Uniform Post Selection Inference for Least Absolute Deviation Regression and Other Z-estimation Problems

BY A. BELLONI

Fuqua School of Business, Duke University,
100 Fuqua Drive, Durham, North Carolina 27708, U.S.
abn5@duke.edu

AND V. CHERNOZHUKOV

Department of Economics, Massachusetts Institute of Technology,
50 Memorial Drive, Cambridge, Massachusetts 02142, U.S.
vchern@mit.edu

AND K. KATO

Graduate School of Economics, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0013, Japan
kkato@e.u-tokyo.ac.jp

SUMMARY

We develop uniformly valid confidence regions for regression coefficients in a high-dimensional sparse median regression model with homoscedastic errors. Our methods are based on a moment equation that is immunized against non-regular estimation of the nuisance part of the median regression function by using Neyman’s orthogonalization. We establish that the resulting instrumental median regression estimator of a target regression coefficient is asymptotically normally distributed uniformly with respect to the underlying sparse model and is semiparametrically efficient. We also generalize our method to a general non-smooth Z-estimation framework with the number of target parameters being possibly much larger than the sample size. We extend Huber’s results on asymptotic normality to this setting, demonstrating uniform asymptotic normality of the proposed estimators over rectangles, constructing simultaneous confidence bands on all of the target parameters, and establishing asymptotic validity of the bands uniformly over underlying approximately sparse models.

Some key words: Instrument; Post-selection inference; Sparsity; Neyman’s Orthogonal Score test; Uniformly valid inference; Z-estimation.

1. INTRODUCTION

We consider independent and identically distributed data vectors \((y_i, x_i^T, d_i)^T\) that obey the regression model

\[
y_i = d_i \alpha_0 + x_i^T \beta_0 + \epsilon_i \quad (i = 1, \ldots, n),
\]

where \(d_i\) is the main regressor and coefficient \(\alpha_0\) is the main parameter of interest. The vector \(x_i\) denotes other high-dimensional regressors or controls. The regression error \(\epsilon_i\) is independent
of \(d_i\) and \(x_i\) and has median zero, that is, \(\text{pr}(\epsilon_i \leq 0) = 1/2\). The distribution function of \(\epsilon_i\) is denoted by \(F_i\) and admits a density function \(f_i\) such that \(f_i(0) > 0\). The assumption motivates the use of the least absolute deviation or median regression, suitably adjusted for use in high-dimensional settings. The framework (1) is of interest in program evaluation, where \(d_i\) represents the treatment or policy variable known a priori and whose impact we would like to infer (Robinson, 1988; Liang et al., 2004; Imbens, 2004). We shall also discuss a generalization to the case where there are many parameters of interest, including the case where the identity of a regressor of interest is unknown a priori.

The dimension \(p\) of controls \(x_i\) may be much larger than \(n\), which creates a challenge for inference on \(\alpha_0\). Although the unknown nuisance parameter \(\beta_0\) lies in this large space, the key assumption that will make estimation possible is its sparsity, namely \(T = \text{supp}(\beta_0)\) has \(s < n\) elements, where the notation \(\text{supp}(\delta) = \{ j \in \{1, \ldots, p\} : \delta_j \neq 0 \}\) denotes the support of a vector \(\delta \in \mathbb{R}^p\). Here \(s\) can depend on \(n\), as we shall use array asymptotics. Sparsity motivates the use of regularization or model selection methods.

A non-robust approach to inference in this setting would be first to perform model selection via the \(\ell_1\)-penalized median regression estimator

\[
(\hat{\alpha}, \hat{\beta}) \in \arg \min_{\alpha, \beta} E_n(||y_i - d_i\alpha - x_i^T\beta||) + \frac{\lambda}{n} \|\Psi(\alpha, \beta^T)^T\|_1,
\]

where \(\lambda\) is a penalty parameter and \(\Psi^2 = \text{diag}\{E_n(d_i^2), E_n(x_i^2), \ldots, E_n(x_{ip}^2)\}\) is a diagonal matrix with normalization weights, where the notation \(E_n(\cdot)\) denotes the average \(n^{-1} \sum_{i=1}^n\) over the index \(i = 1, \ldots, n\). Then one would use the post-model selection estimator

\[
(\bar{\alpha}, \bar{\beta}) \in \min_{\alpha, \beta} \left\{ E_n(||y_i - d_i\alpha - x_i^T\beta||) : \beta_j = 0, j \notin \text{supp}(\beta) \right\},
\]

to perform inference for \(\alpha_0\).

This approach is justified if (2) achieves perfect model selection with probability approaching unity, so that the estimator (3) has the oracle property. However conditions for perfect selection are very restrictive in this model, and, in particular, require strong separation of non-zero coefficients away from zero. If these conditions do not hold, the estimator \(\bar{\alpha}\) does not converge to \(\alpha_0\) at the \(n^{-1/2}\) rate, uniformly with respect to the underlying model, and so the usual inference breaks down (Leeb & Pötscher, 2005). We shall demonstrate the breakdown of such naive inference in Monte Carlo experiments where non-zero coefficients in \(\beta_0\) are not significantly separated from zero.

The breakdown of standard inference does not mean that the aforementioned procedures are not suitable for prediction. Indeed, the estimators (2) and (3) attain essentially optimal rates \(\left\{ (s \log p) / n \right\}^{1/2}\) of convergence for estimating the entire median regression function (Belloni & Chernozhukov, 2011; Wang, 2013). This property means that while these procedures will not deliver perfect model recovery, they will only make moderate selection mistakes, that is, they omit controls only if coefficients are local to zero.

In order to provide uniformly valid inference, we propose a method whose performance does not require perfect model selection, allowing potential moderate model selection mistakes. The latter feature is critical in achieving uniformity over a large class of data generating processes, similarly to the results for instrumental regression and mean regression studied in Zhang & Zhang (2014) and Belloni et al. (2012; 2013; 2014a). This allows us to overcome the impact of moderate model selection mistakes on inference, avoiding in part the criticisms in Leeb & Pötscher (2005), who prove that the oracle property achieved by the naive estimators implies the failure of uniform validity of inference and their semiparametric inefficiency (Leeb & Pötscher, 2008).
Uniform Post Selection Inference for Z-problems

In order to achieve robustness with respect to moderate selection mistakes, we shall construct an orthogonal moment equation that identifies the target parameter. The following auxiliary equation,

\[ d_i = x_i^T \theta_0 + v_i, \quad E(v_i \mid x_i) = 0 \quad (i = 1, \ldots, n), \]

(4)

which describes the dependence of the regressor of interest \( d_i \) on the other controls \( x_i \), plays a key role. We shall assume the sparsity of \( \theta_0 \), that is, \( T_\theta = \text{supp}(\theta_0) \) has at most \( s < n \) elements, and estimate the relation (4) via lasso or post-lasso least squares methods described below. We shall use \( v_i \) as an instrument in the following moment equation for \( \alpha_0 \):

\[ E\{\varphi(y_i - d_i \alpha_0 - x_i^T \beta_0)v_i\} = 0 \quad (i = 1, \ldots, n), \]

(5)

where \( \varphi(t) = 1/2 - 1\{t \leq 0\} \). We shall use the empirical analog of (5) to form an instrumental median regression estimator of \( \alpha_0 \), using a plug-in estimator for \( x_i^T \beta_0 \). The moment equation (5) has the orthogonality property

\[ \frac{\partial}{\partial \beta} E\{\varphi(y_i - d_i \alpha_0 - x_i^T \beta)v_i\} \bigg|_{\beta=\beta_0} = 0 \quad (i = 1, \ldots, n), \]

(6)

so the estimator of \( \alpha_0 \) will be unaffected by estimation of \( x_i^T \beta_0 \) even if \( \beta_0 \) is estimated at a slower rate than \( n^{-1/2} \), that is, the rate of \( o(n^{-1/4}) \) would suffice. This slow rate of estimation of the nuisance function permits the use of non-regular estimators of \( \beta_0 \), such as post-selection or regularized estimators that are not \( n^{-1/2} \) consistent uniformly over the underlying model. The orthogonalization ideas can be traced back to Neyman (1959) and also play an important role in doubly robust estimation (Robins & Rotnitzky, 1995).

Our estimation procedure has three steps: (i) estimation of the confounding function \( x_i^T \beta_0 \) in (1); (ii) estimation of the instruments \( v_i \) in (4); and (iii) estimation of the target parameter \( \alpha_0 \) via empirical analog of (5). Each step is computationally tractable, involving solutions of convex problems and a one-dimensional search.

Step (i) estimates for the nuisance function \( x_i^T \beta_0 \) via either the \( \ell_1 \)-penalized median regression estimator (2) or the associated post-model selection estimator (3).

Step (ii) provides estimates \( \hat{v}_i \) of \( v_i \) in (4) as \( \hat{v}_i = d_i - x_i^T \hat{\theta} \) or \( \hat{v}_i = d_i - x_i^T \hat{\theta} \) (i = 1, \ldots, n). The first is based on the heteroscedastic lasso estimator \( \hat{\theta} \), a version of the lasso of Tibshirani (1996), designed to address non-Gaussian and heteroscedastic errors (Belloni et al., 2012),

\[ \hat{\theta} \in \arg \min_\theta E_n\{(d_i - x_i^T \theta)^2\} + \frac{\lambda}{n} \|\hat{\Gamma} \theta\|_1, \]

(7)

where \( \lambda \) and \( \hat{\Gamma} \) are the penalty level and data-driven penalty loadings defined in the Supplementary Material. The second is based on the associated post-model selection estimator and \( \theta \), called the post-lasso estimator:

\[ \tilde{\theta} \in \arg \min_\theta \left[ E_n\{(d_i - x_i^T \theta)^2\} : \theta_j = 0, \ j \notin \text{supp}(\hat{\theta}) \right]. \]

(8)

Step (iii) constructs an estimator \( \hat{\alpha} \) of the coefficient \( \alpha_0 \) via an instrumental median regression (Chernozhukov & Hansen, 2008), using \( (\hat{v}_i)^n \) as instruments, defined by

\[ \hat{\alpha} \in \arg \min_{\alpha \in \hat{A}} L_n(\alpha), \quad L_n(\alpha) = \frac{4E_n\{|\varphi(y_i - x_i^T \beta_0 - d_i \alpha)\hat{v}_i|\}^2}{E_n(|\hat{v}_i|)^2}, \]

(9)

\[ \alpha = \frac{1}{n} \sum \varphi\left(\frac{d_i - x_i^T \alpha}{\hat{\theta}}\right) \quad (i = 1, \ldots, n), \]

where \( \varphi(t) = 1/2 - 1\{t \leq 0\} \). We shall use the empirical analog of (5) to form an instrumental median regression estimator of \( \alpha_0 \), using a plug-in estimator for \( x_i^T \beta_0 \). The moment equation (5) has the orthogonality property

\[ \frac{\partial}{\partial \beta} E\{\varphi(y_i - d_i \alpha_0 - x_i^T \beta)v_i\} \bigg|_{\beta=\beta_0} = 0 \quad (i = 1, \ldots, n), \]

(6)

so the estimator of \( \alpha_0 \) will be unaffected by estimation of \( x_i^T \beta_0 \) even if \( \beta_0 \) is estimated at a slower rate than \( n^{-1/2} \), that is, the rate of \( o(n^{-1/4}) \) would suffice. This slow rate of estimation of the nuisance function permits the use of non-regular estimators of \( \beta_0 \), such as post-selection or regularized estimators that are not \( n^{-1/2} \) consistent uniformly over the underlying model. The orthogonalization ideas can be traced back to Neyman (1959) and also play an important role in doubly robust estimation (Robins & Rotnitzky, 1995).

Our estimation procedure has three steps: (i) estimation of the conounding function \( x_i^T \beta_0 \) in (1); (ii) estimation of the instruments \( v_i \) in (4); and (iii) estimation of the target parameter \( \alpha_0 \) via empirical analog of (5). Each step is computationally tractable, involving solutions of convex problems and a one-dimensional search.

Step (i) estimates for the nuisance function \( x_i^T \beta_0 \) via either the \( \ell_1 \)-penalized median regression estimator (2) or the associated post-model selection estimator (3).

Step (ii) provides estimates \( \hat{v}_i \) of \( v_i \) in (4) as \( \hat{v}_i = d_i - x_i^T \hat{\theta} \) or \( \hat{v}_i = d_i - x_i^T \hat{\theta} \) (i = 1, \ldots, n). The first is based on the heteroscedastic lasso estimator \( \hat{\theta} \), a version of the lasso of Tibshirani (1996), designed to address non-Gaussian and heteroscedastic errors (Belloni et al., 2012),

\[ \hat{\theta} \in \arg \min_\theta E_n\{(d_i - x_i^T \theta)^2\} + \frac{\lambda}{n} \|\hat{\Gamma} \theta\|_1, \]

(7)

where \( \lambda \) and \( \hat{\Gamma} \) are the penalty level and data-driven penalty loadings defined in the Supplementary Material. The second is based on the associated post-model selection estimator and \( \theta \), called the post-lasso estimator:

\[ \tilde{\theta} \in \arg \min_\theta \left[ E_n\{(d_i - x_i^T \theta)^2\} : \theta_j = 0, \ j \notin \text{supp}(\hat{\theta}) \right]. \]

(8)

Step (iii) constructs an estimator \( \hat{\alpha} \) of the coefficient \( \alpha_0 \) via an instrumental median regression (Chernozhukov & Hansen, 2008), using \( (\hat{v}_i)^n \) as instruments, defined by

\[ \hat{\alpha} \in \arg \min_{\alpha \in \hat{A}} L_n(\alpha), \quad L_n(\alpha) = \frac{4E_n\{|\varphi(y_i - x_i^T \beta_0 - d_i \alpha)\hat{v}_i|\}^2}{E_n(|\hat{v}_i|)^2}, \]

(9)
where $\hat{A}$ is a possibly stochastic parameter space for $\alpha_0$. We suggest $\hat{A} = [\tilde{\alpha} - 10/b, \tilde{\alpha} + 10/b]$ with $b = \{E_n(d_i^2)\}^{1/2} \log n$, though we allow for other choices.

