inequality, we see that

$$\begin{aligned} &\sum_{j=1}^{p}\mathbb{P}(|\widetilde{W}_j-w_j|\ge C\sqrt{n^{-1}\log p}) \\ &\le \sum_{j=1}^{p}\mathbb{P}\Big(\sqrt{n}\big||\widetilde{\beta}_j-\beta_j|-|\widetilde{\beta}_{j+p}-\beta_{j+p}|\big|\ge C\sqrt{\log p}\Big) \qquad \text{(A.179)} \\ &\le \sum_{j=1}^{p}\left[\mathbb{P}\Big(\sqrt{n}|\widetilde{\beta}_j-\beta_j|\ge C\sqrt{\log p}/2\Big)+\mathbb{P}\Big(\sqrt{n}|\widetilde{\beta}_{j+p}-\beta_{j+p}|\ge C\sqrt{\log p}/2\Big)\right]. \end{aligned}$$

The main idea of the proof is to exploit the decomposition in (A.11) and the observation that the main term therein follows the normal distribution. Let us start with bounding the error term in (A.11). We claim that with probability $1-o(p^{-1})$,

$$\max_{1\le j\le 2p}\left|\sum_{k\ne j}\frac{\sqrt{n}\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_k^{\mathrm{aug}}(\beta_k^{\mathrm{aug}}-\widetilde{\beta}_k^{\mathrm{init}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\mathrm{aug}}}\right| \le \frac{Cm_n^{1/2}s\log p}{\sqrt{n}}. \tag{A.180}$$

From the fact that $\beta^{\text{aug}}_{j+p} = 0$ for $1 \le j \le p$ and the bound in (A.180), since $\frac{m_n^{1/2} s \log p}{\sqrt{n}} \ll \sqrt{\log p}$ we can deduce through the union bound that

$$\begin{aligned}
&\sum_{j=1}^{p} \mathbb{P}\Big(\sqrt{n}|\widetilde{\beta}_j - \beta_j| \ge C\sqrt{\log p}/2\Big) \\
&\le \sum_{j=1}^{p} \mathbb{P}\Big(\frac{|\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}|}{\|\widetilde{\mathbf{z}}_j\|_2} \cdot \sqrt{n}\tau_j \ge C\sqrt{\log p}/3\Big) \\
&\quad + \sum_{j=1}^{p} \mathbb{P}\Big(\max_{1\le j\le 2p} \Big|\sum_{k\ne j} \frac{\sqrt{n}\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\text{aug}} (\beta_k - \widetilde{\beta}_k^{\text{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\text{aug}}}\Big| > \frac{C m_n^{1/2} s \log p}{\sqrt{n}}\Big) \\
&\le \sum_{j=1}^{p} \mathbb{P}\Big(\frac{|\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}|}{\|\widetilde{\mathbf{z}}_j\|_2} \cdot \sqrt{n}\tau_j \ge C\sqrt{\log p}/3\Big) + o(1).
\end{aligned} \tag{A.181}$$

Recall the result (A.265) in Lemma 10 and that $\frac{\mathbf{z}_j^T \boldsymbol{\varepsilon}}{\|\mathbf{z}_j\|_2} \sim N(0, \sigma^2)$. As $\frac{m_n \log p}{n} = o(1)$, it holds that for some large constant $C > 0$,

$$\begin{aligned}
&\sum_{j=1}^{p} \mathbb{P}\Big(\frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2} \cdot \sqrt{n}\tau_j \ge C\sqrt{\log p}/3\Big) \\
&\le \sum_{j=1}^{p} \mathbb{P}\Big(\frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2} \ge \widetilde{C}\sqrt{\log p}\Big) \\
&= p \exp\{-\widetilde{C}^2 \log p/2\} \to 0.
\end{aligned}$$

Similarly, we can show that

$$\sum_{j=1}^{p} \mathbb{P}\Big(\frac{\widetilde{\mathbf{z}}_{j+p}^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_{j+p}\|_2} \cdot \sqrt{n}\tau_j \ge C\sqrt{\log p}\Big) \to 0. \tag{A.182}$$

Plugging the two inequalities above into (A.179) leads to the desired result in Lemma 12. It remains to establish (A.180).

Proof of (A.180). Observe that for $k \ne j$,

$$\begin{aligned}
n^{-1}\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\text{aug}} &= n^{-1}(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}} \widetilde{\boldsymbol{\gamma}}_j)^T \widetilde{\mathbf{X}}_k^{\text{aug}} \\
&= n^{-1}\mathbf{e}_j^T \widetilde{\mathbf{X}}_k^{\text{aug}} + n^{-1}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)^T (\widetilde{\mathbf{X}}_{-j}^{\text{aug}})^T \widetilde{\mathbf{X}}_k^{\text{aug}}.
\end{aligned} \tag{A.183}$$

Since $\mathbf{e}_j$ and $\widetilde{\mathbf{X}}_k^{\text{aug}}$ are uncorrelated, it follows from the sub-Gaussian assumption in Condition 13 that for some constant $C > 0$,

$$\mathbb{P}\Big(n^{-1}|\mathbf{e}_j^T \widetilde{\mathbf{X}}_k^{\text{aug}}| \ge C\sqrt{\frac{\log p}{n}}\Big) \le 2p^{-3}.$$

In light of lemma 10 and the sub-Gaussian assumption on $\widetilde{\mathbf{X}}_j$, we can deduce that with probability $1 - o(p^{-3})$,

$$\begin{aligned}
|n^{-1}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)^T (\widetilde{\mathbf{X}}_{-j}^{\text{aug}})^T \widetilde{\mathbf{X}}_k^{\text{aug}}| &\le \|n^{-1/2}\widetilde{\mathbf{X}}_{-j}^{\text{aug}}(\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j)\|_2 \|n^{-1/2}\widetilde{\mathbf{X}}_k^{\text{aug}}\|_2 \\
&\le C\sqrt{\frac{m_n \log p}{n}}.
\end{aligned} \tag{A.184}$$

Plugging the above two results into (A.183), when $m_n \log p = o(n)$ an application of the union bound shows that with probability $1 - o(p^{-1})$,

$$\begin{aligned} \max_{1\le j\le p} \max_{k\neq j} n^{-1} |\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\mathrm{aug}}| &\le C\sqrt{\frac{\log p}{n}} + C\sqrt{\frac{m_n \log p}{n}} \\ &\le C\sqrt{\frac{m_n \log p}{n}}. \end{aligned} \tag{A.185}$$

Similarly, when $\sqrt{\frac{\log p}{n}} = o(1)$, we can show that there exists some constant $C > 0$ such that with probability $1 - o(p^{-1})$,

$$\min_{1\le j\le p} n^{-1} \widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\mathrm{aug}} \ge C. \tag{A.186}$$

Consequently, plugging (A.185), (A.186), and (A.259) into Lemma 9 yields that with probability $1 - o(p^{-1})$,

$$\begin{aligned} &\max_{1\le j\le p} \left| \sum_{k\neq j} \frac{\sqrt{n}\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\mathrm{aug}} (\beta_k - \widetilde{\beta}_k^{\mathrm{init}})}{\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\mathrm{aug}}} \right| \\ &\le \sqrt{n} \frac{\max_{1\le j\le p} \max_{k\neq j} |\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_k^{\mathrm{aug}}|}{\min_{1\le j\le p} |\widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{X}}_j^{\mathrm{aug}}|} \cdot \|\boldsymbol{\beta}^{\mathrm{aug}} - \widetilde{\boldsymbol{\beta}}^{\mathrm{init}}\|_1 \\ &\le C\sqrt{m_n \log p} \cdot s\sqrt{\frac{\log p}{n}} = \frac{C m_n^{1/2} s \log p}{\sqrt{n}}, \end{aligned} \tag{A.187}$$

which establishes (A.180). This concludes the proof of Lemma 12.

B.11. Proof of Lemma 13. The intuition of the proof is that the sparsity of $\boldsymbol{\Omega}^A$ implies the weak dependence among the components of the knockoff statistic vector $\widetilde{\mathbf{W}} = (\widetilde{W}_1, \cdots, \widetilde{W}_p)$, which entails the weak dependence among the indicator functions $\mathbb{1}(\widetilde{W}_j > t)$'s. For $1 \le j \le p$, let us define

$$N_j = \{l \in \mathcal{H}_0 : \boldsymbol{\Omega}_{j,l}^A \neq 0\}.$$

From the sparsity assumption on $\boldsymbol{\Omega}^A$ in Condition 11, we see that $|N_j| \le m_n$ for any $1 \le j \le p$. Then we can obtain through expanding the variance that

$$\begin{aligned} \mathrm{Var}\Big(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j > t)\Big) &= \sum_{j\in\mathcal{H}_0} \sum_{\substack{l\in N_j^c\cap\mathcal{H}_0 \\ l\neq j}} \Big(\mathbb{P}(\widetilde{W}_j \ge t, \widetilde{W}_l \ge t) - \mathbb{P}(\widetilde{W}_j \ge t)\mathbb{P}(\widetilde{W}_l \ge t)\Big) \\ &\quad + \sum_{j\in\mathcal{H}_0} \sum_{l\in N_j\cup\{j\}} \Big(\mathbb{P}(\widetilde{W}_j \ge t, \widetilde{W}_l \ge t) - \mathbb{P}(\widetilde{W}_j \ge t)\mathbb{P}(\widetilde{W}_l \ge t)\Big) \\ &:= V_1(t) + V_2(t). \end{aligned} \tag{A.188}$$

We will deal with terms $V_1(t)$ and $V_2(t)$ above separately.

Regarding the second term $V_2(t)$, it follows from $|N_j \cup \{j\}| \leq m_n + 1$ that

$$\begin{aligned}\sup_{t\in(0,G^{-1}(\frac{c_1 q a_n}{p})]} \frac{V_2(t)}{p_0 G(t)} &\leq \sup_{t\in(0,G^{-1}(\frac{c_1 q a_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0}\sum_{l\in N_j\cup\{j\}}\mathbb{P}(\widetilde{W}_j \geq t)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)} \\ &\leq \sup_{t\in(0,G^{-1}(\frac{c_1 q a_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0}(m_n+1)\mathbb{P}(\widetilde{W}_j \geq t)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)} \\ &\leq m_n + 1.\end{aligned} \tag{A.189}$$

We claim that as $\frac{m_n^{1/2} s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$,

$$(\log p)^{1/\gamma} \sup_{t\in(0,G^{-1}(\frac{c_1 q a_n}{p})]} \frac{V_1(t)}{[p_0 G(t)]^2} \to 0. \tag{A.190}$$

Therefore, combining (A.188), (A.189), and (A.190) leads to the desired result of Lemma 13. It remains to establish (A.190).

Proof of (A.190). Let $\{\eta_j\}_{j=1}^p$ be a sequence of independent random variables with η_j having density function given by

$$\begin{aligned}h_j(t) = &\frac{\sqrt{2}}{\sqrt{\pi}a_j}\big[1-\Phi(b_j^{-1}t)\big]\exp\big\{-t^2/(2v_j^2)\big\} \\ &+ \frac{\sqrt{2}}{\sqrt{\pi}b_j}\big[1-\Phi(v_j^{-1}t)\big]\exp\big\{-t^2/(2b_j^2)\big\},\end{aligned} \tag{A.191}$$

where $v_j = \sqrt{2(\mathbb{E}e_j^2)^{-1}(1-\mathrm{corr}(e_j, e_{j+p}))}$ and $b_j = \sqrt{2(\mathbb{E}e_j^2)^{-1}(1+\mathrm{corr}(e_j, e_{j+p}))}$. For $1 \leq j \leq 2p$, let us define $\xi_j = \sqrt{n}\tau_j \cdot \frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2}$. The essential step in the proof is to show that for $l \in N_j^c \cap \mathcal{H}_0$,

$$(|\xi_j| - |\xi_{j+p}|, |\xi_l| - |\xi_{l+p}|) \xrightarrow{d} (\eta_j, \eta_l).$$

We proceed with proving such result. Define $\delta_n = C\frac{m_n^{1/2} s\log p}{\sqrt{n}}$. We claim that for $l \neq j$ and $l \in N_j^c \cap \mathcal{H}_0$,

$$\begin{aligned}&\mathbb{P}(\widetilde{W}_j \geq t, \widetilde{W}_l \geq t) \\ &\leq \mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t - \delta_n)\bigg(1 + O\bigg(\sqrt{\frac{m_n(\log p)^3}{n}}\bigg)\bigg) + O(p^{-3}),\end{aligned} \tag{A.192}$$

$$\begin{aligned}&\mathbb{P}(\widetilde{W}_j \geq t, \widetilde{W}_l \geq t) \\ &\geq \mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t + \delta_n)\bigg(1 + O\bigg(\sqrt{\frac{m_n(\log p)^3}{n}}\bigg)\bigg) + O(p^{-3}),\end{aligned} \tag{A.193}$$

$$\mathbb{P}(\widetilde{W}_j \geq t) \geq \mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)\bigg(1 + O\bigg(\sqrt{\frac{m_n(\log p)^3}{n}}\bigg)\bigg) + O(p^{-3}), \tag{A.194}$$

$$\mathbb{P}(\widetilde{W}_j \geq t) \leq \mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)\Big(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\Big) + O(p^{-3}). \tag{A.195}$$

The proofs for (A.192)–(A.195) above are analogous. Without loss of generality, we will present only the proof of (A.192) and postpone it to the end of the proof for Lemma 13. In view of (A.192)–(A.195) above and the definition of $V_1(t)$ in (A.188), we can deduce that

(A.196)

$$\begin{aligned}
V_1(t) &= \sum_{j\in\mathcal{H}_0}\sum_{\substack{l\in N_j^c\cap\mathcal{H}_0\\ l\neq j}}\Big(\mathbb{P}(\widetilde{W}_j \geq t, \widetilde{W}_l \geq t) - \mathbb{P}(\widetilde{W}_j \geq t)\mathbb{P}(\widetilde{W}_l \geq t)\Big)\\
&\leq \sum_{j\in\mathcal{H}_0}\sum_{l\neq j}\Bigg\{\mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t - \delta_n)\Big(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\\
&\quad - \mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)\mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)\Big(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\Big)\Bigg\} + O(p^{-1})\\
&= \sum_{j\in\mathcal{H}_0}\sum_{l\neq j}\mathbb{P}(\sqrt{n}t - \delta_n \leq \eta_j \leq \sqrt{n}t + \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t - \delta_n)\\
&\quad + \sum_{j\in\mathcal{H}_0}\sum_{l\neq j}\mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)\mathbb{P}(\sqrt{n}t - \delta_n \leq \eta_l \leq \sqrt{n}t + \delta_n)\\
&\quad + \sum_{j\in\mathcal{H}_0}\sum_{l\neq j}\mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t - \delta_n)\cdot O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big) + O(p^{-1})\\
&:= V_{11}(t) + V_{12}(t) + V_{13}(t) + O(p^{-1}).
\end{aligned}$$

Recall that $p_0G(t) = \sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$. Then it follows from the definition of $V_{11}(t)$ and (A.194) that

$$\frac{V_{11}(t)}{[p_0G(t)]^2} \leq \frac{\sum_{j\in\mathcal{H}_0}\sum_{l\neq j}\mathbb{P}(\sqrt{n}t - \delta_n \leq \eta_j \leq \sqrt{n}t + \delta_n)\mathbb{P}(\eta_l \geq \sqrt{n}t - \delta_n)}{\big[\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)\big(1 + O(\sqrt{\frac{m_n(\log p)^3}{n}})\big) + O(p^{-2})\big]^2}. \tag{A.197}$$

We will consider two ranges $t \in (0, 4n^{-1/2}\max_{1\leq j\leq p}(v_j \vee b_j))$ and $t \in [4n^{-1/2}\max_{1\leq j\leq p}(v_j \vee b_j), G^{-1}(\frac{c_1 q a_n}{p})]$ separately. For the first range $t \in (0, 4n^{-1/2}\max_{1\leq j\leq p}(v_j \vee b_j))$, we can see that $\sqrt{n}t$ is upper bounded by a constant. Since $\delta_n = o(1)$ by the assumption that $\frac{m_n^{1/2}s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$, it follows that $\sqrt{n}t + \delta_n$ and $\sqrt{n}t - \delta_n$ are both of a constant order. Hence, by the definition of the density function $h_j(\cdot)$ of η_j shown in (A.191), $\max_{1\leq j\leq p}h_j(u)$ is bounded by a constant for $u \in [\sqrt{n}t - \delta_n, \sqrt{n}t + \delta_n]$, and

$$C_1 \leq \min_{1\leq j\leq p}\mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n) \leq \max_{1\leq j\leq p}\mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n) \leq C_2$$

for some positive constants $C_1 < C_2$. Thus, it is easy to see that

$$\begin{aligned}
&\sup_{t\in(0,4n^{-1/2}\max_{1\leq j\leq p}(v_j\vee b_j))}\frac{V_{11}(t)}{[p_0G(t)]^2}\\
&\leq C\frac{p_0^2\delta_n\max_{1\leq j\leq p}\sup_{u\in[\sqrt{n}t-\delta_n,\sqrt{n}t+\delta_n]}h_j(u)\max_{1\leq j\leq p}\mathbb{P}(\eta_j \geq \sqrt{n}t - \delta_n)}{p_0^2[\min_{1\leq j\leq p}\mathbb{P}(\eta_j \geq \sqrt{n}t + \delta_n)]^2} \qquad \text{(A.198)}\\
&\leq C\delta_n = C\frac{m_n^{1/2}s\log p}{\sqrt{n}}.
\end{aligned}$$