Our main result establishes that under homoscedasticity, provided that $(s^3 \log^3 p)/n \to 0$ and other regularity conditions hold, despite possible model selection mistakes in Steps (i) and (ii), the estimator $\hat{\alpha}$ obeys

$$\sigma_n^{-1} n^{1/2}(\hat{\alpha} - \alpha_0) \to N(0,1)$$

in distribution, where $\sigma_n^2 = 1/\{4 f_i^2 E(v_i^2)\}$ with $f_i = f_i(0)$ is the semi-parametric efficiency bound for regular estimators of $\alpha_0$. In the low-dimensional case, if $p^3 = o(n)$, the asymptotic behavior of our estimator coincides with that of the standard median regression without selection or penalization, as derived in He & Shao (2000), which is also semi-parametrically efficient in this case. However, the behaviors of our estimator and the standard median regression differ dramatically, otherwise, with the standard estimator even failing to be consistent when $p > n$. Of course, this improvement in the performance comes at the cost of assuming sparsity.

An alternative, more robust expression for $\sigma_n^2$ is given by

$$\sigma_n^2 = J^{-1} \Omega J^{-1}, \quad \Omega = E(v_i^2)/4, \quad J = E(f_i d_i v_i).$$

We estimate $\Omega$ by the plug-in method and $J$ by Powell’s (1986) method. Furthermore, we show that the score statistic $n L_n(\alpha)$ can be used for testing the null hypothesis $\alpha = \alpha_0$, and converges in distribution to a $\chi^2_1$ variable under the null hypothesis, that is,

$$n L_n(\alpha_0) \to \chi^2_1$$

(12)

in distribution. This allows us to construct a confidence region with asymptotic coverage $1 - \xi$ based on inverting the score statistic $n L_n(\alpha)$:

$$\hat{A}_\xi = \{\alpha \in \hat{A} : n L_n(\alpha) \leq q_{1-\xi}\}, \quad \Pr(\alpha_0 \in \hat{A}_\xi) \to 1 - \xi,$$

(13)

where $q_{1-\xi}$ is the $(1 - \xi)$-quantile of the $\chi^2_1$-distribution.

The robustness with respect to moderate model selection mistakes, which is due to (6), allows (10) and (12) to hold uniformly over a large class of data generating processes. Throughout the paper, we use array asymptotics, asymptotics where the model changes with $n$, to better capture finite-sample phenomena such as small coefficients that are local to zero. This ensures the robustness of conclusions with respect to perturbations of the data-generating process along various model sequences. This robustness, in turn, translates into uniform validity of confidence regions over many data-generating processes.

The second set of main results addresses a more general setting by allowing $p_1$-dimensional target parameters defined via Huber’s Z-problems to be of interest, with dimension $p_1$ potentially much larger than the sample size $n$, and also allowing for approximately sparse models instead of exactly sparse models. This framework covers a wide variety of semi-parametric models, including those with smooth and non-smooth score functions. We provide sufficient conditions to derive a uniform Bahadur representation, and establish uniform asymptotic normality, using central limit theorems and bootstrap results of Chernozhukov et al. (2013), for the entire $p_1$-dimensional vector. The latter result holds uniformly over high-dimensional rectangles of dimension $p_1 \gg n$ and over an underlying approximately sparse model, thereby extending previous results from the setting with $p_1 \ll n$ (Huber, 1973; Portnoy, 1984, 1985; He & Shao, 2000) to that with $p_1 \gg n$.

In what follows, the $\ell_2$ and $\ell_1$ norms are denoted by $\| \cdot \|$ and $\| \cdot \|_1$, respectively, and the $\ell_0$-norm, $\| \cdot \|_0$, denotes the number of non-zero components of a vector. We use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Denote by $\Phi(\cdot)$ the distribution function of the standard
normal distribution. We assume that the quantities such as $p, s$, and hence $y_i, x_i, \beta_0, \theta_0, T$ and $T_d$ are all dependent on the sample size $n$, and allow for the case where $p = p_n \to \infty$ and $s = s_n \to \infty$ as $n \to \infty$. We shall omit the dependence of these quantities on $n$ when it does not cause confusion. For a class of measurable functions $F$ on a measurable space, let $cn(\varepsilon, F, \| \cdot \|_{Q,2})$ denote its $\varepsilon$-covering number with respect to the $L^2(Q)$ seminorm $\| \cdot \|_{Q,2}$, where $Q$ is a finitely discrete measure on the space, and let $\text{ent}(\varepsilon, F) = \log \sup_Q cn(\varepsilon\|F\|_{Q,2}, F, \| \cdot \|_{Q,2})$ denote the uniform entropy number where $F = \sup_{f \in F} |f|$.

2. The Methods, Conditions, and Results

2.1. The methods

Each of the steps outlined in Section 1 could be implemented by several estimators. Two possible implementations are the following.

Algorithm 1. The algorithm is based on post-model selection estimators.

Step (i). Run post-$\ell_1$-penalized median regression (3) of $y_i$ on $d_i$ and $x_i$; keep fitted value $x_i^T \hat{\beta}$.

Step (ii). Run the post-lasso estimator (8) of $d_i$ on $x_i$; keep the residual $\hat{v}_i = d_i - x_i^T \hat{\beta}$.

Step (iii). Run instrumental median regression (9) of $y_i - x_i^T \hat{\beta}$ on $d_i$ using $\hat{v}_i$ as the instrument. Report $\hat{\alpha}$ and perform inference based upon (10) or (13).

Algorithm 2. The algorithm is based on regularized estimators.

Step (i). Run $\ell_1$-penalized median regression (3) of $y_i$ on $d_i$ and $x_i$; keep fitted value $x_i^T \hat{\beta}$.

Step (ii). Run the lasso estimator (7) of $d_i$ on $x_i$; keep the residual $\hat{v}_i = d_i - x_i^T \hat{\beta}$.

Step (iii). Run instrumental median regression (9) of $y_i - x_i^T \hat{\beta}$ on $d_i$ using $\hat{v}_i$ as the instrument. Report $\hat{\alpha}$ and perform inference based upon (10) or (13).

In order to perform $\ell_1$-penalized median regression and lasso, one has to choose the penalty levels suitably. We record our penalty choices in the Supplementary Material. Algorithm 1 relies on the post-selection estimators that refit the non-zero coefficients without the penalty term to reduce the bias, while Algorithm 2 relies on the penalized estimators. In Step (ii), instead of the lasso or the post-lasso estimators, Dantzig selector (Candes & Tao, 2007) and Gauss-Dantzig estimators could be used. Step (iii) of both algorithms relies on instrumental median regression (9). Alternatively, in this step, we can use a one-step estimator $\tilde{\alpha}$ defined by

$$\tilde{\alpha} = \hat{\alpha} + E_n\{f_t(0)\hat{\beta}_t^2\}^{-1}E_n\{\varphi(y_i - d_i\hat{\alpha} - x_i^T \hat{\beta})\hat{v}_i\},$$

where $\hat{\alpha}$ is the $\ell_1$-penalized median regression estimator (2). Another possibility is to use the post-double selection median regression estimation, which is simply the median regression of $y_i$ on $d_i$ and the union of controls selected in both Steps (i) and (ii), as $\tilde{\alpha}$. The Supplemental Material shows that these alternative estimators also solve (9) approximately.

2.2. Regularity conditions

We state regularity conditions sufficient for validity of the main estimation and inference results. The behavior of sparse eigenvalues of the population Gram matrix $E(\bar{x}_i \bar{x}_i^T)$ with $\bar{x}_i = (d_i, x_i^T)^T$ plays an important role in the analysis of $\ell_1$-penalized median regression and lasso. Define the minimal and maximal $m$-sparse eigenvalues of the population Gram matrix as

$$\hat{\phi}_{\text{min}}(m) = \min_{1 \leq \| \delta \|_0 \leq m} \frac{\delta^T E(\bar{x}_i \bar{x}_i^T) \delta}{\| \delta \|^2}, \quad \hat{\phi}_{\text{max}}(m) = \max_{1 \leq \| \delta \|_0 \leq m} \frac{\delta^T E(\bar{x}_i \bar{x}_i^T) \delta}{\| \delta \|^2},$$

(15)
where \( m = 1, \ldots, p \). Assuming that \( \tilde{\sigma}_{\min}(m) > 0 \) requires that all population Gram submatrices formed by any \( m \) components of \( \bar{x}_i \) are positive definite.

The main condition, Condition 1, imposes sparsity of the vectors \( \beta_0 \) and \( \theta_0 \) as well as other more technical assumptions. Below let \( c_1 \) and \( C_1 \) be given positive constants, and let \( \ell_n \uparrow \infty \), \( \delta_n \downarrow 0 \), and \( \Delta_n \downarrow 0 \) be given sequences of positive constants.

**Condition 1.** Suppose that (i) \( \{(y_i, d_i, x_i^T)\}_{i=1}^n \) is a sequence of independent and identically distributed random vectors generated according to models (1) and (4), where \( \epsilon_i \) has distribution with distribution function \( F \) such that \( F_\epsilon(0) = 1/2 \) and is independent of the random vector \( (d_i, x_i^T) \); (ii) \( E(v_i^2 \mid x_i) \geq c_1 \) and \( E(|v_i|^3 \mid x_i) \leq C_1 \) almost surely; moreover, \( E(d_i^4) + E(v_i^4) + \max_{j=1,\ldots,p} E(x_{ij}^2d_i^2) + E(|x_{ij}v_i|^3) \leq C_1 \); (iii) there exists \( s = s_n \geq 1 \) such that \( \|\beta_0\|_0 \leq s \) and \( \|\theta_0\|_0 \leq s \); (iv) the error distribution \( F_\epsilon \) is absolutely continuous with continuously differentiable density \( f_\epsilon(\cdot) \) such that \( f_\epsilon(0) \geq c_1 \) and \( f_\epsilon(t) \vee |f'_\epsilon(t)| \leq C_1 \) for all \( t \in \mathbb{R} \); (v) there exist constants \( K_n \) and \( M_n \) such that \( K_n \geq \max_{j=1,\ldots,p} |x_{ij}| \) and \( M_n \geq 1 \vee |x_{ij}\theta_0| \) almost surely, and they obey the growth condition \( \{K_n^4 + (K_n^2 \vee M_n^2)s^2 + M_n^2s^3\} \log^3(p \vee n) \leq n\delta_n \); (vi) \( c_1 \leq \phi_{\min}(\ell_n, s) \leq \phi_{\max}(\ell_n, s) \leq C_1 \).

Condition 1 (i) imposes the setting discussed in the previous section with the zero conditional error distribution. Condition 1 (ii) imposes moment conditions on the structural errors and regressors to ensure good model selection performance of lasso applied to equation (4). Condition 1 (iii) imposes sparsity of the high-dimensional vectors \( \beta_0 \) and \( \theta_0 \). Condition 1 (iv) is a set of standard assumptions in median regression (Koenker, 2005) and in instrumental quantile regression. Condition 1 (v) restricts the sparsity index, namely \( s^3 \log^3(p \vee n) = o(n) \) is required; this is analogous to the restriction \( p^3 \log^3(p \vee n) = o(n) \) made in He & Shao (2000) in the low-dimensional setting. The uniformly bounded regressors condition can be relaxed with minor modifications provided the bound holds with probability approaching unity. Most importantly, no assumptions on the separation from zero of the non-zero coefficients of \( \theta_0 \) and \( \beta_0 \) are made. Condition 1 (vi) is quite plausible for many designs of interest. Conditions 1 (iv) and (v) imply the equivalence between the norms induced by the empirical and population Gram matrices over \( s \)-sparse vectors by Rudelson & Zhou (2013).

### 2.3. Results

The following result is derived as an application of a more general Theorem 2 given in Section 3; the proof is given in the Supplementary Material.

**Theorem 1.** Let \( \hat{\alpha} \) and \( L_n(\alpha_0) \) be the estimator and statistic obtained by applying either Algorithm 1 or 2. Suppose that Condition 1 is satisfied for all \( n \geq 1 \). Moreover, suppose that with probability at least \( 1 - \Delta_n \), \( \|\hat{\beta}\|_0 \leq C_1s \). Then, as \( n \to \infty \), \( \sigma_n^{-1}n^{1/2}(\hat{\alpha} - \alpha_0) \to N(0,1) \) and \( nL_n(\alpha_0) \to \chi^2_\delta \) in distribution, where \( \sigma_n^2 = 1/\{4f^2\epsilon E(v^2_\epsilon)\} \).

Theorem 1 shows that Algorithms 1 and 2 produce estimators \( \hat{\alpha} \) that perform equally well, to the first order, with asymptotic variance equal to the semi-parametric efficiency bound; see the Supplementary Material for further discussion. Both algorithms rely on sparsity of \( \hat{\beta} \) and \( \hat{\theta} \). Sparsity of the latter follows immediately under sharp penalty choices for optimal rates. The sparsity for the former potentially requires a higher penalty level, as shown in Belloni & Chernozhukov (2011); alternatively, sparsity for the estimator in Step 1 can also be achieved by truncating the smallest components of \( \hat{\beta} \). The Supplementary Material shows that suitable truncation leads to the required sparsity while preserving the rate of convergence.
An important consequence of these results is the following corollary. Here \( \mathcal{P}_n \) denotes a collection of distributions for \( \{(y_i, d_i, x_i^T)^T\}_{i=1}^n \) and for \( P_n \in \mathcal{P}_n \) the notation \( \text{pr}_{P_n} \) means that under \( \text{pr}_{P_n} \), \( \{(y_i, d_i, x_i^T)^T\}_{i=1}^n \) is distributed according to the law determined by \( P_n \).

**Corollary.** Let \( \hat{\alpha} \) be the estimator of \( \alpha_0 \) constructed according to either Algorithm 1 or 2, and for every \( n \geq 1 \), let \( \mathcal{P}_n \) be the collection of all distributions of \( \{(y_i, d_i, x_i^T)^T\}_{i=1}^n \) for which Condition 1 holds and \( \|\hat{\beta}\|_0 \leq C_1 s \) with probability at least \( 1 - \Delta_n \). Then for \( \hat{A}_\xi \) defined in (13),

\[
\sup_{P_n \in \mathcal{P}_n} \left\{ \text{pr}_{P_n} \left\{ \alpha_0 \in [\hat{\alpha} \pm \sigma_n n^{-1/2} \Phi^{-1}(1 - \xi / 2)] - (1 - \xi) \right\} \to 0, \right.
\]

\[
\sup_{P_n \in \mathcal{P}_n} \left\{ \text{pr}_{P_n} \left( \alpha_0 \in \hat{A}_\xi \right) - (1 - \xi) \right\} \to 0, \quad n \to \infty.
\]

Corollary 1 establishes the second main result of the paper. It highlights the uniform validity of the results, which hold despite the possible imperfect model selection in Steps (i) and (ii). Condition 1 explicitly characterizes regions of data-generating processes for which the uniformity result holds. Simulations presented below provide additional evidence that these regions are substantial. Here we rely on exactly sparse models, but these results extend to approximately sparse model in what follows.

Both of the proposed algorithms exploit the homoscedasticity of the model (1) with respect to the error term \( \epsilon \). The generalization to the heteroscedastic case can be achieved but we need to consider the density-weighted version of the auxiliary equation (4) in order to achieve the semiparametric efficiency bound. The analysis of the impact of estimation of weights is delicate and is developed in our working paper “Robust Inference in High-Dimensional Approximate Sparse Quantile Regression Models” (arXiv:1312.7186).

### 2.4. Generalization to many target coefficients

We consider the generalization to the previous model:

\[
y = \sum_{j=1}^{p_1} d_j \alpha_j + g(u) + \epsilon, \quad \epsilon \sim F_\epsilon, \quad F_\epsilon(0) = 1/2,
\]

where \( d, u \) are regressors, and \( \epsilon \) is the noise with distribution function \( F_\epsilon \) that is independent of regressors and has median zero, that is, \( F_\epsilon(0) = 1/2 \). The coefficients \( \alpha_1, \ldots, \alpha_{p_1} \) are now the high-dimensional parameter of interest.

We can rewrite this model as \( p_1 \) models of the previous form:

\[
y = \alpha_j d_j + g_j(z_j) + \epsilon, \quad d_j = m_j(z_j) + v_j, \quad E(v_j | z_j) = 0 \quad (j = 1, \ldots, p_1),
\]

where \( \alpha_j \) is the target coefficient,

\[
g_j(z_j) = \sum_{k \neq j}^{p_1} d_k \alpha_k + g(u), \quad m_j(z_j) = E(d_j | z_j),
\]

and where \( z_j = (d_1, \ldots, d_{j-1}, d_{j+1}, \ldots, d_{p_1}, u^T)^T \). We would like to estimate and perform inference on each of the \( p_1 \) coefficients \( \alpha_1, \ldots, \alpha_{p_1} \) simultaneously.

Moreover, we would like to allow regression functions \( h_j = (g_j, m_j)^T \) to be of infinite dimension, that is, they could be written only as infinite linear combinations of some dictionary with respect to \( z_j \). However, we assume that there are sparse estimators \( \hat{h}_j = (\hat{g}_j, \hat{m}_j)^T \) that can estimate \( h_j = (g_j, m_j)^T \) at sufficiently fast \( o(n^{-1/4}) \) rates in the mean square error sense, as stated precisely in Section 3. Examples of functions \( h_j \) that permit such estimation by sparse methods
include the standard Sobolev spaces as well as more general rearranged Sobolev spaces (Bickel et al., 2009; Belloni et al., 2014b). Here sparsity of estimators \( \hat{g}_j \) and \( \hat{m}_j \) means that they are formed by \( O_p(s) \)-sparse linear combinations chosen from \( p \) technical regressors generated from \( z_j \), with coefficients estimated from the data. This framework is general; in particular it contains as a special case the traditional linear sieve/series framework for estimation of \( h_j \), which uses a small number \( s = o(n) \) of predetermined series functions as a dictionary.

Given suitable estimators for \( h_j = (g_j, m_j)^T \), we can then identify and estimate each of the target parameters \( (\alpha_j)_{j=1}^{p_1} \) via the empirical version of the moment equations

\[
E[\psi_j \{ w, \alpha, h_j(z_j) \}] = 0 \quad (j = 1, \ldots, p_1),
\]

where \( \psi_j(w, \alpha, t) = \varphi(y - d_j \alpha - t_1)(d_j - t_2) \) and \( w = (y, d_1, \ldots, d_{p_1}, w^T)^T \). These equations have the orthogonality property:

\[
[\partial E \{ \psi_j \{ w, \alpha, t \} \} / \partial t]_{t = h_j(z_j)} = 0 \quad (j = 1, \ldots, p_1).
\]

The resulting estimation problem is subsumed as a special case in the next section.

3. INFEERENCE ON MANY TARGET PARAMETERS IN Z-PROBLEMS

In this section we generalize the previous example to a more general setting, where \( p_1 \) target parameters defined via Huber’s Z-problems are of interest, with dimension \( p_1 \) potentially much larger than the sample size. This framework covers median regression, its generalization discussed above, and many other semi-parametric models.

The interest lies in \( p_1 = p_1 n \) real-valued target parameters \( \alpha_1, \ldots, \alpha_{p_1} \). We assume that each \( \alpha_j \in A_j \), where each \( A_j \) is a non-stochastic bounded closed interval. The true parameter \( \alpha_j \) is identified as a unique solution of the moment condition:

\[
E[\psi_j \{ w, \alpha, h_j(z_j) \}] = 0. \tag{16}
\]

Here \( w \) is a random vector taking values in \( \mathcal{W} \), a Borel subset of a Euclidean space, which contains vectors \( z_j \ (j = 1, \ldots, p_1) \) as subvectors, and each \( z_j \) takes values in \( Z_j \); here \( z_j \) and \( z_j \) with \( j \neq j' \) may overlap. The vector-valued function \( z \mapsto h_j(z) = \{h_{jm}(z)\}_{m=1}^M \) is a measurable map from \( Z_j \) to \( \mathbb{R}^M \), where \( M \) is fixed, and the function \( (w, \alpha, t) \mapsto \psi_j(w, \alpha, t) \) is a measurable map from an open neighborhood of \( \mathcal{W} \times A_j \times \mathbb{R}^M \) to \( \mathbb{R} \). The former map is a possibly infinite-dimensional nuisance parameter.

Suppose that the nuisance function \( h_j = (h_{jm})_{m=1}^M \) admits a sparse estimator \( \hat{h}_j = (\hat{h}_{jm})_{m=1}^M \) of the form

\[
\hat{h}_{jm}(\cdot) = \sum_{k=1}^p f_{jmk}(\cdot)\hat{\theta}_{jmk}, \quad \| (\hat{\theta}_{jmk})_{k=1}^p \|_0 \leq s \quad (m = 1, \ldots, M),
\]

where \( p = p_n \) may be much larger than \( n \) while \( s = s_n \), the sparsity level of \( \hat{h}_j \), is small compared to \( n \), and \( f_{jmk} : Z_j \to \mathbb{R} \) are given approximating functions.

The estimator \( \hat{\alpha}_j \) of \( \alpha_j \) is then constructed as a Z-estimator, which solves the sample analogue of the equation (16):

\[
|E_n[\psi_j \{ w, \hat{\alpha}_j, h_j(z_j) \}]| \leq \inf_{\alpha \in A_j} |E_n[\psi \{ w, \alpha, \hat{h}_j(z_j) \}]| + \epsilon_n, \tag{17}
\]
where \( \epsilon_n = o(n^{-1/2}b_n^{-1}) \) is the numerical tolerance parameter and \( b_n = \{ \log(ep_1) \}_{1/2} \); \( \hat{\mathcal{A}}_j \) is a possibly stochastic interval contained in \( \mathcal{A}_j \) with high probability. Typically, \( \mathcal{A}_j = \hat{\mathcal{A}}_j \) or can be constructed by using a preliminary estimator of \( \alpha_j \).

In order to achieve robust inference results, we shall need to rely on the condition of orthogonality, or immunity, of the scores with respect to small perturbations in the value of the nuisance parameters, which we can express in the following condition:

\[
\partial_t E\{\psi_j(w, \alpha_j, t) \mid z_j\} \big|_{t = h_j(z_j)} = 0, \tag{18}
\]

where we use the symbol \( \partial_t \) to abbreviate \( \partial / \partial t \). It is important to construct the scores \( \psi_j \) to have property (18) or its generalization given in Remark 1 below. Generally, we can construct the scores \( \psi_j \) that obey such properties by projecting some initial non-orthogonal scores onto the orthogonal complement of the tangent space for the nuisance parameter (van der Vaart & Wellner, 1996; van der Vaart, 1998; Kosorok, 2008). Sometimes the resulting construction generates additional nuisance parameters, for example, the auxiliary regression function in the case of the median regression problem in Section 2.

In Conditions 2 and 3 below, \( \varsigma, n_0, c_1, \) and \( C_1 \) are given positive constants; \( M \) is a fixed positive integer; \( \delta_n \downarrow 0 \) and \( \rho_n \downarrow 0 \) are given sequences of constants. Let \( a_n = \max(p_1, p, n, c) \) and \( b_n = \{ \log(ep_1) \}_{1/2} \).

**Condition 2.** For every \( n \geq 1 \), we observe independent and identically distributed copies \( (w_i)_{i=1}^n \) of the random vector \( w \), whose law is determined by the probability measure \( P \in \mathcal{P}_n \). Uniformly in \( n \geq n_0, P \in \mathcal{P}_n \), and \( j = 1, \ldots, p \), the following conditions are satisfied: (i) the true parameter \( \alpha_j \) obeys (16); \( \hat{\mathcal{A}}_j \) is a possibly stochastic interval such that with probability \( 1 - \delta_n \), \( \{ \alpha_j + c_1 n^{-1/2} \log^2 a_n \} \subset \hat{\mathcal{A}}_j \subset \mathcal{A}_j \); (ii) for \( P \)-almost every \( z_j \), the map \( (\alpha, t) \to E\{\psi_j(w, \alpha, t) \mid z_j\} \) is twice continuously differentiable, and for every \( \nu \in \{\alpha, t_1, \ldots, t_M\} \), \( E(\sup_{\alpha_j \in \hat{\mathcal{A}}_j} |\partial \partial_{\nu} E\{\psi_j(w, \alpha, h_j(z_j)) \mid z_j\}| \leq C_1 \) and, moreover, there exist constants \( L_{1n} \geq 1, L_{2n} \geq 1, \) and a cube \( T_j(z_j) = \times_{m=1}^M T_{jm}(z_j) \) in \( \mathbb{R}^M \) with center \( h_j(z_j) \) such that for every \( \nu, \nu' \in \{\alpha, t_1, \ldots, t_M\} \), \( \sup_{(\alpha, t) \in \hat{\mathcal{A}}_j \times T_j(z_j)} |\partial \partial_{\nu} E\{\psi_j(w, \alpha, t) \mid z_j\}| \leq L_{1n} \), and for every \( \alpha, \alpha' \in \mathcal{A}_j, t, t' \in T_j(z_j), E[|\psi_j(w, \alpha, t) - \psi_j(w, \alpha', t')|^2 \mid z_j] \leq L_{2n}(|\alpha - \alpha'| + |t - t'|^c) \); (iii) the orthogonality condition (18) holds; (iv) the following global and local identifiability conditions hold: \( 2E[|\psi_j(w, \alpha, h_j(z_j))|] \geq |\Gamma_j(\alpha - \alpha_j) \wedge c_1 \) for all \( \alpha \in \mathcal{A}_j \), where \( \Gamma_j = \partial_{\nu} E\{\psi_j(w, \alpha_j, h_j(z_j)) \} \), and \( |\Gamma_j| \geq c_1 \); and (v) the second moments of scores are bounded away from zero: \( E[\psi_j^2(w, \alpha_j, h_j(z_j))] \geq c_1 \).

Condition 2 states rather mild assumptions for Z-estimation problems, in particular, allowing for non-smooth scores \( \psi_j \) such as those arising in median regression. They are analogous to assumptions imposed in the setting with \( p = o(n) \), for example, in He & Shao (2000). The following condition uses a notion of pointwise measurable classes of functions (van der Vaart & Wellner, 1996, p.110).