We proceed with considering the second range $t \in [4n^{-1/2}\max_{1\le j\le p}(v_j \vee b_j), G^{-1}(\frac{c_1 q a_n}{p}))$. An application of similar arguments as for (A.135) shows that

$$\begin{aligned}(\text{A.199}) \quad &\max_{1\le j\le p} \sup_{t\in[4n^{-1/2}\max_{1\le j\le p}(v_j\vee b_j)), G^{-1}(\frac{c_1 q a_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\sqrt{n}t-\delta_n\le\eta_j\le\sqrt{n}t+\delta_n)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n)} \\ &\le C\sqrt{n}G^{-1}(\frac{c_1 q a_n}{p})\cdot\delta_n.\end{aligned}$$

Moreover, it follows from plugging $t = G^{-1}(\frac{c_1 q a_n}{p})$ into (A.195) and taking summation over $j \in \mathcal{H}_0$ that

$$\begin{aligned}(\text{A.200}) \quad \frac{c_1 q a_n p_0}{p} &\le \sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}G^{-1}(\frac{c_1 q a_n}{p})-\delta_n)\Big(1+O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\Big) \\ &\quad + O(p^{-3}).\end{aligned}$$

Then from the density function $h_j(t)$ for η_j, we can obtain through some direct calculations that

$$(\text{A.201}) \qquad \mathbb{P}(\eta_j \ge t) = 2[1-\Phi(v_j^{-1}t)][1-\Phi(b_j^{-1}t)].$$

Further, combining (A.200) and (A.201) yields that

$$G^{-1}(\frac{c_1 q a_n}{p}) = O(\sqrt{\frac{\log p}{n}}).$$

Substituting this bound into (A.199) implies that

$$\begin{aligned}(\text{A.202}) \quad &\max_{1\le j\le p} \sup_{t\in[4n^{-1/2}\max_{1\le j\le p}(v_j\vee b_j)), G^{-1}(\frac{c_1 q a_n}{p})]} \frac{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\sqrt{n}t-\delta_n\le\eta_j\le\sqrt{n}t+\delta_n)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n)} \\ &\le C\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}},\end{aligned}$$

where in the last inequality above we have utilized the definition of δ_n. Thus as $\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}} \to 0$, it holds that

$$\begin{aligned}(\text{A.203}) \quad &\max_{1\le j\le p} \sup_{t\in[4n^{-1/2}\max_{1\le j\le p}(v_j\vee b_j)), G^{-1}(\frac{c_1 q a_n}{p})]} \Big|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t-\delta_n)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n)}-1\Big| \\ &\le C\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}} \to 0.\end{aligned}$$

Since $p_0 G(t) \ge c_1 q a_n p_0/p \to \infty$ for $0 \le t \le G^{-1}(\frac{c_1 q a_n}{p})$, it follows from taking summation over $j \in \mathcal{H}_0$ on both sides of (A.195) that as $m_n^{1/2}(\log p)^{3/2}/\sqrt{n} \to 0$,

$$\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t-\delta_n) \ge C\Big(\frac{c_1 q a_n p_0}{p}+O(p^{-2})\Big)\to\infty,$$

which along with (A.203) implies that

$$\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n) \ge C\Big(\frac{c_1 q a_n p_0}{p}+O(p^{-2})\Big)\to\infty.$$

Combining this with (A.202), we can further bound the ratio in (A.197) in the second range of $t \in [4n^{-1/2}\max_{1\le j\le p}(v_j \vee b_j)), G^{-1}(\frac{c_1 q a_n}{p}))$ as

$$\begin{aligned}
&\sup_{t\in[4n^{-1/2}\max_{1\le j\le p}(v_j\vee b_j)),G^{-1}(\frac{c_1qa_n}{p})]} \frac{V_{11}(t)}{[p_0G(t)]^2} \\
&\le \left\{ \frac{\left[\sum_{j\in\mathcal{H}_0}\mathbb{P}(\sqrt{n}t-\delta_n\le\eta_j\le\sqrt{n}t+\delta_n)\right]^2}{\left[\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n)\right]^2} + \frac{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\sqrt{n}t-\delta_n\le\eta_j\le\sqrt{n}t+\delta_n)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j\ge\sqrt{n}t+\delta_n)} \right\} \\
&\quad \times \left(1+O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}+p^{-2}\Big)\right) \\
&\le C\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}}.
\end{aligned}$$

Hence, we see from the above result and (A.198) that

$$\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p}))} \frac{V_{11}(t)}{[p_0G(t)]^2} \le C\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}}. \tag{A.204}$$

In a similar manner, we can deduce that

$$\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p}))} \frac{V_{12}(t)}{[p_0G(t)]^2} \le C\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}} \tag{A.205}$$

and

$$\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p}))} \frac{V_{13}(t)}{[p_0G(t)]^2} \le C\sqrt{\frac{m_n(\log p)^3}{n}}. \tag{A.206}$$

Combining (A.196) and (A.204)–(A.206) yields (A.190) as $\frac{m_n^{1/2}s(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$. This completes the proof of (A.190). It remains to establish (A.192).

Proof of (A.192). Note that for $j \in \mathcal{H}_0$, it holds that $\beta_j^{\mathrm{aug}} = \beta_{j+p}^{\mathrm{aug}} = 0$ under the setting of the linear model. Then it follows that

$$\widetilde{W}_j = |\widetilde{\beta}_j| - |\widetilde{\beta}_{j+p}| = |\widetilde{\beta}_j - \beta_j^{\mathrm{aug}}| - |\widetilde{\beta}_{j+p} - \beta_{j+p}^{\mathrm{aug}}|.$$

For $1 \le j \le 2p$, let us define $\xi_j = \sqrt{n}\tau_j \cdot \frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2}$. In view of the expression in (A.11) and the bound of the remainder term established in (A.180), an application of the total probability inequality gives that

$$\begin{aligned}
\mathbb{P}\big(\widetilde{W}_j \ge t, \widetilde{W}_l \ge t\big) &\le \mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \ge \sqrt{n}t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \ge \sqrt{n}t - \delta_n\big) \\
&\quad + \mathbb{P}\left(\max_{1\le j\le 2p}\left|\sum_{k\ne j}\frac{\sqrt{n}\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_k^{\mathrm{aug}}(\beta_k - \widetilde{\beta}_k^{\mathrm{init}})}{\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\mathrm{aug}}}\right| > \delta_n\right) \qquad \text{(A.207)} \\
&= \mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \ge \sqrt{n}t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \ge \sqrt{n}t - \delta_n\big) + O(p^{-3}).
\end{aligned}$$

It suffices to consider probability $\mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \geq t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \geq t - \delta_n\big)$ for $t \in (0, \sqrt{n}G^{-1}(\frac{c_1 q a_n}{p})]$. A useful observation is that

$$
\begin{aligned}
&\mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \geq t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \geq t - \delta_n\big) \\
&\leq \mathbb{P}\Big(|\xi_j| - |\xi_{j+p}| \geq t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \geq t - \delta_n, \\
&\qquad \max\{|\xi_j|, |\xi_{j+p}|, |\xi_l|, |\xi_{l+p}|\} \leq C\sqrt{\log p}\Big) \\
&\quad + \mathbb{P}\big(\max\{|\xi_j|, |\xi_{j+p}|, |\xi_l|, |\xi_{l+p}|\} > C\sqrt{\log p}\big) \\
&:= P_1 + P_2.
\end{aligned} \tag{A.208}
$$

We will consider terms P_1 and P_2 above separately.

Let us first deal with term P_2. From the definition of ξ_j, (A.265) in Lemma 10, and the fact that $\frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2} \overset{d}{\sim} N(0,1)$, we can obtain through the union bound that as $\frac{m_n \log p}{n} \to 0$ and for some large constant $C > 4(\mathbb{E}e_j^2)^{-1/2}$,

$$
\begin{aligned}
\mathbb{P}(|\xi_j| \geq C\sqrt{\log p}) &\leq \mathbb{P}\left(\left|\frac{\widetilde{\mathbf{z}}_j^T \boldsymbol{\varepsilon}}{\|\widetilde{\mathbf{z}}_j\|_2}\right| \geq 2C(\mathbb{E}e_j^2)^{1/2}\sqrt{\log p}/3\right) \\
&\quad + \mathbb{P}(\sqrt{n}\tau_j \geq 3(\mathbb{E}e_j^2)^{-1/2}/2) \\
&= O(p^{-3}).
\end{aligned}
$$

Hence, the inequality above implies that

$$
P_2 = O(p^{-3}). \tag{A.209}
$$

We next proceed with analyzing term P_1. Given $\widetilde{\mathbf{X}}^{\text{aug}}$, denote by $f_{\xi,\xi_{j+p}}(x,y)$ the density of (ξ_i, ξ_{j+p}) and $f_{\xi_l,\xi_{l+p}|(\xi_j,\xi_{j+p})}(u,w|x,y)$ the conditional density of $(\xi_l, \xi_{l+p})|(\xi_j, \xi_{j+p})$. Then probability P_2 can be written as

$$
\begin{aligned}
&\mathbb{P}\Big(|\xi_j| - |\xi_{j+p}| \geq t - \delta_n,\, |\xi_l| - |\xi_{l+p}| \geq t - \delta_n, \\
&\qquad \max\{|\xi_j|, |\xi_{j+p}|, |\xi_l|, |\xi_{l+p}| \leq C\sqrt{\log p}\Big) \\
&= \mathbb{E}_{\widetilde{\mathbf{X}}^{\text{aug}}}\Bigg[\int_{\substack{|x|-|y|\geq t-\delta_n \\ |x|\leq C\sqrt{\log p} \\ |y|\leq C\sqrt{\log p}}} f_{\xi,\xi_{j+p}}(x,y) \\
&\qquad \cdot \int_{\substack{|u|-|w|\geq t-\delta_n \\ |u|\leq C\sqrt{\log p} \\ |w|\leq C\sqrt{\log p}}} f_{\xi_l,\xi_{l+p}|(\xi_j,\xi_{j+p})}(u,w|x,y)\, du\, dv\, dx\, dy\Bigg].
\end{aligned} \tag{A.210}
$$

Since $\boldsymbol{\varepsilon} \overset{d}{\sim} N(\mathbf{0}, I_n)$ and is independent of $\widetilde{\mathbf{X}}^{\text{aug}}$, it is easy to see that for $j \neq l$, conditional on $\widetilde{\mathbf{X}}^{\text{aug}}$ we have

$$
(\xi_j, \xi_{j+p}, \xi_l, \xi_{l+p})^T \big|\widetilde{\mathbf{X}}^{\text{aug}} \overset{d}{\sim} N(\mathbf{0}, \mathbf{V}),
$$

where the covariance matrix is given by $\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11}\mathbf{V}_{12} \\ \mathbf{V}_{21}\mathbf{V}_{22} \end{pmatrix}$ with

$$\mathbf{V}_{11} = \begin{pmatrix} n\tau_j^2 & \frac{n\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_{j+p}}{|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}||\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{X}}_{j+p}^{\text{aug}}|} \\ \frac{n\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_{j+p}}{|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}||\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{X}}_{j+p}^{\text{aug}}|} & n\tau_{j+p}^2 \end{pmatrix},$$

$$\mathbf{V}_{12} = \mathbf{V}_{21}^T = \begin{pmatrix} \frac{n\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_l}{|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}||\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{X}}_l^{\text{aug}}|} & \frac{n\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_{l+p}}{|\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}}||\widetilde{\mathbf{z}}_{l+p}^T\widetilde{\mathbf{X}}_{l+p}^{\text{aug}}|} \\ \frac{n\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{z}}_{j+p}}{|\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{X}}_l^{\text{aug}}||\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{X}}_{j+p}^{\text{aug}}|} & \frac{n\widetilde{\mathbf{z}}_{l+p}^T\widetilde{\mathbf{z}}_{j+p}}{|\widetilde{\mathbf{z}}_{l+p}^T\widetilde{\mathbf{X}}_{l+p}^{\text{aug}}||\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{X}}_{j+p}^{\text{aug}}|} \end{pmatrix},$$

$$\mathbf{V}_{22} = \begin{pmatrix} n\tau_l^2 & \frac{n\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{z}}_{l+p}}{|\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{X}}_l^{\text{aug}}||\widetilde{\mathbf{z}}_{l+p}^T\widetilde{\mathbf{X}}_{l+p}^{\text{aug}}|} \\ \frac{n\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{z}}_{l+p}}{|\widetilde{\mathbf{z}}_l^T\widetilde{\mathbf{X}}_l^{\text{aug}}||\widetilde{\mathbf{z}}_{l+p}^T\widetilde{\mathbf{X}}_{l+p}^{\text{aug}}|} & n\tau_{l+p}^2 \end{pmatrix}.$$

It follows from the conditional distribution of the multivariate normal distribution that given $\widetilde{\mathbf{X}}^{\text{aug}}$,

$$\begin{aligned} & f_{\xi_l,\xi_{l+p}|(\xi_j,\xi_{j+p})}(u,v|x,y) \\ &= \frac{1}{2\pi|\mathbf{V}_{22}-\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}|^{1/2}}\times \\ & \exp\left\{-\frac{1}{2}\left[\begin{pmatrix}u\\v\end{pmatrix}-\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right]^T(\mathbf{V}_{22}-\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})^{-1}\right. \\ & \left.\cdot\left[\begin{pmatrix}u\\v\end{pmatrix}-\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\begin{pmatrix}x\\y\end{pmatrix}\right]\right\}. \end{aligned} \tag{A.211}$$

For $l \neq j$ and $l \in N_j^c$, it holds that

$$\mathbb{E}(e_j, e_l) = \frac{\mathbf{\Omega}_{j,l}^A}{\mathbf{\Omega}_{j,j}^A\mathbf{\Omega}_{l,l}^A} = 0.$$

Since $\mathbf{\Omega}_{j,l}^A = \mathbf{\Omega}_{j,l+p}^A = \mathbf{\Omega}_{j+p,l}^A = \mathbf{\Omega}_{j+p,l+p}^A$ due to the symmetric structure of $\mathbf{\Omega}$, we also have

$$\mathbb{E}(e_j, e_{l+p}) = \mathbb{E}(e_{j+p}, e_l) = \mathbb{E}(e_{j+p}, e_{l+p}) = 0$$

for $l \neq j$ and $l \in N_j^c$. Then it follows from (A.266) in Lemma 10 that for $l \neq j$ and $l \in N_j^c$, with probability $1 - O(p^{-3})$

$$\begin{aligned} & n^{-1}\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_l \leq C\sqrt{\frac{m_n\log p}{n}},\ \ n^{-1}\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{z}}_{l+p} \leq C\sqrt{\frac{m_n\log p}{n}}, \\ & n^{-1}\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{z}}_l \leq C\sqrt{\frac{m_n\log p}{n}},\ \ n^{-1}\widetilde{\mathbf{z}}_{j+p}^T\widetilde{\mathbf{z}}_{l+p} \leq C\sqrt{\frac{m_n\log p}{n}}. \end{aligned} \tag{A.212}$$

Similarly, for $1 \leq j \leq 2p$ we can show that with probability $1 - O(p^{-3})$,

$$n^{-1}\widetilde{\mathbf{z}}_j^T\widetilde{\mathbf{X}}_j^{\text{aug}} \geq C. \tag{A.213}$$

Then from (A.212), (A.213), and the definition of $\mathbf{V}_{12}$, we can obtain that with probability $1 - O(p^{-3})$,

$$\|\mathbf{V}_{12}\|_{\max} \leq C\sqrt{\frac{m_n\log p}{n}}. \tag{A.214}$$

We have shown in (A.214) that

$$\|\mathbf{V}_{12}\|_{\max} \leq C\sqrt{\frac{m_n \log p}{n}}$$

with probability $1 - O(p^{-3})$. Similarly, when $\mathbb{E}e_j^2\mathbb{E}e_{j+p}^2 - (\mathbb{E}[e_j e_{j+p}])^2 > C$ for some constant $C > 0$, it can be shown that $|V_{11}| \geq C$ and $|V_{22}| \geq C$ with probability $1 - O(p^{-3})$. Let us define an event

$$\mathcal{C} = \Big\{\widetilde{\mathbf{X}}^{\text{aug}} : \|\mathbf{V}_{12}\|_{\max} \leq C_1\sqrt{\frac{m_n \log p}{n}}, |\mathbf{V}_{22}| \geq C_2, |\mathbf{V}_{11}| \geq C_2,$$
$$\|\mathbf{V}_{11}\|_{\max} \leq C_3, \|\mathbf{V}_{22}\|_{\max} \leq C_3\Big\}.$$

We have shown that $\mathbb{P}(\mathcal{C}) \geq 1 - O(p^{-3})$. Then it is straightforward to see that conditional on event $\mathcal{C}$, we have

$$\frac{1}{2\pi|\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}|^{1/2}} = \frac{1}{2\pi|V_{22}|^{-1/2}}\Big(1 + O\big(\frac{m_n \log p}{n}\big)\Big) \tag{A.215}$$

and

$$\|\mathbf{V}_{22}^{-1} - (\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})^{-1}\|_{\max} \leq C\frac{m_n \log p}{n}. \tag{A.216}$$

In addition, given event $\mathcal{C}$ and the range that $|x| \leq C\sqrt{\log p}$ and $|y| \leq C\sqrt{\log p}$, it holds that

$$\left\|\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right\|_2 \leq C\sqrt{\frac{m_n}{n}}\log p. \tag{A.217}$$

Further, given event $\mathcal{C}$ and that $\max\{|u|, |w|, |x|, |y|\} \leq C\sqrt{\log p}$, it follows from (A.215)–(A.217) that as $\frac{m_n(\log p)^3}{n} = o(1)$,

$$\begin{aligned}&\Bigg|\left[\begin{pmatrix} u \\ w \end{pmatrix} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right]^T(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12})^{-1}\left[\begin{pmatrix} u \\ w \end{pmatrix} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right] \\ &\quad - \begin{pmatrix} u \\ w \end{pmatrix}^T\mathbf{V}_{22}^{-1}\begin{pmatrix} u \\ w \end{pmatrix}\Bigg| \leq C\sqrt{\frac{m_n(\log p)^3}{n}}.\end{aligned} \tag{A.218}$$

Hence, substituting the bounds in (A.215) and (A.218) into (A.211) yields that as $\frac{m_n(\log p)^3}{n} = o(1)$,

$$\begin{aligned}&f_{\xi_l,\xi_{l+p}|(\xi_j,\xi_{j+p})}(u, w|x, y) \\ &= \frac{1}{2\pi|\mathbf{V}_{22}|^{1/2}}\exp\Big\{-\frac{1}{2}\begin{pmatrix} u \\ w \end{pmatrix}^T\mathbf{V}_{22}^{-1}\begin{pmatrix} u \\ w \end{pmatrix}\Big\}\cdot\Big(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\Big) \\ &= f_{\xi_l,\xi_{l+p}}(u, w)\Big(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\Big),\end{aligned} \tag{A.219}$$

which entails that (ξ_l, ξ_{l+p}) is asymptotically independent of (ξ_j, ξ_{j+p}) for $l \neq j$ and $l \in N_j^c$. By plugging (A.219) into (A.210), we can deduce that

$$\begin{aligned} P_1 \leq \mathbb{E}\Big\{ & \mathbb{1}(\mathcal{C})\mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \geq t - \delta_n, \max\{|\xi_j|, |\xi_{j+p}|\} \leq C\sqrt{\log p} \,|\, \widetilde{\mathbf{X}}^{\text{aug}}\big) \\ & \times \mathbb{P}\big(|\xi_l| - |\xi_{l+p}| \geq t - \delta_n, \max\{|\xi_l|, |\xi_{l+p}|\} \leq C\sqrt{\log p} \,|\, \widetilde{\mathbf{X}}^{\text{aug}}\big)\Big\} \\ & \times \left(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\right) + \mathbb{P}(\mathcal{C}^c), \end{aligned} \tag{A.220}$$

where $\mathbb{P}(\mathcal{C}^c) = O(p^{-3})$.

We next show that given $\widetilde{\mathbf{X}}^{\text{aug}}$, $|\xi_j| - |\xi_{j+p}|$ converges in distribution to η_j. Given $\widetilde{\mathbf{X}}^{\text{aug}}$, we see that

$$(\xi_j, \xi_{j+p}) \overset{d}{\sim} N(\mathbf{0}, \mathbf{V}_{11}).$$

Without ambiguity, let us denote by

$$\mathbf{V}_{11} = \begin{pmatrix} \sigma_{1,n}^2 & \rho_n \sigma_{1,n}\sigma_{2,n} \\ \rho_n \sigma_{1,n}\sigma_{2,n} & \sigma_{2,n}^2 \end{pmatrix}$$

for simpler notation, where $\sigma_{1,n}^2 = n\tau_j^2$, $\sigma_{2,n}^2 = n\tau_{j+p}^2$, and $\rho_n = \widetilde{\mathbf{z}}_j^T \widetilde{\mathbf{z}}_{j+p} / (\|\widetilde{\mathbf{z}}_j\|_2 \|\widetilde{\mathbf{z}}_{j+p}\|_2)$. We define an event

$$\begin{aligned} \mathcal{E} = \Big\{ & |\sigma_{1,n}^2 - (\mathbb{E}e_j^2)^{-1}| \leq C\sqrt{\frac{m_n \log p}{n}}, |\sigma_{2,n}^2 - (\mathbb{E}e_{j+p}^2)^{-1}| \leq C\sqrt{\frac{m_n \log p}{n}}, \\ & \text{and } \big|\rho_n - \text{corr}(e_j, e_{j+p})\big| \leq C\sqrt{\frac{m_n \log p}{n}}\Big\}. \end{aligned}$$

It follows from Lemma 10 that $\mathbb{P}(\mathcal{E}) \geq 1 - O(p^{-3})$. Some straightforward calculations show that for $t > 0$, given $\widetilde{\mathbf{X}}^{\text{aug}}$ the density of $|\xi_j| - |\xi_{j+p}|$ can be written as

$$\begin{aligned} f_{|\xi_j| - |\xi_{j+p}|}(t) = & \frac{\sqrt{2}}{\sqrt{\pi} a_{1,n}} \big[1 - \Phi(a_{2,n}^{-1} t)\big] \exp\big\{ - t^2/(2a_{1,n}^2)\big\} \\ & + \frac{\sqrt{2}}{\sqrt{\pi} a_{3,n}} \big[1 - \Phi(a_{4,n}^{-1} t)\big] \exp\big\{ - t^2/(2a_{3,n}^2)\big\}, \end{aligned} \tag{A.221}$$

where

$$\begin{aligned} a_{1,n} &= \sqrt{\sigma_{1,n}^2 + \sigma_{2,n}^2 - 2\rho_n \sigma_{1,n}\sigma_{2,n}}, \quad a_{2,n} = \frac{\sigma_{1,n}\sigma_{2,n} a_{1,n} \sqrt{(1 - \rho_n^2)}}{\sigma_{2,n}^2 - \rho_n \sigma_{1,n}\sigma_{2,n}}, \\ a_{3,n} &= \sqrt{\sigma_{1,n}^2 + \sigma_{2,n}^2 + 2\rho_n \sigma_{1,n}\sigma_{2,n}}, \quad a_{4,n} = \frac{\sigma_{1,n}\sigma_{2,n} a_{3,n} \sqrt{(1 - \rho_n^2)}}{\sigma_{2,n}^2 + \rho_n \sigma_{1,n}\sigma_{2,n}}. \end{aligned}$$

Recall the notation

$$v_j = \sqrt{2(\mathbb{E}e_j^2)^{-1}(1 - \text{corr}(e_j, e_{j+p}))}$$

and

$$b_j = \sqrt{2(\mathbb{E}e_j^2)^{-1}(1 + \text{corr}(e_j, e_{j+p}))}.$$

It holds that $\mathbb{E}(e_j^2) = (\mathbf{\Omega}^A_{j,j})^{-1} = (\mathbf{\Omega}^A_{j+p,j+p})^{-1} = \mathbb{E}(e_{j+p}^2)$ due to the symmetry of $\mathbf{\Omega}^A$. On event $\mathcal{E}$, we have that

$$|a_{1,n}/v_j - 1| \le C\sqrt{\frac{m_n \log p}{n}}, \quad |a_{2,n}/b_j - 1| \le C\sqrt{\frac{m_n \log p}{n}},$$
$$|a_{3,n}/b_j - 1| \le C\sqrt{\frac{m_n \log p}{n}}, \quad \big|a_{4,n}/v_j - 1\big| \le C\sqrt{\frac{m_n \log p}{n}}.$$

Thus, in view of the definition of $h_j(t)$ in (A.191) and (A.221), it follows that as $|t| \le C\sqrt{\log p}$,

$$f_{|\xi_j|-|\xi_{j+p}|}(t) = h_j(t)\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg).$$

With the aid of the above result, we can deduce that on event $\mathcal{E}$,

$$\begin{aligned}
&\mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \ge t - \delta_n, \max\{|\xi_j|, |\xi_{j+p}|\} \le C\sqrt{\log p} \,|\, \widetilde{\mathbf{X}}^{\mathrm{aug}}\big) \\
&\le \mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \ge t - \delta_n, |\xi_j| - |\xi_{j+p}| \le C\sqrt{\log p} \,|\, \widetilde{\mathbf{X}}^{\mathrm{aug}}\big) \\
&\le \bigg(\int_{t-\delta_n}^{C\sqrt{\log p}} h_j(u)\,du\bigg)\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg) \\
&= \mathbb{P}(t - \delta_n \le \eta_j \le C\sqrt{\log p})\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg) \\
&= \big[\mathbb{P}(\eta_j \ge t - \delta_n) - \mathbb{P}(\eta_j > C\sqrt{\log p})\big]\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg).
\end{aligned} \tag{A.222}$$

Moreover, in light of (A.201) it is easy to see that

$$\mathbb{P}(\eta_j > C\sqrt{\log p}) = O(p^{-3})$$

for some large constant C, which together with (A.223) leads to

$$\begin{aligned}
&\mathbb{P}\big(|\xi_j| - |\xi_{j+p}| \ge t - \delta_n, \max\{|\xi_j|, |\xi_{j+p}|\} \le C\sqrt{\log p} \,|\, \widetilde{\mathbf{X}}^{\mathrm{aug}}\big) \\
&\le \mathbb{P}(\eta_j \ge t - \delta_n)\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg) + O(p^{-3}).
\end{aligned} \tag{A.223}$$

Plugging (A.223) into (A.220) shows that

$$P_1 \le \mathbb{P}(\eta_j \ge t - \delta_n)\mathbb{P}(\eta_l \ge t - \delta_n)\bigg(1 + O\Big(\sqrt{\frac{m_n(\log p)^3}{n}}\Big)\bigg) + O(p^{-3}). \tag{A.224}$$

Finally, combining (A.207), (A.208), (A.209), and (A.224) yields (A.192). Similarly, we can also establish (A.193)–(A.195). This completes the proof of Lemma 13.

B.12. Proof of Lemma 14. Let us first prove (A.274). In the proof of Lemma 13 in Section B.11, we have established the lower bound and upper bound for $\mathbb{P}(\widetilde{W}_j \ge t)$ in (A.194) and (A.195), respectively. Recall the definitions that $\delta_n = C\frac{m_n^{1/2} s \log p}{\sqrt{n}}$ and $b_n =$

$C\Delta_n s\sqrt{\frac{\log p}{n}}$. For the numerator and denominator in (A.274), we can write that

$$p_0(G(t-b_n)-G(t+b_n)) = \sum_{j\in\mathcal{H}_0}\big[\mathbb{P}(\widetilde{W}_j \geq t-b_n)-\mathbb{P}(\widetilde{W}_j \geq t+b_n)\big]$$

$$\leq \sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j \geq \sqrt{n}t-\sqrt{n}b_n-\delta_n)\big(1+O\big(\sqrt{\frac{m_n(\log p)^3}{n}}\big)\big)$$

$$-\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j \geq \sqrt{n}t+\sqrt{n}b_n+\delta_n)\big(1+O\big(\sqrt{\frac{m_n(\log p)^3}{n}}\big)\big)+O(p^{-2}) \tag{A.225}$$

$$\leq \sum_{j\in\mathcal{H}_0}\mathbb{P}(\sqrt{n}t-\sqrt{n}b_n-\delta_n \leq \eta_j \leq \sqrt{n}t+\sqrt{n}b_n+\delta_n)$$

$$+\sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j \geq \sqrt{n}t-\sqrt{n}b_n-\delta_n)\cdot O\big(\sqrt{\frac{m_n(\log p)^3}{n}}\big)+O(p^{-2})$$

and

$$p_0G(t) \geq \sum_{j\in\mathcal{H}_0}\mathbb{P}(\eta_j \geq \sqrt{n}t+\delta_n)\big(1+O\big(\sqrt{\frac{m_n(\log p)^3}{n}}\big)\big)+O(p^{-2}), \tag{A.226}$$

respectively.

It follows from (A.225)–(A.226), similar arguments as for (A.204), and $G^{-1}(\frac{c_1qa_n}{p}) = O(\sqrt{\frac{\log p}{n}})$ in the proof of Lemma 13 that as $\sqrt{n}G^{-1}(\frac{c_1qa_n}{p})(\sqrt{n}b_n+\delta_n)\to 0$,

$$\sup_{t\in(0,G^{-1}(\frac{c_1qa_n}{p})]}\frac{G(t-b_n)-G(t+b_n)}{G(t)} \leq C\sqrt{\log p}(\sqrt{n}b_n+\delta_n)+C\sqrt{\frac{m_n(\log p)^3}{n}}$$

$$\leq C\Big(\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}}+\Delta_n s\log p\Big).$$

Thus, we see that when $\frac{m_n^{1/2}s(\log p)^{3/2+1/\gamma}}{\sqrt{n}}+\Delta_n s(\log p)^{1+1/\gamma}\to 0$, the desired result (A.274) holds.

We next proceed with establishing (A.275). In view of Condition 10, it holds that

$$p_1^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}(\widetilde{W}_j < -t) \leq G(t)$$

for $t = O(\sqrt{n^{-1}\log p})$. Moreover, we have

$$b_n = C\Delta_n s\sqrt{\frac{\log p}{n}} = o(G^{-1}(\frac{c_1qa_n}{p}))$$

due to the assumption $\Delta_n s\to 0$ and $G^{-1}(\frac{c_1qa_n}{p}) = O(\sqrt{\frac{\log p}{n}})$. Then it follows that

$$a_n^{-1}\sum_{j\in\mathcal{H}_1}\mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1qa_n}{p})+b_n\Big)$$

$$\leq a_n^{-1}(p-p_0)G\Big(G^{-1}(\frac{c_1qa_n}{p})-b_n\Big) \tag{A.227}$$

$$=\frac{c_1q(p-p_0)}{p}+a_n^{-1}(p-p_0)\Big[G\Big(G^{-1}(\frac{c_1qa_n}{p})-b_n\Big)-G\Big(G^{-1}(\frac{c_1qa_n}{p})\Big)\Big].$$

For notational simplicity, let us define

$$t_n = G^{-1}(\frac{c_1 q a_n}{p}).$$

With the aid of the upper and lower bounds for $\mathbb{P}(\widetilde{W}_j \geq t)$ given in (A.194) and (A.195), we can deduce that

$$\begin{aligned}
&G(t_n - b_n) - G(t_n) \\
&\leq p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\eta_j \geq \sqrt{n} t_n - \sqrt{n} b_n - \delta_n) \Big(1 + O\Big(\sqrt{\frac{m_n (\log p)^3}{n}}\Big)\Big) \\
&\quad - p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\eta_j \geq \sqrt{n} t_n + \delta_n) \Big(1 + O\Big(\sqrt{\frac{m_n (\log p)^3}{n}}\Big)\Big) + O(p^{-2}) \\
&= p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\sqrt{n} t_n - \sqrt{n} b_n - \delta_n \leq \eta_j \leq \sqrt{n} t_n + \delta_n) \\
&\quad + p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\eta_j \geq \sqrt{n} t_n - \sqrt{n} b_n - \delta_n) \cdot O\Big(\sqrt{\frac{m_n (\log p)^3}{n}}\Big) + O(p^{-2}).
\end{aligned} \tag{A.228}$$

An application of similar arguments as for (A.135) leads to

$$\begin{aligned}
&\frac{\mathbb{P}(\sqrt{n} t_n - \sqrt{n} b_n - \delta_n \leq \eta_j \leq \sqrt{n} t_n + \delta_n)}{\mathbb{P}(\eta_j \geq \sqrt{n} t_n + \delta_n)} \\
&\leq C \sqrt{n} t_n (\sqrt{n} b_n + \delta_n) \\
&\leq C \Big(\frac{m_n^{1/2} s (\log p)^{3/2}}{\sqrt{n}} + \Delta_n s \log p \Big)
\end{aligned} \tag{A.229}$$

and

$$\Big| \frac{\mathbb{P}(\eta_j \geq \sqrt{n} t_n - \sqrt{n} b_n - \delta_n)}{\mathbb{P}(\eta_j \geq \sqrt{n} t_n + \delta_n)} - 1 \Big| \leq C \Big(\frac{m_n^{1/2} s (\log p)^{3/2}}{\sqrt{n}} + \Delta_n s \log p \Big). \tag{A.230}$$

It follows from the lower bound in (A.194) and $G(t_n) = G(G^{-1}(\frac{c_1 q a_n}{p})) = \frac{c_1 q a_n}{p}$ that as $\frac{m_n (\log p)^3}{n} \to 0$,

$$p_0^{-1} \sum_{j \in \mathcal{H}_0} \mathbb{P}(\eta_j \geq \sqrt{n} t_n + \delta_n) \leq C(\frac{c_1 q a_n}{p} + O(p^{-3})) \leq C \frac{c_1 q a_n}{p}. \tag{A.231}$$

Therefore, combining (A.228)–(A.231) shows that

$$\begin{aligned}
G(t_n - b_n) - G(t_n) &\leq C \Big(\frac{m_n^{1/2} s (\log p)^{3/2}}{\sqrt{n}} + \Delta_n s \log p \Big) \cdot \frac{c_1 q a_n}{p} + C \sqrt{\frac{m_n (\log p)^3}{n}} \cdot \frac{c_1 q a_n}{p} \\
&\quad + O(p^{-2}) \\
&\leq C \Big(\frac{m_n^{1/2} s (\log p)^{3/2}}{\sqrt{n}} + \Delta_n s \log p \Big) \cdot \frac{c_1 q a_n}{p} + O(p^{-2}).
\end{aligned}$$

Finally, substituting the above bound into (A.227) yields that as $\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}} + \Delta_n s(\log p) \to 0$,

$$\begin{aligned} & a_n^{-1} \sum_{j \in \mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p} + \Delta_n\Big) \\ & \leq \frac{c_1 q(p-p_0)}{p} + C\Big(\frac{m_n^{1/2}s(\log p)^{3/2}}{\sqrt{n}} + \Delta_n s \log p\Big) \cdot \frac{c_1 q(p-p_0)}{p} \\ & \quad + O(\frac{p-p_0}{a_n p^2}) \\ & \to 0, \end{aligned}$$

where we have used the assumption that $p_0/p \to 1$. This establishes (A.275), which concludes the proof of Lemma 14.