**Condition 3.** Uniformly in \( n \geq n_0, P \in \mathcal{P}_n \), and \( j = 1, \ldots, p \), the following conditions are satisfied: (i) the nuisance function \( h_j = (h_{jm})_{m=1}^M \) has an estimator \( \hat{h}_j = (\hat{h}_{jm})_{m=1}^M \) with good sparsity and rate properties, namely, with probability \( 1 - \delta_n \), \( \hat{h}_j \in \mathcal{H}_j \), where \( \mathcal{H}_j = \times_{m=1}^M \mathcal{H}_{jm} \) and each \( \mathcal{H}_{jm} \) is the class of functions \( \hat{h}_{jm} : \mathcal{Z}_j \to \mathbb{R} \) of the form \( \hat{h}_{jm}(\cdot) = \sum_{k=1}^p f_{jmk}(\cdot) \theta_{mk} \) such that \( ||(\theta_{mk})_{k=1}^p || \leq s, \hat{h}_{jm}(z) \in \mathcal{F}_{jm}(z) \) for all \( z \in \mathcal{Z}_j \), and \( E[|\hat{h}_{jm}(z) - h_{jm}(z)|^2] \leq C_1 s \log(a_n) / n \), where \( s = s_n \geq 1 \) is the sparsity level, obeying (iv) ahead; (ii) the class of functions \( \mathcal{F}_j = \{ w \to \psi_j(w, \alpha, \hat{h}(z_j)) : \alpha \in \mathcal{A}_j, \hat{h} \in \mathcal{H}_j \cup \{ h_j \} \} \) is pointwise measurable and obeys the entropy condition \( \text{ent}(\epsilon, \mathcal{F}_j) \leq C_1 M s \log(a_n / \epsilon) \) for all \( 0 < \epsilon \leq 1 \); (iii) the class \( \mathcal{F}_j \)

\[
\mathcal{F}_j = \{ w \to \psi_j(w, \alpha, \hat{h}(z_j)) : \alpha \in \mathcal{A}_j, \hat{h} \in \mathcal{H}_j \cup \{ h_j \} \}.
\]
has measurable envelope $F_j \geq \sup_{f \in F_j} |f|$, such that $F = \max_{j=1,\ldots,p_1} F_j$ obeys $E\{ F_q^q(w) \} \leq C_1$ for some $q \geq 4$; and (iv) the dimensions $p_1, p$, and $s$ obey the growth conditions:

$$n^{-1/2} \{(s \log a_n)^{1/2} + n^{-1/2+1/q_s \log a_n} \} \leq \rho_n, \quad \rho_n^{5/2}(L_{2n_s \log a_n})^{1/2} + n^{1/2}L_{1n_s \rho_n^2} \leq \delta_n h_n^{-1}.$$  

Condition 3 (i) requires reasonable behavior of sparse estimators $\hat{h}_j$. In the previous section, this type of behavior occurred in the cases where $h_j$ consisted of a part of a median regression function and a conditional expectation function in an auxiliary equation. There are many conditions in the literature that imply these conditions from primitive assumptions. For the case with $q = \infty$, Condition 3 (vi) implies the following restrictions on the sparsity indices:

$$(s^2 \log^3 a_n)/n \to 0$$

for the case where $\zeta = 2$, which typically happens when $\psi_j$ is smooth, and

$$(s^2 \log^5 a_n)/n \to 0$$

for the case where $\zeta = 1$, which typically happens when $\psi_j$ is non-smooth. Condition 3 (iii) bounds the moments of the envelopes, and it can be relaxed to a bound that grows with $n$, with an appropriate strengthening of the growth conditions stated in (iv).

Condition 3 (ii) implicitly requires $\psi_j$ not to increase entropy too much; it holds, for example, when $\psi_j$ is a monotone transformation, as in the case of median regression, or a Lipschitz transformation; see van der Vaart & Wellner (1996). The entropy bound is formulated in terms of the upper bound $s$ on the sparsity of the estimators and $p$ the dimension of the overall approximating model appearing via $a_n$. In principle our main result below applies to non-sparse estimators as well, as long as the entropy bound specified in Condition 3 (ii) holds, with index $(s, p)$ interpreted as measures of effective complexity of the relevant function classes.

Recall that $\Gamma_j = \partial_n E[\psi_j \{ w, \alpha, h_j(z_j) \}]$; see Condition 2 (iii). Define

$$\sigma_j^2 = E[\Gamma_j^{-2} \psi_j^2 \{ w, \alpha, h_j(z_j) \}], \quad \phi_j(w) = -\sigma_j^{-1} \Gamma_j^{-1} \psi_j \{ w, \alpha, h_j(z_j) \} \quad (j = 1, \ldots, p_1).$$

The following is the main theorem of this section; its proof is found in Appendix A.

**THEOREM 2.** Under Conditions 2 and 3, uniformly in $P \in \mathcal{P}_n$, with probability $1 - o(1),

$$\max_{j=1,\ldots,p_1} \left| n^{1/2} \sigma_j^{-1}(\hat{\alpha}_j - \alpha_j) - n^{-1/2} \sum_{i=1}^n \phi_j(w_i) \right| = o(b_n^{-1}), \quad n \to \infty.$$  

An immediate implication is a corollary on the asymptotic normality uniform in $P \in \mathcal{P}_n$ and $j = 1, \ldots, p_1$, which follows from Lyapunov’s central limit theorem for triangular arrays.

**COROLLARY 2.** Under the conditions of Theorem 2,

$$\max_{j=1,\ldots,p_1} \sup_{P \in \mathcal{P}_n} \sup_{t \in \mathbb{R}} \left| \Pr_P \left\{ \frac{n^{1/2} \sigma_j^{-1}(\hat{\alpha}_j - \alpha_j)}{\Phi(t)} \right\} \right| = o(1), \quad n \to \infty.$$  

This implies, provided $\max_{j=1,\ldots,p_1} |\hat{\alpha}_j - \alpha_j| = o_P(1)$ uniformly in $P \in \mathcal{P}_n$, that

$$\max_{j=1,\ldots,p_1} \sup_{P \in \mathcal{P}_n} \left| \Pr_P \left\{ \alpha_j \in \left[ \hat{\alpha}_j \pm \hat{\sigma}_j n^{-1/2} \Phi^{-1}(1 - \xi/2) \right] \right\} - (1 - \xi) \right| = o(1), \quad n \to \infty.$$  

This result leads to marginal confidence intervals for $\alpha_j$, and shows that they are valid uniformly in $P \in \mathcal{P}_n$ and $j = 1, \ldots, p_1$.

Another useful implication is the high-dimensional central limit theorem uniformly over rectangles in $\mathbb{R}^{p_1}$, provided that $(\log p_1)^2 = o(n)$, which follows from Corollary 2.1 in Chernozhukov et al. (2013). Let $\mathcal{N} = (\mathcal{N}_j)_{j=1}^{p_1}$ be a normal random vector in $\mathbb{R}^{p_1}$ with mean zero and covariance matrix $[E\{ \phi_j(w) \phi'_j(w) \}]_{j,j'=1}^{p_1}$. Let $\mathcal{R}$ be a collection of rectangles $R$ in $\mathbb{R}^{p_1}$ of the
form
\[ R = \left\{ z \in \mathbb{R}^{p_1} : \max_{j \in A} z_j \leq t, \max_{j \in B} (-z_j) \leq t \right\} \quad (t \in \mathbb{R}, A, B \subset \{1, \ldots, p_1\}). \]

For example, when \( A = B = \{1, \ldots, p_1\} \), \( R = \{ z \in \mathbb{R}^{p_1} : \max_{j=1,\ldots,p_1} |z_j| \leq t \}. \)

**Corollary 3.** Under the conditions of Theorem 2, provided that \( (\log p_1)^7 = o(n) \),
\[ \sup_{P \in \mathcal{P}_n} \sup_{R \in \mathcal{R}} \left| \Pr_P \left[ n^{1/2} \{ \sigma_j^{-1}(\hat{\alpha}_j - \alpha_j) \}^{p_1}_{j=1} \in R \right] - \Pr_P(\mathcal{N} \in R) \right| = o(1), \quad n \to \infty. \]

This implies, in particular, that for \( c_{1-\xi} = (1 - \xi) \)-quantile of \( \max_{j=1,\ldots,p_1} |\mathcal{N}_j| \),
\[ \sup_{P \in \mathcal{P}_n} \left| \Pr_P \left( \alpha_j \in [\hat{\alpha}_j \pm c_{1-\xi} \sigma_j n^{-1/2}], \ j = 1, \ldots, p_1 \right) - (1 - \xi) \right| = o(1), \quad n \to \infty. \]

This result leads to simultaneous confidence bands for \( (\alpha_j)_{j=1}^{p_1} \) that are valid uniformly in \( P \in \mathcal{P}_n \). Moreover, Corollary 3 is immediately useful for testing multiple hypotheses about \( (\alpha_j)_{j=1}^{p_1} \) via the step-down methods of Romano & Wolf (2005) which control the family-wise error rate; see Chernozhukov et al. (2013) for further discussion of multiple testing with \( p_1 \gg n \).

In practice the distribution of \( \mathcal{N} \) is unknown, since its covariance matrix is unknown, but it can be approximated by the Gaussian multiplier bootstrap, which generates a vector
\[ \mathcal{N}^* = (\mathcal{N}_j^*)_{j=1}^{p_1} = \left\{ \frac{1}{n^{1/2}} \sum_{i=1}^{n} \xi_i \hat{\phi}_j(w_i) \right\}_{j=1}^{p_1}, \tag{19} \]
where \( (\xi_i)_{i=1}^{n} \) are independent standard normal random variables, independent of the data \( (w_i)_{i=1}^{n} \), and \( \hat{\phi}_j \) are any estimators of \( \phi_j \), such that \( \max_{j,j' \in \{1,\ldots,p_1\}} |E_n \{ \hat{\phi}_j(w) \hat{\phi}_{j'}(w) \} - E_n \{ \phi_j(w) \phi_{j'}(w) \}| = o_P(b_n^{-1}) \) uniformly in \( P \in \mathcal{P}_n \). Let \( \tilde{\sigma}_j^2 = E_n \{ \hat{\phi}_j^2(w) \} \). Theorem 3.2 in Chernozhukov et al. (2013) then implies the following result.

**Corollary 4.** Under the conditions of Theorem 2, provided that \( (\log p_1)^7 = o(n) \), with probability \( 1 - o(1) \) uniformly in \( P \in \mathcal{P}_n \),
\[ \sup_{P \in \mathcal{P}_n} \sup_{R \in \mathcal{R}} \left| \Pr_P \{ \mathcal{N}^* \in R \mid (w_i)_{i=1}^{n} \} - \Pr_P(\mathcal{N} \in R) \right| = o(1). \]

This implies, in particular, that for \( \hat{c}_{1-\xi} = (1 - \xi) \)-conditional quantile of \( \max_{j=1,\ldots,p_1} |\mathcal{N}_j^*| \),
\[ \sup_{P \in \mathcal{P}_n} \left| \Pr_P \left( \alpha_j \in [\hat{\alpha}_j \pm \hat{c}_{1-\xi} \tilde{\sigma}_j n^{-1/2}], \ j = 1, \ldots, p_1 \right) - (1 - \xi) \right| = o(1). \]

**Remark 1.** The proof of Theorem 2 shows that the orthogonality condition (18) can be replaced by a more general orthogonality condition:
\[ E[\eta(z_j)^T (\tilde{h}_j(z_j) - h_j(z_j))] = 0, \quad (\tilde{h}_j \in \mathcal{H}_j, \ j = 1, \ldots, p_1), \tag{20} \]
where \( \eta(z_j) = \partial_t E \{ \psi_j(w, \alpha_j, t) \mid z_j \}_{t=\tilde{h}_j(z_j)} \), or even more general condition of approximate orthogonality: \( E[\eta(z_j)^T \{ \tilde{h}_j(z_j) - h_j(z_j) \}] = o(n^{-1/2}b_n^{-1}) \) uniformly in \( \tilde{h}_j \in \mathcal{H}_j \) and \( j = 1, \ldots, p_1 \). The generalization (20) has a number of benefits, which could be well illustrated by the median regression model of Section 1, where the conditional moment restriction \( E(v_i \mid x_i) = 0 \) could be now replaced by the unconditional one \( E(v_i, x_i) = 0 \), which allows for more general forms of data-generating processes.
We consider the regression model

\[ y_i = d_i \alpha_0 + x_i^T (c_\theta \theta_0) + \epsilon_i, \quad d_i = x_i^T (c_d \theta_0) + v_i, \]

where \( \alpha_0 = 1/2, \theta_{0j} = 1/j^2 (j = 1, \ldots, 10) \), and \( \theta_{0j} = 0 \) otherwise, \( x_i = (1, z_i^T)^T \) consists of an intercept and covariates \( z_i \sim N(0, \Sigma) \), and the errors \( \epsilon_i \) and \( v_i \) are independently and identically distributed as \( N(0, 1) \). The dimension \( p \) of the controls \( x_i \) is 300, and the sample size \( n \) is 250. The covariance matrix \( \Sigma \) has entries \( \Sigma_{ij} = \rho^{|i-j|} \) with \( \rho = 0.5 \). The coefficients \( c_y \) and \( c_d \) determine the \( R^2 \) in the equations \( y_i - d_i \alpha_0 = x_i^T (c_\theta \theta_0) + \epsilon_i \) and \( d_i = x_i^T (c_d \theta_0) + v_i \). We vary the \( R^2 \) in the two equations, denoted by \( R_{y_1}^2 \) and \( R_{d_1}^2 \) respectively, in the set \{0, 0.1, \ldots, 0.9\}, which results in 100 different designs induced by the different pairs of \( (R_{y_1}^2, R_{d_1}^2) \); we performed 500 Monte Carlo repetitions for each.

The first equation in (21) is a sparse model. However, unless \( c_y \) is very large, the decay of the components of \( \theta_0 \) rules out the typical assumption that the coefficients of important regressors are well separated from zero. Thus we anticipate that the standard post-selection inference procedure, discussed around (3), would work poorly in the simulations. In contrast, from the prior theoretical arguments, we anticipate that our instrumental median estimator would work well.

The simulation study focuses on Algorithm 1, since Algorithm 2 performs similarly. Standard errors are computed using (11). As the main benchmark we consider the standard post-model selection estimator \( \tilde{\alpha} \) based on the post \( \ell_1 \)-penalized median regression method (3).

In Figure 1, we display the empirical false rejection probability of tests of a true hypothesis \( \alpha = \alpha_0 \), with nominal size 5%. The false rejection probability of the standard post-model selection inference procedure based upon \( \tilde{\alpha} \) deviates sharply from the nominal size. This confirms the anticipated failure, or lack of uniform validity, of inference based upon the standard post-model selection procedure in designs where coefficients are not well separated from zero so that perfect model selection does not happen. In sharp contrast, both of our proposed procedures, based on estimator \( \hat{\alpha} \) and the result (10) and on the statistic \( L_{\hat{\alpha}} \) and the result (13), closely track the nominal size. This is achieved uniformly over all the designs considered in the study, and confirms the theoretical results of Corollary 1.