B.13. Proof of Lemma 15. The proof of this lemma relies on the definitions of T_v and $\widetilde{T}_v$, with the intuition that $\widetilde{T}_v$ resembles the vth order statistic of $-\widetilde{W}_j$, while T_v resembles the vth order statistic of $-\widehat{W}_j$. Intuitively, this means that if the distance between $\widetilde{W}_j$ and $\widehat{W}_j$ is bounded by b_n, the distance between the corresponding order statistics should also be bounded by b_n. We will formalize such argument next.

Let us define an event

$$\mathscr{C} := \{\max_{1 \leq j \leq p} |\widehat{W}_j - \widetilde{W}_j| \leq b_n\}.$$

Condition 1 assumes that $\mathbb{P}(\mathscr{C}) \to 1$. Denote by

$$\widehat{S}_v = \big\{1 \leq j \leq p : -\widehat{W}_j \geq T_v\big\} \tag{A.232}$$

and

$$\widetilde{S}_v = \big\{1 \leq j \leq p : -\widetilde{W}_j \geq \widetilde{T}_v\big\}. \tag{A.233}$$

Observe that $|\widehat{S}_v| = v$ and $|\widetilde{S}_v| = v$ by the definitions of T_v and $\widetilde{T}_v$. If $j_0 \in \widehat{S}_v$, on event $\mathscr{C}$ we have that

$$-\widetilde{W}_{j_0} = -\widehat{W}_{j_0} + (\widehat{W}_{j_0} - \widetilde{W}_{j_0}) \geq T_v - b_n, \tag{A.234}$$

which entails that $\sum_{j=1}^p \mathbb{1}(-\widetilde{W}_j \geq T_n - b_n) \geq v$. Moreover, since $\widetilde{T}_v$ satisfies $\sum_{j=1}^p \mathbb{1}(-\widetilde{W}_j \geq \widetilde{T}_v) = v$, it follows that

$$\widetilde{T}_v \geq T_v - b_n$$

by the monotonicity of the indicator function. Similarly, we can also show that

$$T_v \geq \widetilde{T}_v - b_n$$

on event $\mathscr{C}$. Thus, (A.26) is derived. This concludes the proof of Lemma 15.

B.14. Proof of Lemma 16. Note that k is the number of failures before v successes in a binomial process with success probability $\frac{1}{2}$. The major intuition of the desired result (A.27) is that by the law of large numbers, the number of failures and successes should become asymptotically comparable as the number of trials tends to infinity. Let D_{k+v-1} be a binomial random variable with distribution $B(k+v-1, \frac{1}{2})$ and L_v the negative binomial random variable with distribution $NB(v, \frac{1}{2})$. Observe that (52) is equivalent to $\mathbb{P}(L_v \geq k) \leq$

q. According to the relationship between the negative binomial distribution and binomial distribution, we have that

$$\begin{aligned}\mathbb{P}(L_v \geq k) &= 1 - \mathbb{P}(L_v \leq k-1) \\ &= 1 - \mathbb{P}(D_{k+v-1} \geq v) \\ &= \mathbb{P}(D_{k+v-1} \leq v-1).\end{aligned} \tag{A.235}$$

By the central limit theorem, it holds that when $k+v \to \infty$,

$$\mathbb{P}(D_{k+v-1} \leq v-1) = \Phi\Big(\frac{v-1-k}{\sqrt{k+v-1}}\Big) + o(1).$$

Therefore, (52) implies that

$$\frac{v-1-k}{\sqrt{k+v-1}} \leq \Phi^{-1}(q - o(1)). \tag{A.236}$$

In addition, since v is the largest integer such that (52) holds, we have that

$$\mathbb{P}(L_{v+1} \geq k) > q.$$

Using similar arguments as for (A.236), it follows that as $k+v \to \infty$,

$$\mathbb{P}(L_{v+1} \geq k) = \mathbb{P}(D_{k+v} \leq v) = \Phi\Big(\frac{v-k}{\sqrt{k+v}}\Big) + o(1)$$

and hence

$$\frac{v-k}{\sqrt{k+v}} \geq \Phi^{-1}(q - o(1)), \tag{A.237}$$

which along with (A.236) leads to (A.27). This completes the proof of Lemma 16.

B.15. Proof of Lemma 17. The proof of this lemma consists of two steps. We will first establish the tight bounds below for $\widetilde{T}_v$. In the second step, noting that $\widetilde{T}_{v+M_v+1} < \widetilde{T}_v - 2b_n \leq \widetilde{T}_{v+M_v}$ by the definition of M_v in (A.25), we will show that M_v is bounded as long as b_n is sufficiently small.

LEMMA 18. *For $0 < \varepsilon < 1/8$, under Conditions 1, 15, and 16 we have that*

$$\mathbb{P}\Big(G^{-1}\big(\frac{v(1+\varepsilon)}{p_0}\big) < \widetilde{T}_v < G^{-1}\big(\frac{v(1-\varepsilon)}{p_0}\big)\Big) \to 1. \tag{A.238}$$

LEMMA 19. *Under Condition 16, we have that*

$$2b_n < G^{-1}\big(\frac{v(1+\varepsilon)}{p_0}\big) - G^{-1}\big(\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0}\big).$$

Using similar arguments as in the proof of Lemma 18 below, we can show that under Conditions 1, 15, and 16,

$$\mathbb{P}\Big(G^{-1}\big(\frac{v(1+3\varepsilon)(1+\varepsilon)}{p_0}\big) < \widetilde{T}_{v(1+3\varepsilon)} < G^{-1}\big(\frac{(v(1+3\varepsilon))(1-\varepsilon)}{p_0}\big)\Big) \to 1. \tag{A.239}$$

Then it follows that

$$\widetilde{T}_{v(1+3\varepsilon)} < G^{-1}\big(\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0}\big) < G^{-1}\big(\frac{v(1+\varepsilon)}{p_0}\big) < \widetilde{T}_v.$$

Additionally, applying Lemmas 18 and 19 together with the definition of $\widetilde{T}_v$ gives that with asymptotic probability one,

$$\begin{aligned}\widetilde{T}_{v+M_v} &\geq \widetilde{T}_v - 2b_n \\ &\geq G^{-1}(\frac{v(1+\varepsilon)}{p_0}) - \Big[G^{-1}(\frac{v(1+\varepsilon)}{p_0}) - G^{-1}(\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0})\Big] \\ &= G^{-1}(\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0}) > \widetilde{T}_{v(1+3\varepsilon)}.\end{aligned}$$

Therefore, we can obtain that

$$\mathbb{P}(M_v < 3v\varepsilon) \to 1$$

since $\widetilde{T}_v$ is decreasing with respect to v. This will conclude the proof of Lemma 17.

We will present the formal proofs of Lemmas 18 and 19 below.

Proof of Lemma 18. The main idea of the proof is to establish the convergence of the empirical distribution of $\{\widetilde{W}_j\}$ that $\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \geq t)$ is close to $\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \geq t)$. Using similar arguments as in the proof of Lemma 3 in Section B.3, we can obtain that when $m_n/k \to 0$ (which combined with Lemma 16 implies that $m_n/v \to 0$),

$$\sup_{t\in(G^{-1}(\frac{3k}{2p}),G^{-1}(\frac{k}{2p}))}\left|\frac{\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \leq -t)}{\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \leq -t)} - 1\right| = o_p(1). \tag{A.240}$$

Since $\sum_{j\in\mathcal{H}_0}\mathbb{P}(\widetilde{W}_j \leq -G^{-1}(\frac{v(1+\varepsilon)}{p_0}) = v(1+\varepsilon)$, we see from (A.240) that

$$\begin{aligned}\sum_{j=1}^{p}\mathbb{1}(-\widetilde{W}_j \geq G^{-1}(\frac{v(1+\varepsilon)}{p_0})) &\geq \sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \leq -G^{-1}(\frac{v(1+\varepsilon)}{p_0})) \\ &= v(1+\varepsilon)(1+o_p(1)) > v\end{aligned} \tag{A.241}$$

holds with asymptotic probability one. Hence, from the definition of $\widetilde{T}_v$, we have that

$$\mathbb{P}\Big(\widetilde{T}_v > G^{-1}(\frac{v(1+\varepsilon)}{p_0})\Big) \to 1. \tag{A.242}$$

We next prove the upper bound for $\widetilde{T}_v$. Note that $\sum_{j=1}^{p}\mathbb{1}(\widetilde{W}_j \leq -\widetilde{T}_v) = v$. We will aim to show that with asymptotic probability one,

$$\sum_{j\in\mathcal{H}_1}\mathbb{1}(\widetilde{W}_j \leq -\widetilde{T}_v) < v\varepsilon/2. \tag{A.243}$$

Then with asymptotic probability one, it holds that

$$\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \leq -\widetilde{T}_v) \geq v(1-\varepsilon/2). \tag{A.244}$$

On the other hand, applying (A.240) and similar argument as for (A.241), we can obtain that with asymptotic probability one,

$$\sum_{j\in\mathcal{H}_0}\mathbb{1}(\widetilde{W}_j \leq -G^{-1}(\frac{v(1-\epsilon_n)}{p_0}) < v(1-\varepsilon/2). \tag{A.245}$$

Combining the above two results shows that with asymptotic probability one,

$$\widetilde{T}_v \le G^{-1}(\frac{v(1-\varepsilon)}{p_0}),$$

which completes the proof for the upper bound.

It remains to establish (A.243). Since $p_0/p \to 1$ and $v/k \to 1$ (cf. Lemma 16), we have that

$$G^{-1}(\frac{3k}{2p}) < G^{-1}(\frac{v(1+\varepsilon)}{p_0})$$

when n and p are sufficiently large and $0 < \varepsilon < 1/8$. Then from (A.242), it holds that $G^{-1}(\frac{3k}{2p}) \le \widetilde{T}_v$ and hence with asymptotic probability one,

$$\sum_{j\in\mathcal{H}_1} \mathbb{1}(\widetilde{W}_j \le -\widetilde{T}_v) \le \sum_{j\in\mathcal{H}_1} \mathbb{1}(\widetilde{W}_j < -G^{-1}(\frac{3k}{2p})). \tag{A.246}$$

Moreover, an application of the Markov inequality, Lemma 16, and (55) in Condition 16 yields that as $n \to \infty$,

$$\begin{aligned} &\mathbb{P}\Big(\sum_{j\in\mathcal{H}_1} \mathbb{1}(\widetilde{W}_j < -G^{-1}(\frac{3k}{2p})) > v\varepsilon/2\Big) \\ &\le \frac{2}{v\varepsilon}\sum_{j\in\mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{3k}{2p})\Big) \to 0. \end{aligned} \tag{A.247}$$

Therefore, (A.243) is derived in view of (A.246). This completes the proof of Lemma 18.

Proof of Lemma 19. Let us observe that

$$\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0} - \frac{v(1+\varepsilon)}{p_0} = \frac{v}{p_0}(\varepsilon - 3\varepsilon^2). \tag{A.248}$$

By the assumptions that $p_0/p \to 1$ and $m_n/k \to 0$, and applying Lemma 16 and the observation above, it follows that when k and p are sufficiently large,

$$\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0} - \frac{v(1+\varepsilon)}{p_0} \ge \frac{k\varepsilon}{2p}. \tag{A.249}$$

Note that assumption (54) in Condition 16 entails that

$$\sup_{t\in(G^{-1}(\frac{3k}{2p}),G^{-1}(\frac{k}{2p}))} [G(t-b_n) - G(t+b_n)] = o(\frac{k}{p}). \tag{A.250}$$

Combining the above two results and Lemma 16, we can obtain that

$$\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0} - \frac{v(1+\varepsilon)}{p_0} \gg \sup_{t\in(G^{-1}(\frac{3k}{2p}),G^{-1}(\frac{k}{2p}))} [G(t-b_n) - G(t+b_n)]. \tag{A.251}$$

Notice that

$$G^{-1}(\frac{v(1+3\varepsilon)(1-\varepsilon)}{p_0}) \in (G^{-1}(\frac{3k}{2p}), G^{-1}(\frac{k}{2p}))$$

and

$$G^{-1}(\frac{v(1+\varepsilon)}{p_0}) \in (G^{-1}(\frac{3k}{2p}), G^{-1}(\frac{k}{2p}))$$

when k and p are sufficiently large. Therefore, using proof by contradiction and the monotonicity of function $G(\cdot)$, we can establish the desired result of Lemma 19. This concludes the proof of Lemma 19.

B.16. Lemma 20 and its proof. Following the definitions in Section 3.2, let us consider the marginal correlation approximate knockoff statistics defined as $\widehat{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widehat{\mathbf{X}}_j^T\mathbf{y}|)$ and the coupled perfect knockoff statistics given by $\widetilde{W}_j = (\sqrt{n}\|\mathbf{y}\|_2)^{-1}(|\mathbf{X}_j^T\mathbf{y}| - |\widetilde{\mathbf{X}}_j^T\mathbf{y}|)$ with $1 \le j \le p$. When features $X_1, \cdots, X_p$ are independent, we can obtain the following sharper bound of order $\Delta_n\sqrt{\frac{\log p}{n}}$ for the coupling accuracy of the knockoff statistics, compared to the general bound Δ_n in (13).

LEMMA 20. *Assume that features $\{X_j\}_{j=1}^p$ are independent and follow a Gaussian distribution $X_j \overset{d}{\sim} N(0, \sigma_j^2)$. Let the approximate and coupled knockoff variable matrices be defined as*

$$\widehat{\mathbf{X}} = \mathbf{Z}\,\mathrm{diag}(\widehat{\sigma}_1, \cdots, \widehat{\sigma}_p) \ \text{ and } \ \widetilde{\mathbf{X}} = \mathbf{Z}\,\mathrm{diag}(\sigma_1, \cdots, \sigma_p),$$

where $\mathbf{Z} = (\mathbf{Z}_{ij}) \in \mathbb{R}^{n\times p}$ has i.i.d. standard normal entries and is independent of $(\mathbf{X}, \mathbf{y})$, and $\widehat{\sigma}_j$ is the estimator of σ_j, which can be learned in sample. Then under Condition 6, we have that when $\log p = o(n)$,

$$\max_{1\le j\le p} |\widehat{W}_j - \widetilde{W}_j| \le 4\Delta_n\sqrt{\frac{\log p}{n}}. \tag{A.252}$$

Proof of Lemma 20. We first show that Condition 6 leads to $\mathbb{P}(\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| \le 2\Delta_n) \to 1$. First note that $\mathbb{P}(\min_{1\le j\le p} n^{-1/2}\|\mathbf{Z}_j\|_2 > 1/2) \to 1$ since $\log p = o(n)$, due to the concentration inequality for the sum of i.i.d χ_1^2 random variables. If $\max_{1\le j\le p} n^{-1/2}\|\widehat{\mathbf{X}}_j - \widetilde{\mathbf{X}}_j\|_2 \le \Delta_n$ with probability approaching one, then we have $\Delta_n \ge \max_{1\le j\le p} n^{-1/2}|\widehat{\sigma}_j - \sigma_j|\|\mathbf{Z}_j\|_2 \ge \max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| \min_{1\le j\le p} n^{-1/2}\|\mathbf{Z}_j\|_2 > \frac{1}{2}\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j|$ with asymptotic probability one. This proves that Condition 6 leads to

$$\mathbb{P}(\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| < 2\Delta_n) \to 1.$$

Then we can deduce that

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j| \ge 4\Delta_n\sqrt{\frac{\log p}{n}}\Big)\\
&\le \mathbb{P}\Big(\max_{1\le j\le p}(\sqrt{n}\|\mathbf{y}\|_2)^{-1}\big|(\widehat{\mathbf{X}}_j^T - \widetilde{\mathbf{X}}_j)^T\mathbf{y}\big| \ge 4\Delta_n\sqrt{\frac{\log p}{n}}\Big)\\
&\le \mathbb{P}\Big(\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| \max_{1\le j\le p}(\sqrt{n}\|\mathbf{y}\|_2)^{-1}\big|\mathbf{Z}_j^T\mathbf{y}\big| \ge 4\Delta_n\sqrt{\frac{\log p}{n}}\Big)\\
&\le \mathbb{P}\Big(\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| \ge 2\Delta_n\Big) + \sum_{1\le j\le p}\mathbb{P}\Big((\sqrt{n}\|\mathbf{y}\|_2)^{-1}\big|\mathbf{Z}_j^T\mathbf{y}\big| \ge 2\sqrt{\frac{\log p}{n}}\Big).
\end{aligned}$$

Observing that $(\sqrt{n}\|\mathbf{y}\|_2)^{-1}\mathbf{Z}_j^T\mathbf{y} \overset{d}{\sim} N(0, n^{-1})$ and $\max_{1\le j\le p}|\widehat{\sigma}_j - \sigma_j| \le 2\Delta_n$ with asymptotic probability one, we have

$$\mathbb{P}(\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j| \ge 4\Delta_n\sqrt{\frac{\log p}{n}}) \le o(1) + p^{-1} \to 0.$$

This completes the proof of Lemma 20.

APPENDIX C: RCD WITH DEBIASED LASSO IN GLM

In this section, we extend the results in Section 3.3 to the setting of the generalized linear model (GLM)

$$\mathbb{E}[Y|X] = g^{-1}(X^T\boldsymbol{\alpha}^0),$$

where $\boldsymbol{\alpha}^0 = (\alpha_j^0)_{1\le j\le p} \in \mathbb{R}^p$ is the true regression coefficient vector and g is the link function. Assume that feature vector $X = (X_1, \cdots, X_p)^T$ has zero mean. Define $\widetilde{X}^{\text{aug}} = (X^T, \widetilde{X}^T)^T \in \mathbb{R}^{2p}$ and $\widehat{X}^{\text{aug}} = (X^T, \widehat{X}^T)^T \in \mathbb{R}^{2p}$, where $\widetilde{X}$ and $\widehat{X}$ are the perfect knockoffs and the approximate knockoffs for X, respectively. Denote by $\boldsymbol{\beta}^0 = ((\boldsymbol{\alpha}^0)^T, \mathbf{0}_p^T)^T \in \mathbb{R}^{2p}$ the augmented true parameter vector.

Consider the negative log-likelihood function $\rho(y; a) : a \mapsto \mathbb{R}$ defined as $\rho(y; a) = -ya + b(a)$, up to a constant independent of the unknown parameters, where $b(\cdot)$ is a known strictly convex and twice continuously differentiable function. Define the loss function $\rho_{\boldsymbol{\beta}}(Y; \widetilde{X}^{\text{aug}}) = \rho(Y; (\widetilde{X}^{\text{aug}})^T\boldsymbol{\beta})$. Denote by $\dot{\rho}_{\boldsymbol{\beta}} := \frac{\partial}{\partial\boldsymbol{\beta}}\rho_{\boldsymbol{\beta}}$ and $\ddot{\rho}_{\boldsymbol{\beta}} := \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\rho_{\boldsymbol{\beta}}$ the partial derivatives. Note that $\dot{\rho}_{\boldsymbol{\beta}} = \dot{\rho}(Y; (\widetilde{X}^{\text{aug}})^T\boldsymbol{\beta})\widetilde{X}^{\text{aug}}$ and $\dot{\rho}_{\boldsymbol{\beta}} = \ddot{\rho}(Y; (\widetilde{X}^{\text{aug}})^T\boldsymbol{\beta})\widetilde{X}^{\text{aug}}(\widetilde{X}^{\text{aug}})^T$.

Let $\widehat{\mathbf{b}} = (\widehat{b}_j)_{1\le j\le 2p}$ be the debiased estimator for the GLM given in van de Geer et al. (2014) based on the augmented design matrix $\widehat{\mathbf{X}}^{\text{aug}} := [\mathbf{X}, \widehat{\mathbf{X}}] \in \mathbb{R}^{n\times 2p}$, where $\widehat{\mathbf{X}}$ is the approximate knockoff variable matrix. Assume that Condition 6 is satisfied and $\widetilde{\mathbf{X}}$ is the coupled perfect knockoffs variable matrix. Similarly, define $\widetilde{\mathbf{X}}^{\text{aug}} := [\mathbf{X}, \widetilde{\mathbf{X}}] \in \mathbb{R}^{n\times 2p}$. Then $\widehat{\mathbf{b}}$ can be coupled with the debiased Lasso estimator denoted as $\widetilde{\mathbf{b}} = (\widetilde{b}_j)_{1\le j\le 2p} \in \mathbb{R}^{2p}$ based on $\widetilde{\mathbf{X}}^{\text{aug}}$. The regression coefficient difference knockoff statistics can be defined as

$$\widehat{W}_j = |\widehat{b}_j| - |\widehat{b}_{j+p}| \text{ and } \widetilde{W}_j = |\widetilde{b}_j| - |\widetilde{b}_{j+p}|, \quad 1 \le j \le p \tag{A.253}$$

for the approximate and the coupled perfect knockoffs procedures, respectively..

We provide the explicit definition of the debiased Lasso estimator to assist future presentation. For each $1 \le j \le 2p$, the debiased Lasso estimator $\widehat{\mathbf{b}} = (\widehat{b}_j)_{1\le j\le 2p}$ is a one-step bias correction from the Lasso estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_j)_{1\le j\le 2p} \in \mathbb{R}^{2p}$. First, the Lasso estimator is given by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{2p}}{\arg\min} \Big\{ n^{-1}\sum_{i=1}^{n} \rho_{\boldsymbol{\beta}}(y_i; \widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}) + \lambda\|\boldsymbol{\beta}\|_1 \Big\}, \tag{A.254}$$

where $\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}$ is the ith row (observation) of the augmented design matrix $\widehat{\mathbf{X}}^{\text{aug}}$. To obtain the debiased Lasso estimator, define

$$\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n} \ddot{\rho}_{\widehat{\boldsymbol{\beta}}}(y_i; \widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}) = n^{-1}(\widehat{\mathbf{X}}^{\text{aug}})^T\widehat{\mathbf{D}}\widehat{\mathbf{X}}^{\text{aug}},$$

where $\widehat{\mathbf{D}} = \text{diag}(\ddot{\rho}(y_1; \widehat{\mathbf{X}}_{1,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}}), \cdots, \ddot{\rho}(y_n; \widehat{\mathbf{X}}_{n,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}})) \in \mathbb{R}^{n\times n}$ is a diagonal matrix. Further, for $1 \le j \le 2p$, define

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{2p-1}}{\arg\min}(\widehat{\boldsymbol{\Sigma}}_{j,j} - 2\widehat{\boldsymbol{\Sigma}}_{j,-j}\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\widehat{\boldsymbol{\Sigma}}_{-j,-j}\boldsymbol{\gamma} + 2\lambda_j\|\boldsymbol{\gamma}\|_1),$$

where λ and $\{\lambda_j\}_{j=1}^{2p}$ are the nonnegative regularization parameters. In addition, let

$$\widehat{\tau}_j^2 = \widehat{\boldsymbol{\Sigma}}_{j,j} - \widehat{\boldsymbol{\Sigma}}_{j,-j}\widehat{\boldsymbol{\gamma}}_j.$$

Then the debiased Lasso estimator for GLM (van de Geer et al. (2014)) based on the approximate augmented design matrix $\widehat{\mathbf{X}}^{\text{aug}}$ is defined as

$$\widehat{b}_j = \widehat{\beta}_j - \frac{n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j)}{\widehat{\tau}_j^2}, \quad 1 \le j \le p, \tag{A.255}$$

where $\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}} := (\dot{\rho}(y_1; \widehat{\mathbf{X}}_{1,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}}), \cdots, \dot{\rho}(y_n; \widehat{\mathbf{X}}_{n,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}})) \in \mathbb{R}^n$.

Analogously, the coupled debiased Lasso estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_j)_{1\le j\le 2p}$ based on the perfect augmented design matrix $\widetilde{\mathbf{X}}^{\text{aug}}$ can be defined componentwisely as

$$\widetilde{b}_j = \widetilde{\beta}_j - \frac{n^{-1}\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}}^T(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}, \tag{A.256}$$

where $\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}} := (\dot{\rho}(y_1; \widetilde{\mathbf{X}}_{1,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}}), \cdots, \dot{\rho}(y_n; \widetilde{\mathbf{X}}_{n,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}})) \in \mathbb{R}^n$,

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{2p}}{\arg\min}\Big\{n^{-1}\sum_{i=1}^n \rho_{\boldsymbol{\beta}}(y_i; \widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}) + \lambda\|\boldsymbol{\beta}\|_1\Big\}, \tag{A.257}$$

$$\widetilde{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{2p-1}}{\arg\min}(\widetilde{\boldsymbol{\Sigma}}_{j,j} - 2\widetilde{\boldsymbol{\Sigma}}_{j,-j}\boldsymbol{\gamma} + \boldsymbol{\gamma}^T\widetilde{\boldsymbol{\Sigma}}_{-j,-j}\boldsymbol{\gamma} + 2\lambda_j\|\boldsymbol{\gamma}\|_1), \quad \widetilde{\tau}_j^2 = \widetilde{\boldsymbol{\Sigma}}_{j,j} - \widetilde{\boldsymbol{\Sigma}}_{j,-i}\widetilde{\boldsymbol{\gamma}}_j,$$

and

$$\widetilde{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \ddot{\rho}_{\widetilde{\boldsymbol{\beta}}}(y_i; \widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}) = n^{-1}(\widetilde{\mathbf{X}}^{\text{aug}})^T\widetilde{\mathbf{D}}\widetilde{\mathbf{X}}^{\text{aug}}.$$

In the above, $\widetilde{\mathbf{D}} = \text{diag}(\ddot{\rho}(y_1; \widetilde{\mathbf{X}}_{1,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}}), \cdots, \ddot{\rho}(y_n; \widetilde{\mathbf{X}}_{n,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}})) \in \mathbb{R}^{n\times n}$ is a diagonal matrix.

It is important to emphasize that the *same* regularization parameters λ and λ_j's in defining $\widehat{\mathbf{b}}$ should be used as in defining $\widetilde{\mathbf{b}}$ in (A.256) so that their constructions differ only by the used design matrix; this plays a key role in applying our coupling technique.

Indeed, we prove in Lemma 21 that the coupling technique together with Condition 6 and some other regularity conditions ensures that with asymptotic probability one,

$$\max_{1\le j\le 2p}|\widetilde{b}_j - \widehat{b}_j| \lesssim \Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}. \tag{A.258}$$

The above result guarantees that $\widehat{W}_j$'s and $\widetilde{W}_j$'s are also uniformly close over $1 \le j \le p$ with $\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j| \lesssim \Delta_n s\sqrt{(\log p)/n} + s^{3/2}(\log p)/n$. As long as $s\Delta_n \to 0$ and $s^{3/2}\sqrt{(\log p)/n} \to 0$, this upper bound has a smaller order than the concentration rate δ_n of $\widetilde{W}_j$ (cf. Condition 2), because here $\delta_n \sim \sqrt{n^{-1}\log p}$ as shown in our Lemma 22. As commented after Theorem 2, the assumption that the coupling rate of $\max_{1\le j\le p}|\widehat{W}_j - \widetilde{W}_j|$ is of a smaller order than the concentration rate δ_n plays a key role in establishing our theory on the asymptotic FDR control.

We next introduce some additional notation and formally present the regularity conditions specific to this section. Let $\mathbf{D} = \text{diag}(\ddot{\rho}(y_1; \widetilde{\mathbf{X}}_{1,\cdot}^{\text{aug}}\boldsymbol{\beta}^0), \cdots, \ddot{\rho}(y_n; \widetilde{\mathbf{X}}_{n,\cdot}^{\text{aug}}\boldsymbol{\beta}^0)) \in \mathbb{R}^{n\times n}$ be a diagonal matrix and $\mathbf{U} = \mathbf{D}^{1/2}\widetilde{\mathbf{X}}^{\text{aug}}$ the weighted perfect design matrix. We define $\boldsymbol{\Sigma} = n^{-1}\mathbb{E}[\mathbf{U}^T\mathbf{U}] = n^{-1}\mathbb{E}[(\widetilde{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widetilde{\mathbf{X}}^{\text{aug}}]$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\gamma}_j = (\gamma_{j,l})_{l\ne j}$ with $\gamma_{j,l} = -\boldsymbol{\Omega}_{j,l}/\boldsymbol{\Omega}_{j,j}$. For $1 \le j \le 2p$, denote by $\mathcal{S}_j = \text{supp}(\boldsymbol{\gamma}_j) \cup \text{supp}(\widetilde{\boldsymbol{\gamma}}_j) \cup \text{supp}(\widehat{\boldsymbol{\gamma}}_j)$. Let $J = \text{supp}(\boldsymbol{\beta}^0) \cup \text{supp}(\widetilde{\boldsymbol{\beta}}) \cup \text{supp}(\widehat{\boldsymbol{\beta}})$ and $s := \|\boldsymbol{\beta}^0\|_0 = \|\boldsymbol{\alpha}^0\|_0 = o(n)$. We make the technical assumptions below.

CONDITION 17. *For a large constant $r > 0$, it holds with probability $1 - O(p^{-r})$ that*

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_1 \le Cs\sqrt{\frac{\log p}{n}}, \tag{A.259}$$

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2 \le C\sqrt{\frac{s\log p}{n}}, \tag{A.260}$$

$$\|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2 \le C\sqrt{s\log p}. \tag{A.261}$$

CONDITION 18. *For a large constant $r > 0$, it holds with probability $1 - O(p^{-r})$ that*

$$\max_{1\le j\le 2p} \|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_1 \le C(s + m_n)\sqrt{\frac{\log p}{n}}, \tag{A.262}$$

$$\max_{1\le j\le 2p} \|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\|_2 \le C\sqrt{\frac{(s + m_n)\log p}{n}}, \tag{A.263}$$

$$\max_{1\le j\le 2p} \|\widetilde{\mathbf{X}}^{\text{aug}}_{-j}(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)\|_2 \le C\sqrt{(s + m_n)\log p}, \tag{A.264}$$

$$\max_{1\le j\le 2p} |\widetilde{\tau}_j^2 - \boldsymbol{\Omega}^{-1}_{j,j}| \le C\sqrt{\frac{(s + m_n)\log p}{n}}, \tag{A.265}$$

$$\begin{aligned}&\max_{1\le j,l\le 2p} \left| n^{-1}(\widetilde{\mathbf{X}}^{\text{aug}}_j - \widetilde{\mathbf{X}}^{\text{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)^T\mathbf{D}(\widetilde{\mathbf{X}}^{\text{aug}}_l - \widetilde{\mathbf{X}}^{\text{aug}}_{-l}\widetilde{\boldsymbol{\gamma}}_l) - \frac{\boldsymbol{\Omega}_{j,l}}{\boldsymbol{\Omega}_{j,j}\boldsymbol{\Omega}_{l,l}} \right| \\ &\le C\sqrt{\frac{(s + m_n)\log p}{n}},\end{aligned} \tag{A.266}$$

where m_n is the sparsity level of $\boldsymbol{\Omega}$ defined in Condition 20.

Conditions 17 and 18 are well-known results about the the consistency of the GLM Lasso estimator and hold under some regularity conditions (van de Geer et al. (2014)).