In Figure 2, we compare the performance of the standard post-selection estimator \( \tilde{\alpha} \) and our proposed post-selection estimator \( \hat{\alpha} \). We use three different measures of performance of the two approaches: mean bias, standard deviation, and root mean square error. The significant bias for the standard post-selection procedure occurs when the main regressor \( d_i \) is correlated with other controls \( x_i \). The proposed post-selection estimator \( \hat{\alpha} \) performs well in all three measures. The root mean square errors of \( \hat{\alpha} \) are typically much smaller than those of \( \tilde{\alpha} \), fully consistent with our theoretical results and the semiparametric efficiency of \( \hat{\alpha} \).

**Acknowledgments**

This paper was first presented in 8th World Congress in Probability and Statistics in August 2012. We would like to thank the editor, an associate editor, and anonymous referees for their careful review. We are also grateful to Sara van de Geer, Xuming He, Richard Nickl, Roger Koenker, Vladimir Koltchinskii, Enno Mammen, Steve Portnoy, Philippe Rigollet, Richard Samworth, and Bin Yu for useful comments and discussions. Research support from the National Science Foundation and the Japan Society for the Promotion of Science is gratefully acknowledged.
Fig. 1. The empirical false rejection probabilities of the nominal 5% level tests based on: (a) the standard post-model selection procedure based on $\tilde{\alpha}$, (b) the proposed post-model selection procedure based on $\hat{\alpha}$, (c) the score statistic $L_n$, and (d) an ideal procedure with the false rejection rate equal to the nominal size.

SUPPLEMENTARY MATERIAL

In the supplementary material we provide omitted proofs, technical lemmas, discuss extensions to the heteroscedastic case, and alternative implementations.

A. PROOF OF THEOREM 2

A-1. A maximal inequality

We first state a maximal inequality used in the proof of Theorem 2.

**Lemma A1.** Let $w, w_1, \ldots, w_n$ be independent and identically distributed random variables taking values in a measurable space, and let $F$ be a pointwise measurable class of functions on that space. Suppose that there is a measurable envelope $F \geq \sup_{f \in F} |f|$ such that $E\{F^q(w)\} < \infty$ for some $q \geq 2$. Consider the empirical process indexed by $F$: $G_n(f) = n^{-1/2} \sum_{i=1}^n [f(w_i) - E\{f(w)\}]$, $f \in F$. Let $\sigma > 0$ be any positive constant such that $\sup_{f \in F} E\{f^2(w)\} \leq \sigma^2 \leq E\{F^2(w)\}$. Moreover, suppose that
Fig. 2. Mean bias (top row), standard deviation (middle row), root mean square (bottom row) of the standard post-model selection estimator $\tilde{\alpha}$ (panels (a)-(c)), and of the proposed post-model selection estimator $\hat{\alpha}$ (panels (d)-(f)).

there exist constants $A \geq e$ and $s \geq 1$ such that $\text{ent}(\varepsilon, \mathcal{F}) \leq s \log(A/\varepsilon)$ for all $0 < \varepsilon \leq 1$. Then

\[
E\left\{\sup_{f \in \mathcal{F}} |G_n(f)|\right\} \leq K \left\{\min_{i=1,...,n} \frac{\max_{i=1,...,n} F^2(w_i)}{\sigma} \right\}^{1/2} \\
+ n^{-1/2+1/q} \log(A) \left\{E\left\{F^q(w)\right\}\right\}^{1/2}
\]

where $K$ is a universal constant. Moreover, for every $t \geq 1$, with probability not less than $1 - t^{-q/2}$,

\[
\sup_{f \in \mathcal{F}} |G_n(f)| \leq 2E\left\{\sup_{f \in \mathcal{F}} |G_n(f)|\right\} + K_q \left(\sigma t^{1/2} + n^{-1/2+1/q} \log(A) \left\{E\left\{F^q(w)\right\}\right\}^{1/2}\right),
\]

where $K_q$ is a constant that depends only on $q$.

\textbf{Proof.} The first and second inequalities follow from Corollary 5.1 and Theorem 5.1 in Chernozhukov et al. (2014) applied with $\alpha = 1$, using that $\left\{E\left\{\max_{i=1,...,n} F^2(w_i)\right\}\right\}^{1/2} \leq \left\{E\left\{\max_{i=1,...,n} F^q(w_i)\right\}\right\}^{1/q} \leq n^{1/q} \left\{E\left\{F^q(w)\right\}\right\}^{1/q}. \quad \square
A.2. Proof of Theorem 2

It suffices to prove the theorem under any sequence \( P = P_n \in \mathcal{P}_n \). We shall suppress the dependence of \( P \) on \( n \) in the proof. In this proof, let \( C \) denote a generic positive constant that may differ in each appearance, but that does not depend on the sequence \( P \in \mathcal{P}_n, n \), or \( j = 1, \ldots, p_1 \). Recall that the sequence \( \rho_n \downarrow 0 \) satisfies the growth conditions in Condition 3 (iv). We divide the proof into three steps. Below we use the following notation: for any given function \( g : \mathbb{W} \rightarrow \mathbb{R}, G_n(g) = n^{-1/2} \sum_{i=1}^n [g(w_i) - E\{g(w)\}] \).

Step 1. Let \( \tilde{\alpha}_j \) be any estimator such that with probability \( 1 - o(1) \),

\[
E_n[\psi_j\{w, \tilde{\alpha}_j, \tilde{h}_j(z_j)\}] = E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + E_{\tilde{\alpha}_j}(\tilde{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1}),
\]

uniformly in \( j = 1, \ldots, p_1 \). Expand

\[
E_n[\psi_j\{w, \tilde{\alpha}_j, \tilde{h}_j(z_j)\}] = E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + n^{-1/2}G_n[\psi_j\{w, \alpha_j, \tilde{h}_j(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\}] = I_j + II_j + III_j,
\]

where we have used \( E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = 0 \). We first bound \( III_j \). Observe that, with probability \( 1 - o(1) \), \( \max_{j=1,\ldots,p_1} |III_j| \leq n^{-1/2} \sup_{f \in \mathcal{F}} |G_n(f)| \), where \( \mathcal{F} \) is the class of functions defined by

\[
\mathcal{F} = \{ w \mapsto \psi_j\{w, \alpha, \tilde{h}(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\} : j = 1, \ldots, p_1, \tilde{h} \in H_j, \alpha \in A_j, |\alpha - \alpha_j| \leq C\rho_n \},
\]

which has \( 2\mathcal{F} \) as an envelope. We apply Lemma 1 to this class of functions. By Condition 3 (ii) and a simple covering number calculation, we have \( \text{ent}(\varepsilon, \mathcal{F}) \leq Cs \log(a_n/\varepsilon) \). By Condition 2 (ii), \( \sup_{f \in \mathcal{F}} E\{f^2(w)\} \) is bounded by

\[
\sup_{j=1,\ldots,\max_{j=1,\ldots,p_1}} E\left\{ E\left( \left[ \psi_j\{w, \alpha, \tilde{h}(z_j)\} - \psi_j\{w, \alpha_j, h_j(z_j)\} \right]^2 \right) \left| z_j \right) \right\} \leq C L_2n\rho_n^2,
\]

where we have used the fact that \( E\{[\tilde{h}_m(z_j) - h_{jm}(z_j)]^2\} \leq C\rho_n^2 \) for all \( m = 1, \ldots, M \) whenever \( \tilde{h} = (\tilde{h}_m)_{m=1}^M \in H_j \). Hence applying Lemma 1 with \( t = \log n \), we conclude that, with probability \( 1 - o(1) \),

\[
n^{1/2} \max_{j=1,\ldots,\max_{j=1,\ldots,p_1}} |III_j| \leq C \rho_n^2 (L_2 n \log a_n)^{1/2} + n^{-1/2+1/4} \log a_n = o(b_n^{-1}),
\]

where the last equality follows from Condition 3 (iv).

Next, we expand \( II_j \). Pick any \( \alpha \in A_j \) with \( |\alpha - \alpha_j| \leq C\rho_n, \tilde{h} = (\tilde{h}_m)_{m=1}^M \in H_j \). Then by Taylor’s theorem, for any \( j = 1, \ldots, p_1 \) and \( z_j \in Z_j \), there exists a vector \((\tilde{\alpha}(z_j), \bar{z}(z_j))^T\) on the line segment joining \((\alpha, \tilde{h}(z_j))^T\) and \((\alpha_j, h_j(z_j))^T\) such that \( E[\psi_j\{w, \alpha, \tilde{h}(z_j)\}] \) can be written as

\[
E[\psi_j\{w, \alpha_j, h_j(z_j)\}] + E(\partial_n E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j])(\alpha - \alpha_j) + \sum_{m=1}^M E\{(\partial_m E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j])(\alpha - \alpha_j)\}\left[ \tilde{h}_m(z_j) - h_{jm}(z_j) \right] \}
\]

The third term is zero because of the orthogonality condition (18). Condition 2 (ii) guarantees that the expectation and derivative can be interchanged for the second term, that is,

\[
E(\partial_n E[\psi_j\{w, \alpha_j, h_j(z_j)\} \mid z_j]) = \partial_n E[\psi_j\{w, \alpha_j, h_j(z_j)\}] = \Gamma_j.
\]

Moreover, by the same condition, each of the last three terms is bounded by \( C L_1n\rho_n^2 = o(n^{-1/2}b_n^{-1}) \), uniformly in \( j = 1, \ldots, p_1 \). Therefore, with probability \( 1 - o(1) \),

\[
II_j = \Gamma_j(\tilde{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1}),
\]

uniformly in \( j = 1, \ldots, p_1 \). Combining the previous bound on \( III_j \) with these bounds leads to the desired assertion.
Step 2. We wish to show that with probability $1 - o(1)$, $\inf_{\alpha \in \tilde{A}_j} |E_n[\psi_j\{w, \alpha, \widehat{h}_j(z_j)\}]| = o(n^{-1/2}b_n^{-1})$, uniformly in $j = 1, \ldots, p_1$. Define $\alpha^*_j = \alpha_j - \Gamma_j^{-1}E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]$ ($j = 1, \ldots, p_1$). Then we have $\max_{j=1,\ldots,p_1} |\alpha^*_j - \alpha_j| \leq C \max_{j=1,\ldots,p_1} |E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]|$. Consider the class of functions $F^j = \{w \mapsto \psi_j\{w, \alpha_j, h_j(z_j)\} : j = 1, \ldots, p_1\}$, which has $\tilde{F}$ as an envelope.

Since this class is finite with cardinality $p_1$, we have $\text{ent}(\varepsilon, F^j) \leq \log(p_1/\varepsilon)$. Hence applying Lemma 1 to $F^j$ with $\sigma = |E[|F^2(w)|]|^{1/2} \leq C$ and $t = \log n$, we conclude that with probability $1 - o(1)$,

$$\max_{j=1,\ldots,p_1} |E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}]| \leq Cn^{-1/2}\{\log a_n\}^{1/2} + n^{-1/2+1/\eta}\log a_n \leq Cn^{-1/2}\log a_n.$$  

Since $\tilde{A}_j \supset [\alpha_j \pm c_1n^{-1/2}\log^2 a_n]$ with probability $1 - o(1)$, $\alpha^*_j \in \tilde{A}_j$ with probability $1 - o(1)$.

Therefore, using Step 1 with $\tilde{\alpha}_j = \alpha^*_j$, we have, with probability $1 - o(1)$,

$$\inf_{\alpha \in \tilde{A}_j} \left| E_n[\psi_j\{w, \alpha, \widehat{h}_j(z_j)\}] \right| \leq \left| E_n[\psi_j\{w, \alpha^*_j, \widehat{h}_j(z_j)\}] \right| = o(n^{-1/2}b_n^{-1}),$$

uniformly in $j = 1, \ldots, p_1$, where we have used the fact that $E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + \Gamma_j(\alpha^*_j - \alpha_j) = 0$.

Step 3. We wish to show that with probability $1 - o(1)$, $\max_{j=1,\ldots,p_1} |\tilde{\alpha}_j - \alpha_j| \leq C\rho_n$. By Step 2 and the definition of $\tilde{\alpha}_j$, with probability $1 - o(1)$, we have $\max_{j=1,\ldots,p_1} |\tilde{\alpha}_j - \alpha_j| = o(n^{-1/2}b_n^{-1})$. Consider the class of functions $F'' = \{w \mapsto \psi_j\{w, \alpha, \widehat{h}(z_j)\} : j = 1, \ldots, p_1, \alpha \in \tilde{A}_j, \widehat{h} \in \mathcal{H}_j \cup \{\widehat{h}\} \}$. Then with probability $1 - o(1)$,

$$\left| E_n[\psi_j\{w, \tilde{\alpha}_j, \widehat{h}(z_j)\}] \right| \geq \left| E[\psi_j\{w, \alpha, \widehat{h}(z_j)\}]\right|_{\alpha = \tilde{\alpha}_j, \widehat{h} = \widehat{h}} - n^{-1/2}\sup_{f \in F''} |G_n(f)|,$$

uniformly in $j = 1, \ldots, p_1$. Observe that $F''$ has $\tilde{F}$ as an envelope and, by Condition 3 (ii) and a simple covering number calculation, $\text{ent}(\varepsilon, F'') \leq Cs\log(a_n/\varepsilon)$. Then applying Lemma 1 with $\sigma = |E[|F'(w)|]|^{1/2} \leq C$ and $t = \log n$, we have, with probability $1 - o(1)$,

$$n^{-1/2}\sup_{f \in F''} |G_n(f)| \leq Cn^{-1/2}\{s\log a_n\}^{1/2} + n^{-1/2+1/\eta}s\log a_n = O(\rho_n).$$

Moreover, application of the expansion (A1) with $\alpha_j = \alpha$ together with the Cauchy–Schwarz inequality implies that $|E[\psi_j\{w, \alpha, \widehat{h}(z_j)\}] - E[\psi_j\{w, \alpha_j, h_j(z_j)\}]|$ is bounded by $C(\rho_n + L_1\rho_n^2) = O(\rho_n)$, so that with probability $1 - o(1)$,

$$\left| E[\psi_j\{w, \alpha, \widehat{h}(z_j)\}]\right|_{\alpha = \tilde{\alpha}_j, \widehat{h} = \widehat{h}} \geq \left| E[\psi_j\{w, \alpha_j, h_j(z_j)\}]\right|_{\alpha = \tilde{\alpha}_j} - O(\rho_n),$$

uniformly in $j = 1, \ldots, p_1$, where we have used Condition 2 (ii) together with the fact that $E[|\tilde{h}_m(z_j) - h_m(z_j)|^2] \leq C\rho_n^2$ for all $m = 1, \ldots, M$ whenever $\tilde{h} = (\tilde{h}_m)_{m=1}^M \in \mathcal{H}_j$. By Condition 2 (iv), the first term on the right side is bounded from below by $(1/2)\{\Gamma_j(\alpha_j - \alpha_j)\} \wedge c_1$, which, combined with the fact that $|\Gamma_j| \geq c_1$, implies that with probability $1 - o(1)$, $|\tilde{\alpha}_j - \alpha_j| \leq o(n^{-1/2}b_n^{-1}) + O(\rho_n) = O(\rho_n)$, uniformly in $j = 1, \ldots, p_1$.