CONDITION 19 (Loss function). *The derivatives $\dot{\rho}(y;a) := \frac{\partial}{\partial a}\rho(y;a)$ and $\ddot{\rho}(y;a) := \frac{\partial^2}{\partial a^2}\rho(y;a)$ exist for all (y,a), and for some δ-neighborhood with $\delta > 0$, $\ddot{\rho}(y;a)$ is Lipschitz such that*

$$\max_{a_0\in\{X^T\boldsymbol{\alpha}^0\}} \sup_{|a-a_0|\vee|a'-a_0|\le\delta} \sup_{y\in\mathcal{Y}} \frac{|\ddot{\rho}(y;a) - \ddot{\rho}(y;a')|}{|a - a'|} \le C_4,$$

where $\mathcal{Y}$ is the space in which the response variable Y lives. In addition, the derivatives are bounded such that for constants $K_1, K_2 > 0$,

$$\max_{a_0\in\{X^T\boldsymbol{\alpha}^0\}} \sup_{y\in\mathcal{Y}} |\dot{\rho}(y;a_0)| \le K_1, \quad \max_{a_0\in\{X^T\boldsymbol{\alpha}^0\}} \sup_{y\in\mathcal{Y}} |\ddot{\rho}(y;a)| \le K_2, \quad \min_{a_0\in\{X^T\boldsymbol{\alpha}^0\}} \min_{y\in\mathcal{Y}} |\ddot{\rho}(y;a)| \ge K_3.$$

CONDITION 20 (Sparsity). *(i) For some constant $C_5 > 0$, $\mathbb{P}(|J| \le C_5 s) \to 1$.*
(ii) For some sequence $m_n \lesssim s$, it holds that $\max_{1\le j\le 2p} \|\boldsymbol{\Omega}_j\|_0 \le m_n$ and $\mathbb{P}(\max_{1\le j\le 2p} |\mathcal{S}_j| \le C_6 m_n) \to 1$ with some constant $C_6 > 0$.
(iii) $\max_{1\le j\le 2p} \|\boldsymbol{\gamma}_j\|_2 \le C_7$ and $C_8 < \lambda_{\min}(\boldsymbol{\Omega}) \le \lambda_{\max}(\boldsymbol{\Omega}) < C_9$ with some positive constants C_7, C_8, and C_9.

CONDITION 21 (Compatibility). *Assume that with probability* $1 - o(1)$,

$$\min_{\|\boldsymbol{\beta}\|_0 \le C_9 s} \frac{\boldsymbol{\beta}^T \mathbf{U}^T \mathbf{U} \boldsymbol{\beta}}{n\|\boldsymbol{\beta}\|_2^2} \ge \kappa_1 \tag{A.267}$$

for some large enough constant $C_9 > 0$ *and a small constant* $\kappa_1 > 0$.

CONDITION 22 (boundedness). *Assume that* $\|\widetilde{\mathbf{X}}\|_\infty = \max_{i,j} |\widetilde{\mathbf{X}}_{i,j}| \le M$ *for a constant* $M > 0$. *In addition,* $\|\widetilde{\mathbf{X}}^{\text{aug}}_{-j} \boldsymbol{\gamma}_j\|_\infty \le M$.

Note that the boundedness assumption in Condition 22 is for technical simplicity; it can be replaced with a less stringent sub-Gaussian condition and the results in Theorem 7 remain to hold.

CONDITION 23 (Signal strength). *Let* $\mathscr{A}_n = \{j \in \mathcal{H}_1 : |\beta_j^0| \gg \sqrt{n^{-1}\log p}\}$ *and it holds that* $a_n := |\mathscr{A}_n| \to \infty$.

We are now ready to state our results on the FDR control for the approximate knockoffs inference based on the debiased Lasso coefficients for GLM.

THEOREM 7. *Assume that Conditions 6, 10, and 17–23 hold,* $m_n/a_n \to 0$, *and* $\frac{s^{3/2}(\log p)^{3/2+1/\gamma}}{\sqrt{n}} + \Delta_n s(\log p)^{1+1/\gamma} \to 0$ *for some constant* $0 < \gamma < 1$. *Then we have*

$$\limsup_{n\to\infty} \text{FDR} \le q.$$

C.1. Proof of Theorem 7. The main idea of the proof is to directly apply Theorem 1 by verifying Conditions 1–5 for the knockoff statistics constructed from the debiased Lasso coefficients under the GLM. There are two key observations. The first one is that the Lasso estimators based on the approximate knockoffs and the perfect coupling counterpart should be close if the design matrices $\widehat{\mathbf{X}}^{\text{aug}}$ and $\widetilde{\mathbf{X}}^{\text{aug}}$ are close to each other. The second key observation is that the debiased Lasso coefficients are asymptotically normal (van de Geer et al. (2014)). Let $\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0} := (\dot{\rho}(y_1; \widehat{\mathbf{X}}^{\text{aug}}_{1,\cdot}\boldsymbol{\beta}^0), \cdots, \dot{\rho}(y_n; \widehat{\mathbf{X}}^{\text{aug}}_{n,\cdot}\boldsymbol{\beta}^0)) = (\dot{\rho}(y_1; \widetilde{\mathbf{X}}^{\text{aug}}_{1,\cdot}\boldsymbol{\beta}^0), \cdots, \dot{\rho}(y_n; \widetilde{\mathbf{X}}^{\text{aug}}_{n,\cdot}\boldsymbol{\beta}^0)) \in \mathbb{R}^n$. It follows from the Taylor expansion that $\dot{\rho}(y_i; \widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\boldsymbol{\beta}^0) - \dot{\rho}(y_i; \widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}}) = \ddot{\rho}(y_i; \xi)\widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}(\boldsymbol{\beta}^0 - \widetilde{\boldsymbol{\beta}})$ for some ξ locating between $\widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\boldsymbol{\beta}^0$ and $\widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}}$. By Condition 19, we can obtain that

$$\left|\dot{\rho}(y_i; \widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\boldsymbol{\beta}^0) - \dot{\rho}(y_i; \widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}}) - \ddot{\rho}(y_i; \widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}})\widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}(\boldsymbol{\beta}^0 - \widetilde{\boldsymbol{\beta}})\right| \le C_4[\widetilde{\mathbf{X}}^{\text{aug}}_{i,\cdot}(\boldsymbol{\beta}^0 - \widetilde{\boldsymbol{\beta}})]^2. \tag{A.268}$$

In view of (A.255), the debiased Lasso coefficient can be written as

(A.269)
$$\begin{aligned}\sqrt{n}(\widetilde{b}_j-\beta_j^0)&=\sqrt{n}(\widetilde{\beta}_j-\beta_j^0)-\frac{n^{-1/2}\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\\&=-\frac{n^{-1/2}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}+\sqrt{n}(\widetilde{\beta}_j-\beta_j^0)\\&\quad+\frac{n^{-1/2}(\widetilde{\mathbf{D}}\widetilde{\mathbf{X}}^{\mathrm{aug}}(\boldsymbol{\beta}^0-\widetilde{\boldsymbol{\beta}}))^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}+\frac{C_4n^{-1/2}\widetilde{\mathbf{R}}|\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j|}{\widetilde{\tau}_j^2}\\&=-\frac{n^{-1/2}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}+\frac{C_4n^{-1/2}\widetilde{\mathbf{R}}|\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j|}{\widetilde{\tau}_j^2}\\&\quad+\frac{n^{-1/2}(\boldsymbol{\beta}_{-j}^0-\widetilde{\boldsymbol{\beta}}_{-j})^T(\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2},\end{aligned}$$

where $\widetilde{\mathbf{R}}=([\widetilde{\mathbf{X}}_{1,\cdot}^{\mathrm{aug}}(\boldsymbol{\beta}^0-\widetilde{\boldsymbol{\beta}})]^2,\cdots,[\widetilde{\mathbf{X}}_{n,\cdot}^{\mathrm{aug}}(\boldsymbol{\beta}^0-\widetilde{\boldsymbol{\beta}})]^2)$, and we have used the equality $\widetilde{\tau}_j^2=\widetilde{\boldsymbol{\Sigma}}_{j,j}-\widetilde{\boldsymbol{\Sigma}}_{j,-j}\widetilde{\boldsymbol{\gamma}}_j=(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)$.

By the property of GLM, we have $\mathbb{E}[\dot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\boldsymbol{\beta}^0)|\widetilde{\mathbf{X}}^{\mathrm{aug}}]=0$, and hence $\mathbb{E}[\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)]=0$. In addition, it holds that

$$\begin{aligned}\mathrm{Var}[n^{-1/2}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)|\widetilde{\mathbf{X}}^{\mathrm{aug}}]&=n^{-1}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)^T\mathbf{D}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\\&\approx\widetilde{\tau}_j^2.\end{aligned}$$

Thus, as the remainders in (A.269) are asymptotically negligible, the debiased Lasso estimator is asymptotically normal in the sense that

(A.270)
$$\sqrt{n}\widetilde{\tau}_j(\widetilde{b}_j-\beta_j^0)\xrightarrow{d}N(0,1).$$

Our proof will build mainly on such intuition. Throughout the proof below, constant C may take different values from line to line.

The four lemmas below outline the proof for verifying the general Conditions 1–5. Proofs of Lemma 21–24 are provided in Sections C.2–C.5, respectively.

LEMMA 21. *Assume that Conditions 6 and 17–22 are satisfied. Then as* $\Delta_n s^{1/2}\to 0$ *and* $s\sqrt{\frac{\log p}{n}}\to 0$, *we have that*

(A.271)
$$\mathbb{P}\Big(\max_{1\le j\le 2p}|\widetilde{b}_j-\widehat{b}_j|\ge C\Big(\Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n}\Big)\Big)\to 0.$$

Lemma 21 above indicates that Condition 1 is satisfied with convergence rate $b_n:=C(\Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n})$. Let us define $w_j=|\beta_j^0|$.

LEMMA 22. *Assume that Conditions 17–22 are satisfied. Then as* $s^{3/2}\sqrt{\frac{\log p}{n}}\to 0$, *we have that for some* $C>0$, $\sum_{j=1}^p\mathbb{P}(|\widetilde{W}_j-w_j|\ge C\sqrt{n^{-1}\log p})\to 0$.

Lemma 22 above shows that Condition 2 related to the concentration rate of $\widetilde{W}_j$ is satisfied with $\delta_n = C\sqrt{n^{-1}\log p}$. In addition, it holds that $b_n \ll C\sqrt{n^{-1}\log p}$ due to the assumptions $\Delta_n s \to 0$ and $s\sqrt{\frac{\log p}{n}} \to 0$ in Theorem 7. In addition, in light of the definition of w_j, under Condition 23 we have that the general Condition 3 on the signal strength is also satisfied. We next turn to the verification of Conditions 4–5.

LEMMA 23. *Assume that Conditions 17–22 are satisfied. Then as* $\frac{s^{3/2}(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$, *we have that* $\mathrm{Var}\left(\sum_{j\in\mathcal{H}_0} \mathbb{1}(\widetilde{W}_j > t)\right) \le V_1(t) + V_2(t)$, *where for some* $0<\gamma<1$ *and* $0<c_1<1$,

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p})]} \frac{V_1(t)}{[p_0 G(t)]^2} \to 0 \tag{A.272}$$

and

$$\sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p})]} \frac{V_2(t)}{p_0 G(t)} \lesssim m_n. \tag{A.273}$$

LEMMA 24. *Assume that Conditions 6, 10, and 17–22 are satisfied. Then when* $\frac{s^{3/2}(\log p)^{3/2+1/\gamma}}{\sqrt{n}} \to 0$ *and* $\Delta_n s(\log p)^{1+1/\gamma} \to 0$, *we have that*

$$(\log p)^{1/\gamma} \sup_{t\in(0, G^{-1}(\frac{c_1 q a_n}{p})]} \frac{G(t-b_n) - G(t+b_n)}{G(t)} \to 0 \tag{A.274}$$

and

$$a_n^{-1} \sum_{j\in\mathcal{H}_1} \mathbb{P}\Big(\widetilde{W}_j < -G^{-1}(\frac{c_1 q a_n}{p}) + b_n\Big) \to 0 \tag{A.275}$$

as $n\to\infty$.

Lemma 23 above shows that Condition 4 is satisfied, whereas Lemma 24 implies that Condition 5 is satisfied. Finally, the conclusion of Theorem 7 can be derived by directly applying the general Theorem 1. This completes the proof of Theorem 7.

C.2. Proof of Lemma 21. The proof is analogous to that of Lemma 11. The main idea is to apply the KKT condition to the GLM Lasso and then use Condition 6. From the definitions of $\widehat{b}_j$ in (A.255) and the coupled counterpart $\widetilde{b}_j$ in (A.256), we have that

$$\begin{aligned}\max_{1\le j\le 2p} |\widehat{b}_j - \widetilde{b}_j| &\le \max_{1\le j\le 2p} |\widehat{\beta}_j - \widetilde{\beta}_j| \\ &+ \max_{1\le j\le 2p} \left| \frac{n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}} - \widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)}{\widehat{\tau}_j^2} - \frac{n^{-1}|\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}} - \widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2} \right|.\end{aligned} \tag{A.276}$$

We will show that for some constant $C>0$, it holds that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2 \le C\Big(\Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}\Big)\right) \to 1, \tag{A.277}$$

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le j\le 2p} \left| \frac{n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}} - \widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)}{\widehat{\tau}_j^2} - \frac{n^{-1}|\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}}^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}} - \widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2} \right| \right. \\ &\qquad \left. \le C\Big(\Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}\Big)\right) \to 1.\end{aligned} \tag{A.278}$$

Then combining the two results above can establish the desired conclusion of Lemma 21. We next proceed with proving (A.277) and (A.278).

Proof of (A.277)**.** Recall the definitions of Lasso estimators $\widehat{\boldsymbol{\beta}}$ in (A.254) and $\widetilde{\boldsymbol{\beta}}$ in (A.257). It follows from the KKT condition that

$$n^{-1}\sum_{i=1}^{n}\dot{\rho}(y_i;\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}})(\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}})^T+\lambda\widehat{\boldsymbol{\zeta}}=0, \tag{A.279}$$

$$n^{-1}\sum_{i=1}^{n}\dot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}})(\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}})^T+\lambda\widetilde{\boldsymbol{\zeta}}=0, \tag{A.280}$$

where $\widetilde{\boldsymbol{\zeta}}=(\widetilde{\zeta}_1,\cdots,\widetilde{\zeta}_{2p})$ and $\widehat{\boldsymbol{\zeta}}=(\widehat{\zeta}_1,\cdots,\widehat{\zeta}_{2p})$ with

$$\widetilde{\zeta}_j=\begin{cases}\text{sgn}(\widetilde{\beta}_j) & \text{if } \widetilde{\beta}_j\neq 0,\\ \in[-1,1] & \text{if } \widetilde{\beta}_j=0,\end{cases}\quad\text{and}\quad \widehat{\zeta}_j=\begin{cases}\text{sgn}(\widehat{\beta}_j) & \text{if } \widehat{\beta}_j\neq 0,\\ \in[-1,1] & \text{if } \widehat{\beta}_j=0.\end{cases}$$

Taking the difference between (A.279) and (A.279) above and multiplying both sides by $\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}$ lead to

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n}\dot{\rho}(y_i;\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}\widehat{\boldsymbol{\beta}})(\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}})(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})-n^{-1}\sum_{i=1}^{n}\dot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}\widetilde{\boldsymbol{\beta}})(\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}})(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\\ &=-\lambda(\widehat{\boldsymbol{\zeta}}-\widetilde{\boldsymbol{\zeta}})^T(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\leq 0.\end{aligned}$$

Further applying the Taylor expansion for function $\dot{\rho}$ and Condition 19 yields

$$\begin{aligned}&n^{-1}\sum_{i=1}^{n}\big[\dot{\rho}(y_i;\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}\boldsymbol{\beta}^0)+\ddot{\rho}(y_i;\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}\boldsymbol{\beta}^0)\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\big]\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\\ &-n^{-1}\sum_{i=1}^{n}\big[\dot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}\boldsymbol{\beta}^0)+\ddot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}\boldsymbol{\beta}^0)\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\big]\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\\ &\leq C_4n^{-1}\sum_{i=1}^{n}\big|\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\big|^3,\end{aligned} \tag{A.281}$$

which can be equivalently written in the matrix form as

$$\begin{aligned}&n^{-1}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^T(\widehat{\mathbf{X}}^{\text{aug}}-\widetilde{\mathbf{X}}^{\text{aug}})^T\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}+n^{-1}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^T(\widetilde{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widetilde{\mathbf{X}}^{\text{aug}}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\\ &+n^{-1}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^T[(\widehat{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widehat{\mathbf{X}}^{\text{aug}}-(\widetilde{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widetilde{\mathbf{X}}^{\text{aug}}](\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\\ &+n^{-1}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})^T[(\widehat{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widehat{\mathbf{X}}^{\text{aug}}-(\widetilde{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\widetilde{\mathbf{X}}^{\text{aug}}](\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\\ &\leq C_4n^{-1}\sum_{i=1}^{n}\big|\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\big|^3.\end{aligned} \tag{A.282}$$

Note that by Condition 20, $|\text{supp}(\widehat{\boldsymbol{\beta}})\cup\text{supp}(\widetilde{\boldsymbol{\beta}})\cup\text{supp}(\boldsymbol{\beta}^0)|\leq Cs$ with probability approaching one. Thus, by a similar technique of proving (A.146), we can obtain from Condi-

tions 20 and 21 that

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}\|_2 \lesssim & \max_{J:|J|\leq Cs}\left\|n^{-1}(\widehat{\mathbf{X}}_J^{\text{aug}}-\widetilde{\mathbf{X}}_J^{\text{aug}})^T\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}\right\|_2 \\ & + \max_{J:|J|\leq Cs}\left\|n^{-1}[(\widehat{\mathbf{X}}_J^{\text{aug}})^T\mathbf{D}\widehat{\mathbf{X}}^{\text{aug}}-(\widetilde{\mathbf{X}}_J^{\text{aug}})^T\mathbf{D}\widetilde{\mathbf{X}}^{\text{aug}}](\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\right\|_2 \\ & + \max_{J:|J|\leq Cs} n^{-1}\sum_{i=1}^{n}\left(\widetilde{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\right)^2\left\|\widetilde{\mathbf{X}}_{i,J}\right\|_2 := R_1+R_2+R_3.\end{aligned} \tag{A.283}$$