Step 4. By Steps 1 and 3, with probability $1 - o(1)$,

$$E_n[\psi_j\{w, \tilde{\alpha}_j, \widehat{h}(z_j)\}] = E_n[\psi_j\{w, \alpha_j, h_j(z_j)\}] + \Gamma_j(\tilde{\alpha}_j - \alpha_j) + o(n^{-1/2}b_n^{-1}),$$

uniformly in $j = 1, \ldots, p_1$. Moreover, by Step 2, with probability $1 - o(1)$, the left side is $o(n^{-1/2}b_n^{-1})$ uniformly in $j = 1, \ldots, p_1$. Solving this equation with respect to $\tilde{\alpha}_j - \alpha_j$ leads to the conclusion of the theorem.

REFERENCES

Belloni, A., Chen, D., Chernozhukov, V. & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2430.

Belloni, A. & Chernozhukov, V. (2011). $\ell_1$-penalized quantile regression for high dimensional sparse models. *Ann. Statist.* **39**, 82–130.
Belloni, A., Chernozhukov, V. & Hansen, C. (2013). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: The 2010 World Congress of the Econometric Society* 3, 245–295.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014a). Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.* 81, 608–650.

Belloni, A., Chernozhukov, V. & Wang, L. (2014b). Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.* 42, 757–788.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* 37, 1705–1732.

Candes, E. & Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *Ann. Statist.* 35, 2313–2351.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* 41, 2786–2819.

Chernozhukov, V., Chetverikov, D. & Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* 42, 1564–1597.

Chernozhukov, V. & Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *J. Econometrics* 142, 379–398.

He, X. & Shao, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* 73, 120–135.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* 1, 799–821.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stud.* 86, 4–29.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory* 21, 21–59.

Leeb, H. & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges’ estimator. *J. Econometrics* 142, 201–211.

Liang, H., Wang, S., Robins, J. M. & Carroll, R. J. (2004). Estimation in partially linear models with missing covariates. *J. Amer. Statist. Assoc.* 99, 357–367.

Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics, the Harold Cramer Volume*, U. Grenander, ed. New York: John Wiley and Sons, Inc.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of $p$ regression parameters when $p^2/n$ is large. I. Consistency. *Ann. Statist.* 12, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of $p$ regression parameters when $p^2/n$ is large. II. Normal approximation. *Ann. Statist.* 13, 1251–1638.

Powell, J. L. (1986). Censored regression quantiles. *J. Econometrics* 32, 143–155.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* 90, 122–129.

Robinson, P. M. (1988). Root-$n$-consistent semiparametric regression. *Econometrica* 56, 931–954.

Romano, J. P. & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica* 73, 1237–1282.

Rudelson, M. & Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* 59, 3434–3447.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B* 58, 267–288.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer-Verlag.

Wang, L. (2013). $L_1$ penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.* 120, 135–151.

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *J. R. Statist. Soc. B* 76, 217–242.
CENTRAL LIMIT THEOREMS AND BOOTSTRAP IN HIGH DIMENSIONS

VICTOR CHERNOZHUKOV, DENIS CHETVERIKOV, AND KENGO KATO

Abstract. In this paper, we derive central limit and bootstrap theorems for probabilities that centered high-dimensional vector sums hit rectangles and sparsely convex sets. Specifically, we derive Gaussian and bootstrap approximations for the probabilities $\Pr(n^{-1/2} \sum_{i=1}^n X_i \in A)$ where $X_1, \ldots, X_n$ are independent random vectors in $\mathbb{R}^p$ and $A$ is a rectangle, or, more generally, a sparsely convex set, and show that the approximation error converges to zero even if $p = p_n \to \infty$ and $p \gg n$; in particular, $p$ can be as large as $O(e^{Cn^c})$ for some constants $c, C > 0$. The result holds uniformly over all rectangles, or more generally, sparsely convex sets, and does not require any restrictions on the correlation among coordinates of $X_i$. Sparsely convex sets are sets that can be represented as intersections of many convex sets whose indicator functions depend nontrivially only on a small subset of their arguments, with rectangles being a special case.

1. Introduction

Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^p$ where $p \geq 2$ may be large or even much larger than $n$. Denote by $X_{ij}$ the $j$-th coordinate of $X_i$, so that $X_i = (X_{i1}, \ldots, X_{ip})'$. We assume that each $X_i$ is centered, namely $E[X_{ij}] = 0$, and $E[X_{ij}^2] < \infty$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Define the normalized sum $S_n^X := (S_1^X, \ldots, S_n^X)' := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$.

We consider Gaussian approximation to $S_n^X$, and to this end, let $Y_1, \ldots, Y_n$ be independent centered Gaussian random vectors in $\mathbb{R}^p$ such that each $Y_i$ has the same covariance matrix as $X_i$, that is, $Y_i \sim N(0, E[X_i X_i'])$. Define the normalized sum for the Gaussian random vectors: $S_n^Y := (S_1^Y, \ldots, S_n^Y)' := \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$.

Date: First version: February 07, 2014. This version: March 10, 2015.

Key words and phrases. Central limit theorem, bootstrap limit theorems, high dimensions, rectangles, sparsely convex sets.

We are grateful to Evarist Giné, Ramon van Handel, Vladimir Koltchinskii, Richard Nickl, and Larry Wasserman for useful discussions.
We are interested in bounding the quantity

\[ \rho_n(\mathcal{A}) := \sup_{A \in \mathcal{A}} |P(S_n^X \in A) - P(S_n^Y \in A)|, \tag{1} \]

where \( \mathcal{A} \) is a class of Borel sets in \( \mathbb{R}^p \).

Bounding \( \rho_n(\mathcal{A}) \) for various classes \( \mathcal{A} \) of sets in \( \mathbb{R}^p \), with a special emphasis on explicit dependence on the dimension \( p \) in bounds, has been studied by a number of authors; see, for example, \[6, 7, 8, 21, 27, 33, 34, 35, 36\] (see \[16\] for an exhaustive literature review). Typically, we are interested in how fast \( p = p_n \to \infty \) is allowed to grow while guaranteeing \( \rho_n(\mathcal{A}) \to 0 \). In particular, for \( I \) being the \( p \times p \) identity matrix, Bentkus \[7\] established one of the sharpest results in this direction and proved that when \( X_1, \ldots, X_n \) are i.i.d. with \( E[X_iX_i'] = I \),

\[ \rho_n(\mathcal{A}) \leq c_p(\mathcal{A}) \frac{E[\|X_i\|^2]}{\sqrt{n}}, \tag{2} \]

where \( c_p(\mathcal{A}) \) is a constant that depends only on \( p \) and \( \mathcal{A} \); for example, \( c_p(\mathcal{A}) \) is bounded by a universal constant when \( \mathcal{A} \) is the class of all Euclidean balls in \( \mathbb{R}^p \), and \( c_p(\mathcal{A}) \leq 400^{1/4} \) when \( \mathcal{A} \) is the class of all convex sets in \( \mathbb{R}^p \). Note, however, that this bound does not allow \( p \) to be larger than \( n \) once we require \( \rho_n(\mathcal{A}) \to 0 \). Indeed by Hölder’s inequality, when \( E[X_iX_i'] = I \),

\[ E[\|X_i\|^3] \geq (E[\|X_i\|^2])^{3/2} = p^{3/2}, \]

and hence in order to make the right hand side of (2) to be \( o(1) \), we at least need \( p = o(n^{1/3}) \) when \( \mathcal{A} \) is the class of Euclidean balls, and \( p = o(n^{2/7}) \) when \( \mathcal{A} \) is the class of all convex sets. Similar conditions are needed in other papers cited above. It is worthwhile to mention here that, when \( \mathcal{A} \) is the class of all convex sets, it was proved by \[27\] that \( \rho_n(\mathcal{A}) \geq cE[\|X_i\|^3]/\sqrt{n} \) for some universal constant \( c > 0 \).

In modern statistical applications, such as high dimensional estimation and multiple hypothesis testing, however, \( p \) is often larger or even much larger than \( n \). It is therefore interesting to ask whether it is possible to provide a nontrivial class of sets \( \mathcal{A} \) in \( \mathbb{R}^p \) for which we would have

\[ \rho_n(\mathcal{A}) \to 0 \text{ even if } p \text{ is potentially larger or much larger than } n. \tag{3} \]

In this paper, we derive bounds on \( \rho_n(\mathcal{A}) \) for \( \mathcal{A} = \mathcal{A}_{\text{re}} \) being the class of all rectangles, or more generally for \( \mathcal{A} = \mathcal{A}_{\text{si}} \) being the class of simple convex sets, and show that these bounds lead to (3). We call any convex set a simple convex set if it can be well approximated by an affine transformation of a rectangle. An extension to simple convex sets is interesting because it allows us to derive similar bounds for \( \mathcal{A} = \mathcal{A}_{\text{sp}}(s) \) being the class of \((s-)\)-sparsely convex sets. These are sets that can be represented as an intersection of many convex sets whose indicator functions depend nontrivially at most on \( s \) elements of their arguments (for some small \( s \)).

The sets considered are useful for applications in mathematical statistics. In particular, rectangles and sparsely convex sets are interesting because
they allow us to approximate the probabilities of various key statistics exceeding or falling below certain thresholds. For example, the probability that a collection of Kolmogorov-type statistics falls below a collection of thresholds

$$P \left( \max_{j \in J_k} S^{X}_{nj} \leq t_k \text{ for all } k = 1, \ldots, \kappa \right) = P \left( S^{X}_n \in A \right)$$

can be approximated by $P(S^{Y}_n \in A)$ within the error margin $\rho_n(\mathcal{A}^{re})$; here \( \{J_k\} \) are subsets of \( \{1, \ldots, p\} \), \( \{t_k\} \) are thresholds in the interval \((-\infty, \infty)\), \( 1 \leq \kappa < 2^p \) is an integer, and \( A \in \mathcal{A}^{re} \) is a rectangle of the form \( \{w \in \mathbb{R}^p : \max_{j \in J_k} w_j \leq t_k \text{ for all } k = 1, \ldots, \kappa\} \). Another example is the probability that a collection of Pearson-type statistics falls below a collection of thresholds

$$P \left( \| (S^{X}_{nj})_{j \in J_k} \|^2 \leq t_k \text{ for all } k = 1, \ldots, \kappa \right) = P \left( S^{X}_n \in A \right)$$

can be approximated by $P(S^{Y}_n \in A)$ within the error margin $\rho_n(\mathcal{A}^{sp}(s))$; here \( \{J_k\} \) are subsets of \( \{1, \ldots, p\} \) of fixed cardinality \( s \), \( \{t_k\} \) are thresholds in the interval \((0, \infty)\), \( 1 \leq \kappa \leq C^p_s \) is an integer, and \( A \in \mathcal{A}^{sp}(s) \) is a sparsely convex set of the form \( \{w \in \mathbb{R}^p : \| (w_j)_{j \in J_k} \|^2 \leq t_k \text{ for all } k = 1, \ldots, \kappa\} \). In practice, as we demonstrate, the approximations above could be estimated using the empirical or multiplier bootstraps.

The results in this paper substantially extend those obtained in [15] where we considered the class \( \mathcal{A} = \mathcal{A}^m \) of sets of the form \( A = \{w \in \mathbb{R}^p : \max_{j \in J} w_j \leq a\} \) for some \( a \in \mathbb{R} \) and \( J \subset \{1, \ldots, p\} \), but in order to obtain much better dependence on \( n \), we employ new techniques. Most notably, we employ an induction argument as the main ingredient in the new proof, as inspired by Bolthausen [9]. Our paper builds upon our previous work [15], which in turn builds on a number of works listed in the bibliography (see [16] for a detailed review and links to the literature).