Observe that given $(\mathbf{X},\widetilde{\mathbf{X}})$, $\dot{\rho}_{\boldsymbol{\beta}^0}$ is a vector consisting of i.i.d bounded random variables with zero mean and bounded variance. Following the same technique of proving (A.147) and (A.148), we can obtain that

$$\mathbb{P}\left(R_1 \leq C\Delta_n\sqrt{\frac{s\log n}{n}}\right)\to 1 \tag{A.284}$$

and

$$\mathbb{P}\left(R_2 \leq C\Delta_n s\sqrt{\frac{\log p}{n}}\right)\to 1. \tag{A.285}$$

Regarding R_3, it follows from Conditions 17 and 22 that with probability $1-o(1)$,

$$R_3 \leq C\sqrt{s}Mn^{-1}\|\widetilde{\mathbf{X}}^{\text{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2 \leq C\sqrt{s}Mn^{-1}s\log p \leq C\frac{s^{3/2}\log p}{n}. \tag{A.286}$$

Combining (A.284)–(A.286) derives (A.277). Further, applying (A.282) again with the bounds in (A.284)–(A.286) and (A.277) yields that

$$\mathbb{P}\left(n^{-1/2}\|\mathbf{D}^{1/2}\widetilde{\mathbf{X}}^{\text{aug}}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\|_2 \leq C\left(\Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n}\right)\right)\to 1. \tag{A.287}$$

Therefore, it follows by Condition 19 that

$$\mathbb{P}\left(n^{-1/2}\|\widetilde{\mathbf{X}}^{\text{aug}}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\|_2 \leq C\left(\Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n}\right)\right)\to 1. \tag{A.288}$$

Proof of (A.278)**.** Observe that $\widehat{\boldsymbol{\gamma}}_j$ and $\widetilde{\boldsymbol{\gamma}}_j$ can be equivalently written as

$$\widehat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}\in\mathbb{R}^{2p-1}}{\arg\min}\|\widehat{\mathbf{D}}^{1/2}\widehat{\mathbf{X}}_j^{\text{aug}}-\widehat{\mathbf{D}}^{1/2}\widehat{\mathbf{X}}_{-j}^{\text{aug}}\boldsymbol{\gamma}\|_2^2+\lambda_j\|\boldsymbol{\gamma}\|_1. \tag{A.289}$$

In addition, it can be obtained from Conditions 17, 19, 20, and 22 that with probability $1-o(1)$,

$$|\ddot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}\widetilde{\boldsymbol{\beta}})-\ddot{\rho}(y_i;\widetilde{\mathbf{X}}_{i,\cdot}\boldsymbol{\beta}^0)| \leq C|\widetilde{\mathbf{X}}_{i,\cdot}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)| \leq CM\sqrt{s}\sqrt{s\frac{\log p}{n}} = CMs\sqrt{\frac{\log p}{n}}\to 0.$$

Hence, under Conditions 6, 17, 19, 20, and 22, we have that with probability $1-o(1)$,

$$\begin{aligned}
&n^{-1/2}\|\widehat{\mathbf{D}}^{1/2}\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{D}}^{1/2}\widetilde{\mathbf{X}}_j^{\mathrm{aug}}\|_2\\
&\le n^{-1/2}\|(\widehat{\mathbf{D}}^{1/2}-\widetilde{\mathbf{D}}^{1/2})\widehat{\mathbf{X}}_j^{\mathrm{aug}}\|_2+n^{-1/2}\|\widetilde{\mathbf{D}}^{1/2}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})\|_2\\
&\le CMn^{-1/2}\|\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}\|_2+n^{-1/2}\|\mathbf{D}^{1/2}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})\|_2\\
&\quad+n^{-1/2}\|(\widetilde{\mathbf{D}}^{1/2}-\mathbf{D}^{1/2})(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})\|_2 \qquad \text{(A.290)}\\
&\le CMn^{-1/2}\|(\widehat{\mathbf{X}}^{\mathrm{aug}}-\widetilde{\mathbf{X}}^{\mathrm{aug}})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2+CMn^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}})\|_2\\
&\quad+n^{-1/2}\|\mathbf{D}^{1/2}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})\|_2+n^{-1/2}\|(\widetilde{\mathbf{D}}^{1/2}-\mathbf{D}^{1/2})(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})\|_2\\
&\lesssim \Delta_n s\sqrt{\frac{\log p}{n}}+\Delta_n\Big(1+s\sqrt{\frac{\log p}{n}}\Big)\lesssim\Delta_n.
\end{aligned}$$

Consequently, using similar argument as for (A.277) and (A.287), we can obtain that

$$\mathbb{P}\bigg(\max_{1\le j\le 2p}\|\widetilde{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_j\|_2\le Cm_n^{1/2}\Delta_n\bigg)\to 1, \qquad \text{(A.291)}$$

$$\mathbb{P}\bigg(n^{-1/2}\max_{1\le j\le 2p}n^{-1/2}\|\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}(\widetilde{\boldsymbol{\gamma}}_j-\widehat{\boldsymbol{\gamma}}_j)\|_2\le Cm_n^{1/2}\Delta_n\bigg)\to 1. \qquad \text{(A.292)}$$

Moreover, by similar arguments as for (A.152), (A.185), and (A.186), we can deduce that with probability $1-o(p^{-1})$,

$$\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\|_2\lesssim\Delta_n m_n^{1/2}, \qquad \text{(A.293)}$$

$$\min_{1\le j\le p}\widetilde{\tau}_j^2=\min_{1\le j\le p}n^{-1}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\ge C, \qquad \text{(A.294)}$$

and

$$\max_{1\le j\le p}\max_{k\ne j}n^{-1}\big|(\widetilde{\mathbf{X}}_k^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big|\le C\sqrt{\frac{(m_n+s)\log p}{n}}. \qquad \text{(A.295)}$$

Now we are ready to establish (A.278). Specifically, the main term in (A.278) can be decomposed into the three terms below

$$\begin{aligned}
&\max_{1\le j\le 2p}\bigg|\frac{n^{-1}(\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}-\dot{\boldsymbol{\rho}}_{\widetilde{\boldsymbol{\beta}}})^T(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\bigg|\\
&+\max_{1\le j\le 2p}\bigg|\frac{n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T\big(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)}{\widetilde{\tau}_j^2}\bigg| \qquad \text{(A.296)}\\
&+\max_{1\le j\le 2p}\bigg|n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\Big(\frac{1}{\widetilde{\tau}_j^2}-\frac{1}{\widehat{\tau}_j^2}\Big)\bigg|:=I_1+I_2+I_3.
\end{aligned}$$

We will deal with the three terms I_1, I_2, and I_3 separately. First for I_1, it follows from Condition 19 that

$$\begin{aligned}
I_1&\le\max_{1\le j\le 2p}\bigg|\frac{n^{-1}[(\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}})]^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\bigg|\\
&\quad+C\max_{1\le j\le 2p}\bigg|\frac{n^{-1}\sum_{i=1}^n(\widehat{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}})^2|\widetilde{\mathbf{X}}_{i,j}^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{i,-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j|}{\widetilde{\tau}_j^2}\bigg|:=I_{11}+I_{12}. \qquad \text{(A.297)}
\end{aligned}$$

Regarding I_{12}, in view of Conditions 18 and 22, and (A.294), we have

$$\begin{aligned} I_{12} &\le C \max_{1\le j\le 2p} n^{-1}\sum_{i=1}^{n}(\widehat{\mathbf{X}}^{\mathrm{aug}}_{i,\cdot}\widehat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}})^2|\widetilde{\mathbf{X}}^{\mathrm{aug}}_{i,j} - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{i,-j}\boldsymbol{\gamma}_j| \\ &\quad + C \max_{1\le j\le 2p} n^{-1}\sum_{i=1}^{n}(\widehat{\mathbf{X}}^{\mathrm{aug}}_{i,\cdot}\widehat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{i,\cdot}\widetilde{\boldsymbol{\beta}})^2|\widetilde{\mathbf{X}}^{\mathrm{aug}}_{i,-j}(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j)| \\ &\le CM\left(1 + m_n^{1/2}\sqrt{\frac{(s+m_n)\log p}{n}}\right)n^{-1}\|\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}\|_2^2 \\ &\le CMn^{-1}\|\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}\|_2^2, \end{aligned} \tag{A.298}$$

where we have applied the assumption that $s\sqrt{\frac{\log p}{n}} \to 0$ and $m_n \lesssim s$. In addition, noting that $(\widehat{\mathbf{X}}^{\mathrm{aug}} - \widetilde{\mathbf{X}}^{\mathrm{aug}})\boldsymbol{\beta}^0 = \mathbf{0}$ by definition, we obtain from (A.288) and Condition 17 that with probaility $1 - o(1)$,

$$\begin{aligned} n^{-1}\|\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}} - \widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}\|_2^2 &\le n^{-1}\|\widetilde{\mathbf{X}}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 + n^{-1}\|(\widehat{\mathbf{X}}^{\mathrm{aug}} - \widetilde{\mathbf{X}}^{\mathrm{aug}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2^2 \\ &\lesssim \frac{\Delta_n^2 s^2\log p}{n} + \frac{s^3(\log p)^2}{n^2} + \Delta_n^2 s\frac{s\log p}{n} \\ &\lesssim \frac{\Delta_n^2 s^2\log p}{n} + \frac{s^3(\log p)^2}{n^2}, \end{aligned} \tag{A.299}$$

which together with (A.298) yields

$$I_{12} \le C\left(\frac{\Delta_n^2 s^2\log p}{n} + \frac{s^3(\log p)^2}{n^2}\right). \tag{A.300}$$

Now we proceed with examining I_{11}. Observe that it admits the decomposition

$$\begin{aligned} I_{11} &\le \max_{1\le j\le 2p}\left|\frac{n^{-1}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})^T(\widetilde{\mathbf{X}}^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\right| \\ &\quad + \max_{1\le j\le 2p}\left|\frac{n^{-1}\widehat{\boldsymbol{\beta}}^T(\widetilde{\mathbf{X}}^{\mathrm{aug}} - \widehat{\mathbf{X}}^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\right| \\ &\le \max_{1\le j\le 2p}|\widetilde{\beta}_j - \widehat{\beta}_j| + \max_{1\le j\le 2p}\left|n^{-1}(\widetilde{\boldsymbol{\beta}}_{-j} - \widehat{\boldsymbol{\beta}}_{-j})^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)\right| \\ &\quad + \max_{1\le j\le 2p} n^{-1}\left|\widehat{\boldsymbol{\beta}}^T(\widetilde{\mathbf{X}}^{\mathrm{aug}} - \widehat{\mathbf{X}}^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)\right| \\ &:= I_{111} + I_{112} + I_{113}. \end{aligned} \tag{A.301}$$

As for I_{112}, it follows from (A.277) and (A.295) that with probability $1 - o(1)$,

$$\begin{aligned} I_{112} &\le \max_{1\le j\le 2p}\max_{J:|J|\le Cs} n^{-1}\big\|\widetilde{\boldsymbol{\beta}}_{-j} - \widehat{\boldsymbol{\beta}}_{-j}\big\|_2\big\|(\widetilde{\mathbf{X}}^{\mathrm{aug}}_{J\setminus\{j\}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)\big\|_2 \\ &\le Cs^{1/2}\big\|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\big\|_2\max_{1\le j\le 2p}\max_{k\ne j} n^{-1}\big|(\widetilde{\mathbf{X}}^{\mathrm{aug}}_k)^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j - \widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)\big| \\ &\le s\sqrt{\frac{\log p}{n}}\left(\Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}\right) \lesssim \Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}, \end{aligned} \tag{A.302}$$

where we have used the assumption that $s\sqrt{\frac{\log p}{n}} \to 0$.

Regarding I_{113}, noting that $(\widehat{\mathbf{X}}^{\text{aug}} - \widetilde{\mathbf{X}}^{\text{aug}})\boldsymbol{\beta}^0 = \mathbf{0}$, by a similar argument as for (A.168), we have that with probability $1 - o(1)$,

$$\begin{aligned} I_{113} &= \max_{1\le j\le 2p} n^{-1}\Big|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\Big| \\ &\le n^{-1/2}\|(\widetilde{\mathbf{X}}^{\text{aug}} - \widehat{\mathbf{X}}^{\text{aug}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_2 \max_{1\le j\le 2p} n^{-1/2}\|\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\|_2 \\ &\lesssim \Delta_n s\sqrt{\frac{\log p}{n}}. \end{aligned} \tag{A.303}$$

Combining (A.277), (A.301), (A.302), and (A.303), we can derive that with probability $1 - o(1)$,

$$I_{11} \lesssim \Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}, \tag{A.304}$$

which together with (A.297) and (A.300) gives that

$$\mathbb{P}\bigg(I_1 \le C\Big(\Delta_n s\sqrt{\frac{\log p}{n}} + \frac{s^{3/2}\log p}{n}\Big)\bigg) \to 1. \tag{A.305}$$

Next we turn to I_2 in (A.296). Applying the Taylor expansion and Condition 19, we can obtain that

$$\begin{aligned} I_2 &\lesssim \max_{1\le j\le 2p}\Big|n^{-1}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)\Big| \\ &\quad + \max_{1\le j\le 2p}\Big|n^{-1}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^T(\widehat{\mathbf{X}}^{\text{aug}})^T\mathbf{D}\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)\Big| \\ &\quad + \max_{1\le j\le 2p} n^{-1}\sum_{i=1}^n[\widehat{\mathbf{X}}_{i,\cdot}^{\text{aug}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)]^2\big|\widehat{\mathbf{X}}_{i,j}^{\text{aug}} - \widehat{\mathbf{X}}_{i,-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_{i,j} - (\widetilde{\mathbf{X}}_{i,j}^{\text{aug}} - \widetilde{\mathbf{X}}_{i,-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big| \\ &:= I_{21} + I_{22} + I_{23}. \end{aligned} \tag{A.306}$$

Note that $\mathbb{E}[\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}|(\mathbf{X}, \widetilde{\mathbf{X}}, \widehat{\mathbf{X}})] = \mathbf{0}$ and with probability $1 - o(p^{-1})$,

$$\begin{aligned} &\mathrm{Var}(n^{-1}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)|(\mathbf{X}, \widetilde{\mathbf{X}}, \widehat{\mathbf{X}})) \\ &= n^{-2}\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)^T\mathbf{D}\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big) \\ &\lesssim n^{-2}\|\widehat{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_j^{\text{aug}} + \widetilde{\mathbf{X}}_{-j}^{\text{aug}}(\widetilde{\boldsymbol{\gamma}}_j - \widehat{\boldsymbol{\gamma}}_j) + (\widetilde{\mathbf{X}}_{-j}^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}})\widehat{\boldsymbol{\gamma}}_j\|_2^2 \\ &\lesssim n^{-1}(\Delta_n^2 + \Delta_n^2 m_n + \Delta_n^2 m_n) \lesssim n^{-1}\Delta_n^2 m_n. \end{aligned}$$

Since the components of $\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}$ are all bounded by K_1 under Condition 19, we see that $n^{-1}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T\big(\widehat{\mathbf{X}}_j^{\text{aug}} - \widehat{\mathbf{X}}_{-j}^{\text{aug}}\widehat{\boldsymbol{\gamma}}_j - (\widetilde{\mathbf{X}}_j^{\text{aug}} - \widetilde{\mathbf{X}}_{-j}^{\text{aug}}\widetilde{\boldsymbol{\gamma}}_j)\big)$ is sub-Gaussian, which entails that with probability $1 - o(1)$,

$$I_{21} \le C\Delta_n\sqrt{\frac{m_n\log p}{n}}. \tag{A.307}$$

In the same manner of proving (A.156), we can show that with probability $1 - o(1)$,

$$I_{22} \lesssim \sqrt{\frac{s\log p}{n}}\Delta_n m_n^{1/2} \lesssim \Delta_n s\sqrt{\frac{\log p}{n}}. \tag{A.308}$$

Regarding I_{23}, it holds that with probability $1-o(1)$,

$$\begin{aligned} I_{23} &\le n^{-1}\|\widehat{\mathbf{X}}^{\mathrm{aug}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\|_2 \\ &\le \frac{s\log p}{n}\Delta_n m_n^{1/2} \lesssim \Delta_n s\sqrt{\frac{\log p}{n}}. \end{aligned} \tag{A.309}$$

A combination of (A.306)–(A.309) leads to

$$\mathbb{P}\left(I_2 \le C\Delta_n s\sqrt{\frac{\log p}{n}}\right)\to 1. \tag{A.310}$$