The organization of this paper is as follows. In Section 2, we derive a Central Limit Theorem (CLT) for rectangles in high dimensions; that is, we derive a bound on \( \rho_n(\mathcal{A}) \) for \( \mathcal{A} = \mathcal{A}^{re} \) being the class of all rectangles and show that the bound converges to zero under certain conditions even when \( p \) is potentially larger or much larger than \( n \). In Section 3, we extend this result by showing that similar bounds apply for \( \mathcal{A} = \mathcal{A}^{si} \) being a class of simple convex sets and for \( \mathcal{A} = \mathcal{A}^{sp}(s) \) being a class of sparsely convex sets. In Section 4, we derive high dimensional Empirical and Multiplier Bootstrap theorems that allow us to approximate $P(S^{Y}_n \in A)$ for \( A \in \mathcal{A}^{re}, \mathcal{A}^{si}, \) or \( \mathcal{A}^{sp}(s) \) using the data \( X_1, \ldots, X_n \). In Section 5, we state an induction lemma, a key result underlying the derivations in the paper. Finally, we provide all proofs as well as some technical results in the Appendix.

1.1. Notation. For \( a \in \mathbb{R} \), \( [a] \) denotes the largest integer smaller than or equal to \( a \). For \( w = (w_1, \ldots, w_p)' \in \mathbb{R}^p \) and \( y = (y_1, \ldots, y_p)' \in \mathbb{R}^p \), we write \( w \leq y \) if \( w_j \leq y_j \) for all \( j = 1, \ldots, p \). For \( y = (y_1, \ldots, y_p)' \in \mathbb{R}^p \) and \( a \in \mathbb{R} \),
we write $y+a = (y_1+a, \ldots, y_p+a)'$. Throughout the paper, $\mathbb{E}_n[\cdot]$ denotes the average over index $i = 1, \ldots, n$; that is, it simply abbreviates the notation $n^{-1} \sum_{i=1}^n [\cdot]$. For example, $\mathbb{E}_n[x_{ij}] = n^{-1} \sum_{i=1}^n x_{ij}$. We also write $X_i^T := \{X_1, \ldots, X_n\}$. For $v \in \mathbb{R}^p$, we use the notation $\|v\|_0 := \sum_{j=1}^p 1\{v_j \neq 0\}$ and $\|v\| = (\sum_{j=1}^p v_j^2)^{1/2}$. For $\alpha > 0$, we define the function $\psi_\alpha : [0, \infty) \to [0, \infty)$ by $\psi_\alpha(x) := \exp(x^\alpha) - 1$, and for a real-valued random variable $\xi$, we define $\|\xi\|_{\psi_\alpha} := \inf\{\lambda > 0 : \mathbb{E}[\psi_\alpha(|\xi|/\lambda)] \leq 1\}$.

For $\alpha \geq 1$, $\|\cdot\|_{\psi_\alpha}$ is an Orlicz norm, while for $\alpha \in (0, 1)$, $\|\cdot\|_{\psi_\alpha}$ is not a norm but a quasi-norm, that is, there exists a constant $K_\alpha$ depending only on $\alpha$ such that $\|\xi_1 + \xi_2\|_{\psi_\alpha} \leq K_\alpha(\|\xi_1\|_{\psi_\alpha} + \|\xi_2\|_{\psi_\alpha})$. Throughout the paper, we assume that $n \geq 4$ and $p \geq 2$.

## 2. High Dimensional CLT for Rectangles

This section presents a high dimensional CLT for rectangles. We begin with presenting an abstract theorem (Theorem 2.1) that has wide applicability but depends on the tail properties of the distributions of $X_{ij}$'s in a nontrivial way. Then we apply this theorem under simple moment conditions to derive more explicit bounds in Corollary 2.1.

Let $\mathcal{A}^{re}$ be the class of all rectangles in $\mathbb{R}^p$; that is, $\mathcal{A}^{re}$ consists of all sets $A$ of the form

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, p\}$$

for some $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, p$. We will derive a bound on $\rho_n(\mathcal{A}^{re})$, and show that under certain conditions it leads to $\rho_n(\mathcal{A}^{re}) \to 0$ even when $p = p_n$ is potentially larger or much larger than $n$.

To describe the bound, we need to prepare some notation. Define

$$L_n := \max_{1 \leq j \leq p} \sum_{i=1}^n \mathbb{E}|X_{ij}|^3/n,$$

and for $\phi \geq 1$, define

$$M_{n,X}(\phi) := n^{-1} \sum_{i=1}^n \mathbb{E}\left[\max_{1 \leq j \leq p} |X_{ij}|^3 1\left\{\max_{1 \leq j \leq p} |X_{ij}| > \sqrt{n}/(4\phi \log p)\right\}\right].$$

Similarly, define $M_{n,Y}(\phi)$ with $X_{ij}$'s replaced by $Y_{ij}$'s in (5), and let

$$M_n(\phi) := M_{n,X}(\phi) + M_{n,Y}(\phi).$$

The following is the first main result of this paper.

**Theorem 2.1** (Abstract High Dimensional CLT for Rectangles). Suppose that there exists some constant $b > 0$ such that $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \geq b$ for all $j = 1, \ldots, p$. Then there exist constants $K_1, K_2 > 0$ depending only $b$ such that...
that for every constant $L_n \geq L_n$,
\[
\rho_n(\mathcal{A}^{re}) \leq K_1 \left( \left( \frac{L_n^2 \log^2 p}{n} \right)^{\frac{1}{6}} + \frac{M_n(\phi_n)}{L_n} \right)
\]
with
\[
\phi_n : = K_2 \left( \frac{L_n \log^4 p}{n} \right)^{-\frac{1}{6}}
\]
\[ \text{(6)} \]

**Remark 2.1** (Key features of Theorem 2.1). (i) The bound (6) should be contrasted with Bentkus’s [7] bound (2). For the sake of exposition, assume that the vectors $X_1, \ldots, X_n$ are such that $E[X_{ij}^2] = 1$ and for some sequence of constants $B_n \geq 1$, $|X_{ij}| \leq B_n$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Then it can be shown that the bound (6) reduces to
\[
\rho_n(\mathcal{A}^{re}) \leq K \{n^{-1}B_n^2 \log^2 (pn)\}^{\frac{1}{2}}
\]
for some universal constant $K$; see Corollary 2.1 below. Importantly, the right hand side of (8) converges to zero even when $p$ is much larger than $n$; indeed we just need $B_n^2 \log^2 (pn) = o(n)$ to make $\rho_n(\mathcal{A}^{re}) \to 0$, and if in addition $B_n = O(1)$, the condition reduces to $\log p = o(n^{1/7})$. In contrast, Bentkus’s bound (2) requires $p = o(n^{2/7})$ to make $\rho_n(\mathcal{A}) \to 0$ when $\mathcal{A}$ is the class of all convex sets. Hence by restricting the class of sets to the smaller one, $\mathcal{A} = \mathcal{A}^{re}$, we are able to considerably weaken the requirement on $p$.

(ii) On the other hand, the bound in (8) depends on $n$ through $n^{-1/6}$, so that our Theorem 2.1 does not recover the Berry-Esseen bound when $p$ is fixed. However, given that the rate $n^{-1/6}$ is optimal (in a minimax sense) in CLT in infinite dimensional Banach spaces (see [5]), the factor $n^{-1/6}$ seems nearly optimal in terms of dependence on $n$ in the high-dimensional settings as considered here. In addition, examples in [17] suggest that dependence on $B_n$ is also optimal. Hence we conjecture that up to a universal constant,
\[
\{n^{-1}B_n^2 (\log p)^a\}^{\frac{1}{2}}
\]
for some $a > 0$ is an optimal bound (in a minimax sense) in the high dimensional setting as considered here. The value $a = 3$ could be motivated by the theory of moderate deviations for self-normalized sums when all the coordinates of $X_i$ are independent.

**Remark 2.2** (Relation to previous work). Theorem 2.1 extends Theorem 2.2 in [15] where we derived a bound on $\rho_n(\mathcal{A}^{m})$ with $\mathcal{A}^{m} \subset \mathcal{A}^{re}$ consisting of all sets of the form
\[
A = \{w \in \mathbb{R}^p : w_j \leq a \text{ for all } j = 1, \ldots, p\}
\]
for some $a \in \mathbb{R}$. In particular, we improve the dependence on $n$ from $n^{-1/8}$ in [15] to $n^{-6/6}$. In addition, we note that extension to the class $\mathcal{A}^{re}$ from the class $\mathcal{A}^{m}$ is not immediate since in both papers we assume that $\text{Var}(S_{nj}^X)$ is bounded below from zero uniformly in $j = 1, \ldots, p$, so that it is not possible
to directly extend the results in [15] to the class of rectangles $\mathcal{A} = \mathcal{A}^{re}$ by just rescaling the coordinates in $S_n^X$. ■

The bound (6) depends on $M_n(\phi_n)$ whose values are problem specific. Therefore, we now apply Theorem 2.1 in two specific examples that are most useful in mathematical statistics (as well as other related fields such as econometrics). Let $b, q > 0$ be some constants, and let $B_n \geq 1$ be a sequence of constants, possibly growing to infinity as $n \to \infty$. Assume that the following conditions are satisfied:

\begin{enumerate}[(M.1)]
    \item $n^{-1} \sum_{i=1}^{n} E[X_{ij}^2] \geq b$ for all $j = 1, \ldots, p$,
    \item $n^{-1} \sum_{i=1}^{n} E[|X_{ij}|^{2+k}] \leq B_k^n$ for all $j = 1, \ldots, p$ and $k = 1, 2$.
\end{enumerate}

We consider examples where one of the following conditions holds:

\begin{enumerate}[(E.1)]
    \item $E[\exp(|X_{ij}|/B_n)] \leq 2$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$,
    \item $E[\max_{1 \leq j \leq p} |X_{ij}|/B_n]^q \leq 2$ for all $i = 1, \ldots, n$,
\end{enumerate}

An application of Theorem 2.1 under these conditions leads to the following corollary. To avoid the repetitions in stating the results below, let

\begin{equation}
D_n^{(1)} = \left( \frac{B_n^2 \log^7 (pn)}{n} \right)^{1/6}, \quad D_n^{(2)} = \left( \frac{B_n^2 \log^3 p}{n^{1-2/q}} \right)^{1/3}.
\end{equation}

**Corollary 2.1** (High Dimensional CLT for Rectangles). Suppose that conditions (M.1) and (M.2) are satisfied. Then under (E.1), we have

$\rho_n(\mathcal{A}^{re}) \leq CD_n^{(1)}$,

where the constant $C$ depends only on $b$; while under (E.2), we have

$\rho_n(\mathcal{A}^{re}) \leq C\{D_n^{(1)} + D_n^{(2)}\}$,

where the constant $C$ depends only on $b$ and $q$.

3. High Dimensional CLT for Simple and Sparsely Convex Sets

In this section, we extend the results of Section 2 by considering larger classes of sets; in particular, we consider classes of simple convex sets, and obtain, under certain conditions, bounds that are similar to those in Section 2 (Corollary 3.1). Although an extension to simple convex sets is not difficult, in high dimensional spaces, the class of simple convex sets is rather large. In addition, it allows us to derive similar bounds for the classes of sparsely convex sets. These classes in turn may be of interest in mathematical statistics where sparse models and techniques have been of canonical importance in the past years.

3.1. Simple convex sets. Consider a convex set $A \subset \mathbb{R}^p$. This set can be characterized by its support function:

$\mathcal{S}_A : S^{p-1} \to \mathbb{R} \cup \{\infty\}, \quad v \mapsto \mathcal{S}_A(v) := \sup\{w'v : w \in A\}$,

where $S^{p-1} := \{v \in \mathbb{R}^p : \|v\| = 1\}$; in particular, $A = \cap_{v \in S^{p-1}} \{w \in \mathbb{R}^p : w'v \leq \mathcal{S}_A(v)\}$. We say that a convex set $A$ is $m$-generated if it is generated
by intersections of $m$ half-spaces. The support function $S_A$ of such a set $A$

 can be characterized completely by its values $\{S_A(v) : v \in V(A)\}$ for the set $V(A)$ consisting of $m$ unit vectors that are outward normal to the facets of $A$. Indeed,

$$A = \cap_{v \in V(A)} \{w \in \mathbb{R}^p : w'v \leq S_A(v)\}.$$ 

For $\epsilon > 0$ and an $m$-generated convex set $A^m$, we define

$$A^{m,\epsilon} := \cap_{v \in V(A^m)} \{w \in \mathbb{R}^p : w'v \leq S_{A^m}(v) + \epsilon\},$$

and we say that a convex set $A$ admits an approximation with precision $\epsilon$ by an $m$-generated convex set $A^m$ if

$$A^m \subset A \subset A^{m,\epsilon}.$$ 

Let $a, d > 0$ be some constants. Let $A^s$ be a class of sets $A$ in $\mathbb{R}^p$ that satisfy the following condition:

(C) The set $A$ admits an approximation with precision $\epsilon = a/n$ by an $m$-generated convex set $A^m$ where $m \leq (pn)^d$.

We refer to a set $A$ that satisfies condition (C) as a simple convex set because it can be well approximated by affine transformations of rectangles. Note that any rectangle $A \in A^s$ satisfies condition (C) with $a = 0$ and $d = 1$ (recall that $n \geq 4$). Let $A^m(A)$ denote the set $A^m$ appearing in condition (C) applied to the set $A$.

For every $A \in A^s$ with an approximating $m$-generated set $A^m = A^m(A)$ and $\tilde{X}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{im})' = (v'X_i)_{v \in V(A^m)}$, $i = 1, \ldots, n$, we assume that the following conditions are satisfied:

(M.1') $n^{-1} \sum_{i=1}^n E[\tilde{X}_{ij}^2] \geq b$ for all $j = 1, \ldots, m$,

(M.2') $n^{-1} \sum_{i=1}^n E[|\tilde{X}_{ij}|^{2+k}] \leq B_k^s$ for all $j = 1, \ldots, m$ and $k = 1, 2$.

In addition, we assume that one of the following conditions holds:

(E.1') $E[\exp(|\tilde{X}_{ij}|/B_n)] \leq 2$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$,

(E.2') $E[\max_{1 \leq j \leq m} |\tilde{X}_{ij}|/B_n] \leq 2$ for all $i = 1, \ldots, n$.