Now we proceed to deal with I_3 in (A.296). Note that

$$I_3 \le \max_{1\le j\le 2p}\left|n^{-1}\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right|\cdot\max_{1\le j\le 2p}\frac{|\widehat{\tau}_j^2-\widetilde{\tau}_j^2|}{|\widetilde{\tau}_j^2\widehat{\tau}_j^2|} := I_{31}\cdot I_{32}. \tag{A.311}$$

It can be seen that

$$I_{31} \lesssim \max_{1\le j\le 2p}\left|n^{-1}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| + \max_{1\le j\le 2p}\left|n^{-1}(\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}-\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0})^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right|.$$

By a similar argument as for (A.307), we can show that with probability $1-o(1)$,

$$\max_{1\le j\le 2p}\left|n^{-1}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| \lesssim \sqrt{\frac{\log p}{n}}.$$

In addition, we have under Condition 19 that with probability $1-o(1)$,

$$\begin{aligned} &\max_{1\le j\le 2p}\left|n^{-1}(\dot{\boldsymbol{\rho}}_{\widehat{\boldsymbol{\beta}}}-\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0})^T(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| \\ &\lesssim \max_{1\le j\le 2p}|n^{-1}(\widehat{\mathbf{X}}^{\mathrm{aug}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0))^T\mathbf{D}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)| \\ &\lesssim \max_{1\le j\le 2p} n^{-1}\|\widehat{\mathbf{X}}^{\mathrm{aug}}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j\|_2 \lesssim \sqrt{\frac{s\log p}{n}}. \end{aligned}$$

Thus, we can obtain that with probability $1-o(1)$,

$$I_{31} \lesssim \sqrt{\frac{s\log p}{n}}. \tag{A.312}$$

As for I_{32}, by definition it holds that

$$\begin{aligned} |\widehat{\tau}_j^2-\widetilde{\tau}_j^2| &= n^{-1}\left|(\widehat{\mathbf{X}}_j^{\mathrm{aug}})^T\widehat{\mathbf{D}}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\right| \\ &\le n^{-1}\left|(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widehat{\mathbf{D}}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| \\ &\quad + n^{-1}\left|(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T(\widehat{\mathbf{D}}-\widetilde{\mathbf{D}})(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| \\ &\quad + n^{-1}\left|(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j))\right|. \end{aligned} \tag{A.313}$$

Furthermore, it can be shown that with probability $1-o(1)$,

$$\begin{aligned} &n^{-1}\left|(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widehat{\mathbf{D}}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\right| \\ &\lesssim n^{-1}\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_j^{\mathrm{aug}}\|_2\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j\|_2 \\ &\lesssim \Delta_n \end{aligned} \tag{A.314}$$

and

$$
\begin{aligned}
&n^{-1}\big|(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T(\widehat{\mathbf{D}}-\widetilde{\mathbf{D}})(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j)\big| \\
&\le n^{-1}\sum_{i=1}^{n}|\widehat{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}||\widetilde{\mathbf{X}}_{i,j}||\widehat{\mathbf{X}}_{i,j}^{\mathrm{aug}}-\widehat{\mathbf{X}}_{i,-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j| \\
&\le n^{-1}M\sum_{i=1}^{n}|\widehat{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}_{i,\cdot}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}||\widehat{\mathbf{X}}_{i,j}^{\mathrm{aug}}-\widehat{\mathbf{X}}_{i,-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j| \\
&\le n^{-1}M\|\widehat{\mathbf{X}}^{\mathrm{aug}}\widehat{\boldsymbol{\beta}}-\widetilde{\mathbf{X}}^{\mathrm{aug}}\widetilde{\boldsymbol{\beta}}\|_2\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j\|_2 \\
&\lesssim \Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n},
\end{aligned}
\tag{A.315}
$$

where we have applied the bound obtained in (A.299).

Moreover, we have that

$$
\begin{aligned}
&n^{-1}\big|(\widetilde{\mathbf{X}}_j^{\mathrm{aug}})^T\widetilde{\mathbf{D}}(\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j))\big| \\
&\lesssim n^{-1}\|\widetilde{\mathbf{X}}_j^{\mathrm{aug}}\|_2\|\widehat{\mathbf{X}}_j^{\mathrm{aug}}-\widehat{\mathbf{X}}_{-j}^{\mathrm{aug}}\widehat{\boldsymbol{\gamma}}_j-(\widetilde{\mathbf{X}}_j^{\mathrm{aug}}-\widetilde{\mathbf{X}}_{-j}^{\mathrm{aug}}\widetilde{\boldsymbol{\gamma}}_j)\|_2 \\
&\lesssim \Delta_n m_n^{1/2},
\end{aligned}
\tag{A.316}
$$

which together with (A.294), (A.313), (A.314), and (A.316) yields that with probability $1-o(1)$,

$$
I_{32}=\max_{1\le j\le 2p}\frac{|\widehat{\tau}_j^2-\widetilde{\tau}_j^2|}{|\widetilde{\tau}_j^2\widehat{\tau}_j^2|}\lesssim \Delta_n m_n^{1/2}+\frac{s^{3/2}\log p}{n}. \tag{A.317}
$$

Combining (A.311), (A.312), and (A.317) leads to

$$
\mathbb{P}\bigg(I_3\le C\Big(\Delta_n s\sqrt{\frac{\log p}{n}}+\frac{s^{3/2}\log p}{n}\Big)\bigg)\to 1. \tag{A.318}
$$

Consequently, substituting (A.305), (A.310), and (A.318) into (A.296) gives the desired result (A.278). This completes the proof of Lemma 21.

C.3. Proof of Lemma 22. The main idea of the proof is to bound the remainders in the decomposition of $\sqrt{n}(\widetilde{b}_j-\beta_j^0)$ as presented in (A.269) and use the fact that the main term is sub-Gaussian. Note that by the triangle inequality and the fact that $w_j=|\beta_j^0|=|\beta_j^0|-|\beta_{j+p}^0|$, it holds that

$$
\begin{aligned}
&\sum_{j=1}^{p}\mathbb{P}\big(|\widetilde{W}_j-w_j|\ge C\sqrt{n^{-1}\log p}\big) \\
&\le \sum_{j=1}^{p}\Big[\mathbb{P}(\sqrt{n}|\widetilde{b}_j-\beta_j^0|\ge C\sqrt{\log p}/2)+\mathbb{P}(\sqrt{n}|\widetilde{b}_{j+p}-\beta_{j+p}^0|\ge C\sqrt{\log p}/2)\Big] \\
&=\sum_{j=1}^{2p}\mathbb{P}(\sqrt{n}|\widetilde{b}_j-\beta_j^0|\ge C\sqrt{\log p}/2).
\end{aligned}
\tag{A.319}
$$

For the second remainder in (A.269), applying the bounds in (A.295), (A.294), and (A.259), we have that with probability $1-o(p^{-3})$,

$$\begin{aligned}
&\max_{1\le j\le p}\frac{n^{-1/2}(\boldsymbol{\beta}^0_{-j}-\widetilde{\boldsymbol{\beta}}_{-j})^T(\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j})^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)}{\widetilde{\tau}_j^2}\\
&\le \max_{1\le j\le p} n^{-1/2}\sum_{k\neq j}\frac{n^{-1/2}|(\widetilde{\mathbf{X}}^{\mathrm{aug}}_k)^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)|\widetilde{\beta}_k-\beta^0_k|}{\widetilde{\tau}_j^2}\\
&\le \max_{1\le j\le p}\max_{k\neq j} n^{-1/2}|(\widetilde{\mathbf{X}}^{\mathrm{aug}}_k)^T\widetilde{\mathbf{D}}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)|\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1\\
&\lesssim \frac{s^{3/2}\log p}{\sqrt{n}}.
\end{aligned}\tag{A.320}$$

Regarding the first remainder in (A.269), applying (A.261), (A.264), (A.294), and the fact that $\|\widetilde{\mathbf{X}}_j-\widetilde{\mathbf{X}}_{-j}\boldsymbol{\gamma}_j\|\le M$ for some $M>0$, we can obtain that with probability $1-o(p^{-3})$,

$$\frac{n^{-1/2}\widetilde{\mathbf{R}}|\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j|}{\widetilde{\tau}_j^2}\lesssim n^{-1/2}\|\widetilde{\mathbf{X}}^{\mathrm{aug}}(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0)\|_2^2\lesssim\frac{s\log p}{\sqrt{n}}.\tag{A.321}$$

Further, observe that the main term $n^{-1/2}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)$ in (A.269) is sub-Gaussian since $\|\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j\|_\infty\le M$ and $\|\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}\|_\infty\le M$ for some constant $M>0$. Moreover, it holds that

$$\mathrm{Var}(n^{-1/2}\dot{\boldsymbol{\rho}}_{\boldsymbol{\beta}^0}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)|\widetilde{\mathbf{X}}^{\mathrm{aug}})=n^{-1}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)^T\mathbf{D}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)\le M$$

for some constant $M>0$. Therefore, using similar arguments as in the proof of Lemma 12, we can establish the desired result in Lemma 22.

C.4. Proof of Lemma 23. We will apply the moderate deviation result (i.e., the rate of convergence) for multivariate normal approximation (Saulis (1992)), and the remaining proof can proceed by the same technique as used for proving Lemma 13. From the decomposition for $\sqrt{n}(\widetilde{b}_j-\beta^0_j)$ outlined in (A.269) and the bounds in (A.320) and (A.321), it is seen that the main term is $\xi_j:=-\frac{\dot{\boldsymbol{\rho}}^T_{\boldsymbol{\beta}^0}(\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j)}{\sqrt{n}\widetilde{\tau}_j^2}$ and the two remainders in (A.269) are bounded by $C\frac{s^{3/2}\log p}{\sqrt{n}}$ with probability $1-o(p^{-1})$. Denote by $\widetilde{\mathbf{z}}_j=\widetilde{\mathbf{X}}^{\mathrm{aug}}_j-\widetilde{\mathbf{X}}^{\mathrm{aug}}_{-j}\widetilde{\boldsymbol{\gamma}}_j$ for $1\le j\le 2p$. Observe that given $\widetilde{\mathbf{X}}^{\mathrm{aug}}$, $(\xi_j,\xi_{j+p},\xi_l,\xi_{l+p})^T\overset{d}{\sim}N(\mathbf{0},\mathbf{V})$, where the covariance matrix $\mathbf{V}$ is given by $\mathbf{V}=\begin{pmatrix}\mathbf{V}_{11}\mathbf{V}_{12}\\ \mathbf{V}_{21}\mathbf{V}_{22}\end{pmatrix}$ with

$$\mathbf{V}_{11}=\begin{pmatrix}\frac{\widetilde{\mathbf{z}}_j^T\mathbf{D}\widetilde{\mathbf{z}}_j}{n\widetilde{\tau}_j^4} & \frac{\widetilde{\mathbf{z}}_j^T\mathbf{D}\widetilde{\mathbf{z}}_{j+p}}{n\widetilde{\tau}_j^2\widetilde{\tau}_{j+p}^2}\\ \frac{\widetilde{\mathbf{z}}_j^T\mathbf{D}\widetilde{\mathbf{z}}_{j+p}}{n\widetilde{\tau}_j^2\widetilde{\tau}_{j+p}^2} & \frac{\widetilde{\mathbf{z}}_{j+p}^T\mathbf{D}\widetilde{\mathbf{z}}_{j+p}}{n\widetilde{\tau}_{j+p}^4}\end{pmatrix},\quad \mathbf{V}_{12}=\mathbf{V}_{21}^T=\begin{pmatrix}\frac{\widetilde{\mathbf{z}}_j^T\mathbf{D}\widetilde{\mathbf{z}}_l}{n\widetilde{\tau}_j^2\widetilde{\tau}_l^2} & \frac{\widetilde{\mathbf{z}}_j^T\mathbf{D}\widetilde{\mathbf{z}}_{l+p}}{n\widetilde{\tau}_j^2\widetilde{\tau}_{l+p}^2}\\ \frac{\widetilde{\mathbf{z}}_{j+p}^T\mathbf{D}\widetilde{\mathbf{z}}_l}{n\widetilde{\tau}_{j+p}^2\widetilde{\tau}_l^2} & \frac{\widetilde{\mathbf{z}}_{j+p}^T\mathbf{D}\widetilde{\mathbf{z}}_{l+p}}{n\widetilde{\tau}_{j+p}^2\widetilde{\tau}_{l+p}^2}\end{pmatrix},$$

$$\mathbf{V}_{22}=\begin{pmatrix}\frac{\widetilde{\mathbf{z}}_l^T\mathbf{D}\widetilde{\mathbf{z}}_l}{n\widetilde{\tau}_l^4} & \frac{\widetilde{\mathbf{z}}_l^T\mathbf{D}\widetilde{\mathbf{z}}_{l+p}}{n\widetilde{\tau}_l^2\widetilde{\tau}_{l+p}^2}\\ \frac{\widetilde{\mathbf{z}}_l^T\mathbf{D}\widetilde{\mathbf{z}}_{l+p}}{n\widetilde{\tau}_l^2\widetilde{\tau}_{l+p}^2} & \frac{\widetilde{\mathbf{z}}_{j+p}^T\mathbf{D}\widetilde{\mathbf{z}}_{l+p}}{n\widetilde{\tau}_{l+p}^4}\end{pmatrix}.$$

Let us define the event

$$\mathcal{E}:=\left\{\max_{1\le j,l\le 2p}\left|n^{-1}\widetilde{\mathbf{z}}_j\mathbf{D}\widetilde{\mathbf{z}}_l-\frac{\boldsymbol{\Omega}_{j,l}}{\boldsymbol{\Omega}_{j,j}\boldsymbol{\Omega}_{l,l}}\right|\le C\sqrt{\frac{s\log p}{n}}\right\}.$$

By Condition 18, we see that $\mathbb{P}(\mathcal{E}) \geq 1 - o(p^{-3})$.

Let $(Z_1, Z_2, Z_3, Z_4)^T \overset{d}{\sim} N(\mathbf{0}, \mathbf{V})$. Given $\widetilde{\mathbf{X}}^{\text{aug}}$ and event $\mathcal{E}$, it follows from the rate of convergence (i.e., the moderate deviation theorem) for multivariate normal approximate (e.g. Theorem 1 in Saulis (1992)) that for any $1 \leq j \neq l \leq 2p$,

$$\left| \frac{\mathbb{P}(\xi_{j+p} \geq 0, \xi_j - \xi_{j+p} \geq t, \xi_{l+p} \geq 0, \xi_l - \xi_{l+p} \geq t)}{\mathbb{P}(Z_2 \geq 0, Z_1 - Z_2 \geq t, Z_4 \geq 0, Z_3 - Z_4 \geq t)} - 1 \right| \leq C \frac{1+t^3}{\sqrt{n}} \tag{A.322}$$

uniformly for $t \in [0, C\sqrt{\log p}]$ when $\log p = o(n^{1/3})$. Noting that $\mathbb{P}(|\xi_j| - |\xi_{j+p}| \geq t, |\xi_l - \xi_{l+p}| \geq t)$ can be decomposed into 16 probabilities that are similar to the numerator in (A.322), we can deduce that for any $1 \leq j \neq l \leq 2p$,

$$\left| \frac{\mathbb{P}(|\xi_j| - |\xi_{j+p}| \geq t, |\xi_l - \xi_{l+p}| \geq t)}{\mathbb{P}(|Z_1| - |Z_2| \geq t, |Z_3| - |Z_4| \geq t)} - 1 \right| \leq C \frac{1+t^3}{\sqrt{n}} \tag{A.323}$$

uniformly for $t \in [0, C\sqrt{\log p}]$ when $\log p = o(n^{1/3})$. Analogously, we can show that for any $1 \leq j \neq 2p$,

$$\left| \frac{\mathbb{P}(|\xi_j| - |\xi_{j+p}| \geq t)}{\mathbb{P}(|Z_1| - |Z_2| \geq t)} - 1 \right| \leq C \frac{1+t^3}{\sqrt{n}} \tag{A.324}$$

uniformly for $t \in [0, C\sqrt{\log p}]$ when $\log p = o(n^{1/3})$. Therefore, following exactly the same procedure for proving Lemma 13, we can establish Lemma 23. To avoid redundancy, we omit the proof details here.

C.5. Proof of Lemma 24. By the moderate deviation result in (A.324), the probability $\mathbb{P}(\widetilde{W}_j \geq t) \approx \mathbb{P}(|\xi_j| - |\xi_{j+p}| \geq t)$ can be approximated by the probability $\mathbb{P}(|Z_1| - |Z_2| \geq t)$ of normal distribution with controlled relative rate of convergence as $t \leq C\sqrt{\log p}$. By the same technique for proving Lemma 14, but just with slightly different definitions that $\delta_n = \frac{s^{3/2} \log p}{\sqrt{n}}$ and $b_n = C(\Delta_n s \sqrt{\frac{\log p}{n}} + \frac{s^{3/2} \log p}{n})$, we can establish the desired results in Lemma 24. To avoid redundancy, we omit the proof details here.