Conditions (M.1'), (M.2'), (E.1'), and (E.2') are similar to those used in the previous section but they apply to $\tilde{X}_1, \ldots, \tilde{X}_n$ rather than to $X_1, \ldots, X_n$.

Recall the definition of $\rho_n(A)$ in (1). An extension of Corollary 2.1 leads to the following result in the case where $A = A^s$. Recall the definitions of $D_n^{(1)}$ and $D_n^{(2)}$ given in (9).

**Corollary 3.1 (High Dimensional CLT for Simple Convex Sets).** Let $A^s$ be a class of simple convex sets in $\mathbb{R}^p$ such that conditions (M.1'), (M.2'), and (E.1') are satisfied for every $A \in A^s$. Then

$$\rho_n(A^s) \leq CD_n^{(1)},$$

where the constant $C$ depends only on $a$, $b$, and $d$. If, instead of condition (E.1'), condition (E.2') is satisfied for every $A \in A^s$, then

$$\rho_n(A^s) \leq C\{D_n^{(1)} + D_{n,q}^{(2)}\}.$$ 


where the constant $C$ depends only on $a$, $b$, $d$, and $q$.

It is worthwhile to mention that a notable example where the transformed variables $\tilde{X}_i = (v^\prime X_i)_{v \in V(A^m)}$ verify condition (E.1′) is the case where each $X_i$ obeys a log-concave distribution. Recall that a Borel probability measure $\nu$ on $\mathbb{R}^p$ is log-concave if for every Borel subsets $A_1, A_2$ of $\mathbb{R}^p$ and $\lambda \in (0, 1)$,

$$\nu(\lambda A_1 + (1 - \lambda) A_2) \geq \nu(A_1)^\lambda \nu(A_2)^{1 - \lambda},$$

where $\lambda A_1 + (1 - \lambda) A_2 = \{\lambda x + (1 - \lambda)y : x \in A_1, y \in A_2\}$.

**Corollary 3.2** (High Dimensional CLT for Simple Convex Sets with Log-concave Distributions). Suppose that each $X_i$ obeys a centered log-concave distribution on $\mathbb{R}^p$ and that all the eigenvalues of $E[X_iX_i^\prime]$ are bounded from below by a constant $k_1 > 0$ and from above by a constant $k_2 \geq k_1$ for every $i = 1, \ldots, n$. Then for $A^s$ the class of all simple convex sets in $\mathbb{R}^p$, we have

$$\rho_n(A^s) \leq C n^{-1/6} \log^{7/6}(pn),$$

where the constant $C$ depends only on $a, b, d, k_1$, and $k_2$.

### 3.2. Sparsely convex sets

We next consider classes of sparsely convex sets defined as follows.

**Definition 3.1** (Sparsely convex sets). For integer $s > 0$, we say that $A \subset \mathbb{R}^p$ is an $s$-sparsely convex set if there exist an integer $Q > 0$ and convex sets $A_q \subset \mathbb{R}^p, q = 1, \ldots, Q$, such that $A = \cap_{q=1}^{Q} A_q$, and the indicator function of each $A_q$, $w \mapsto I(w \in A_q)$, depends at most on $s$ elements of its argument $w = (w_1, \ldots, w_p)$ (which we call the main components of $A_q$). We also say that $A = \cap_{q=1}^{Q} A_q$ is a sparse representation of $A$.

Observe that for any $s$-sparsely convex set $A \subset \mathbb{R}^p$, the integer $Q$ in Definition 3.1 can be chosen to satisfy $Q \leq C^p_s \leq p^s$. Indeed, if we have a sparse representation $A = \cap_{q=1}^{Q} A_q$ for $Q > C^p_s$, then there are at least two sets $A_{q_1}$ and $A_{q_2}$ with the same main components, and hence we can replace these two sets by one convex set $A_{q_1} \cap A_{q_2}$ with the same main components; this procedure can be repeated until we have $Q \leq C^p_s$.

**Example 3.1.** The simplest example verifying Definition 3.1 is a rectangle as in (4), which is a 1-sparsely convex set. Another example is the set

$$A = \{w \in \mathbb{R}^p : v_k^\prime w \leq a_k \text{ for all } k = 1, \ldots, m\}$$

for some unit vectors $v_k \in S^{p-1}$ and coefficients $a_k, k = 1, \ldots, m$. If the number of non-zero elements of each $v_k$ does not exceed $s$, this $A$ is an $s$-sparsely convex set. Yet another example is the set

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j \text{ for all } j = 1, \ldots, p \text{ and } w_1^2 + w_2^2 \leq c\}$$

for some coefficients $-\infty \leq a_j \leq b_j \leq \infty$, $j = 1, \ldots, p$, and $0 < c \leq \infty$. This $A$ is a 2-sparsely convex set. A more complicated example is the set

$$A = \{w \in \mathbb{R}^p : a_j \leq w_j \leq b_j, w_k^2 + w_l^2 \leq c_{kl}, \text{ for all } j, k, l = 1, \ldots, p\}$$
for some coefficients $-\infty \leq a_j \leq b_j \leq \infty$, $0 < c_{kl} \leq \infty$, $j,k,l = 1,\ldots,p$. This $A$ is a 2-sparsely convex set. Finally, consider the set

$$A = \{ w \in \mathbb{R}^p : \| (w_j)_{j \in J_k} \|^2 \leq t_k \text{ for all } k = 1, \ldots, \kappa \},$$

where $\{ J_k \}$ are subsets of $\{1, \ldots, p\}$ of fixed cardinality $\kappa$, $\{ t_k \}$ are thresholds in $(0, \infty)$, and $1 \leq \kappa \leq \mathbb{C}_p$ is an integer. This $A$ is an $s$-sparsely convex set.

Fix an integer $s > 0$, and let $\mathcal{A}^{sp}(s)$ denote the class of all $s$-sparsely convex sets in $\mathbb{R}^p$. We assume that the following condition is satisfied:

(M.1$''$) $n^{-1} \sum_{i=1}^n E[(v'X_i)^2] \geq b$ for all $v \in \mathbb{S}^{p-1}$ with $\|v\|_0 \leq s$.

Then we have the following corollary:

**Corollary 3.3 (High Dimensional CLT for Sparsely Convex Sets).** Suppose that conditions (M.1$''$) and (M.2) are satisfied. Then under (E.1), we have

$$\rho_n(\mathcal{A}^{sp}(s)) \leq CD_n^{(1)},$$

where the constant $C$ depends only on $b$ and $s$; while under (E.2), we have

$$\rho_n(\mathcal{A}^{sp}(s)) \leq C\{D_n^{(1)} + D_n^{(2)}\},$$

where the constant $C$ depends only on $b$, $q$, and $s$.

**Remark 3.1 (Dependence on $s$).** In many applications, it may be of interest to consider $s$-sparsely convex sets with $s = s_n$ depending on $n$ and potentially growing to infinity: $s = s_n \to \infty$. It is therefore interesting to derive the optimal dependence of the constant $C$ in (12) and (13) on $s$. We leave this question for future work.

### 4. Empirical and Multiplier Bootstrap Theorems

So far we have shown that the probabilities $P(S_n^X \in A)$ can be well approximated by the Gaussian analog $P(S_n^Y \in A)$ under weak conditions uniformly in rectangles $A \in \mathcal{A}^{re}$, simple convex sets $A \in \mathcal{A}^{si}$, or sparsely convex sets $A \in \mathcal{A}^{sp}(s)$. In practice, however, the covariance matrix of $S_n^Y$ is typically unknown, and direct computation of $P(S_n^Y \in A)$ is infeasible. Therefore, in this section, we derive high dimensional bootstrap theorems which allow us to approximate the probabilities $P(S_n^Y \in A)$ (and hence $P(S_n^X \in A)$) by means of the bootstrap. We consider multiplier and empirical bootstrap methods (for various version of bootstraps, we refer to [30]).

#### 4.1. Multiplier bootstrap

We first consider the multiplier bootstrap. Let $e_1, \ldots, e_n$ be a sequence of i.i.d. $N(0,1)$ random variables that are independent of $X^n = \{X_1, \ldots, X_n\}$. Let $\hat{\mu}_n^X := (\hat{\mu}_{n1}^X, \ldots, \hat{\mu}_{np}^X)' := E_n[X]$, and consider the normalized sum:

$$S_{n}^{eX} := (S_{n1}^{eX}, \ldots, S_{np}^{eX})' := \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i (X_i - \hat{\mu}_n^X).$$
We are interested in bounding
\[ \rho_n^{MB}(A) := \sup_{A \in \mathcal{A}} |P(S_n^c \in A | X_1^n) - P(S_n^c \in A)| \]
in the cases where \( A = A^{re}, A^{ai}, \) or \( A^{sp}(s) \).

We begin with the case where \( A = A^{ai} \). Let
\[
\Sigma^c := n^{-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_n^X)(X_i - \hat{\mu}_n^X)', \quad \Sigma^Y := n^{-1} \sum_{i=1}^{n} E[Y_iY_i'].
\]
Observe that \( E[S_n^c(X_n^c) | X_1^n] = \Sigma^c \) and \( E[S_n^c(Y_n^c) | X_1^n] = \Sigma^Y \). Define
\[
\Delta_n := \sup_{A \in \mathcal{A}^i} \max_{v_1, v_2 \in V(A^m(A))} |v_1'(\Sigma^c - \Sigma^Y)v_2|.
\]

Then we have the following theorem for classes of simple convex sets.

**Theorem 4.1** (Abstract Multiplier Bootstrap Theorem for Simple Convex Sets). Suppose that condition (M.1') is satisfied for every \( A \in \mathcal{A}^{ai} \). Then for every constant \( \Delta_n > 0 \), on the event \( \Delta_n \leq \Delta_n \), we have
\[
\rho_n^{MB}(A^{ai}) \leq C \left\{ \Delta_n^{1/3} \log^{2/3}(pn) + n^{-1} \log^{1/2}(pn) \right\},
\]
where the constant \( C \) depends only on \( a, b, \) and \( d \).

**Remark 4.1** (Case of rectangles). From the proof of Theorem 4.1, we have the following bound when \( A = A^{re} \): under (M.1), for every constant \( \Delta_n > 0 \), on the event \( \Delta_n \leq \Delta_n \), we have
\[
\rho_n^{MB}(A^{re}) \leq C \Delta_n^{1/3} \log^{2/3}(pn),
\]
where the constant \( C \) depends only on \( b \), and \( \Delta_n \) is defined by
\[
\Delta_n = \max_{1 \leq j, k \leq p} |\Sigma^c_{jk} - \Sigma^Y_{jk}|,
\]
where \( \Sigma^c_{jk} \) and \( \Sigma^Y_{jk} \) are the \((j, k)\)th elements of \( \Sigma^c \) and \( \Sigma^Y \), respectively. \( \blacksquare \)

We shall derive more explicit bounds on \( \rho_n^{MB}(A^{ai}) \) under suitable moment conditions as in the previous section. We will need to strengthen condition (C) and will assume that all sets \( A \) in \( \mathcal{A}^{ai} \) satisfy the following condition:

**C’** The set \( A \) admits an approximation with precision \( \epsilon = a/n \) by an \( m \)-generated convex set \( A^m \) where \( m \leq (pn)^d \) and \( A^m \) is such that for \( v \in V(A^m) \), \( ||v||_0 \leq s \).

Note that condition (C’) is more restrictive than (C) as it requires that the outward unit normal vectors to the hyperplanes forming the \( m \)-generated convex set \( A^m \) are sparse. We need this extra condition to control \( \Delta_n \). Then we have the following corollary. Here for \( \alpha \in (0, 1) \), define
\[
D^{(1)}_n(\alpha) = \left( \frac{B^2_2 (\log^5(pn)) \log^{2/(1-\alpha)}(1)}{n} \right)^{1/6}, \quad D^{(2)}_{n, q}(\alpha) = \left( \frac{B^2_2 \log^3 p}{\alpha^{2/q} n^{1-2/q}} \right)^{1/3}.
\]
Corollary 4.1 (Multiplier Bootstrap for Simple Convex Sets). Let $\alpha \in (0, e^{-1})$ be a constant. Suppose that conditions $(C')$ and $(M.1')$ are satisfied for every $A \in \mathcal{A}^{si}$. In addition, suppose that condition $(M.2)$ is satisfied. Then under $(E.1)$, we have with probability at least $1 - \alpha$, 

$$\rho_n^{MB}(\mathcal{A}^{si}) \leq C D_n^{(1)}(\alpha),$$

where the constant $C$ depends only on $a, b, d$ and $s$; while under $(E.2)$, we have with probability at least $1 - \alpha$,

$$\rho_n^{MB}(\mathcal{A}^{si}) \leq C \{ D_n^{(1)}(\alpha) + D_{n,q}^{(1)}(\alpha) \},$$

where the constant $C$ depends only on $a, b, d, q$, and $s$.

When each $X_i$ obeys a log-concave distribution, then we have the following corollary analogous to Corollary 3.2. In this case, in stead of condition $(C')$, we will make an alternative assumption that the cardinality of $\bigcup_{A \in \mathcal{A}^{si}} V(\mathcal{A}^{si})$ for each $A$ is at most $(pn)^d$.

Corollary 4.2 (Multiplier Bootstrap for Simple Convex Sets with Log-concave Distributions). Let $\alpha \in (0, e^{-1})$ be a constant. Suppose that each $X_i$ obeys a centered log-concave distribution on $\mathbb{R}^p$ and that all the eigenvalues of $E[X_i X_i']$ are bounded from below by a constant $k_1 > 0$ and from above by a constant $k_2 \geq k_1$ for all $i = 1, \ldots, n$. Moreover, suppose that every $A \in \mathcal{A}^{si}$ satisfies, in addition to condition $(C)$, that the cardinality of the set $\bigcup_{A \in \mathcal{A}^{si}} V(\mathcal{A}^{si}(A))$ is at most $(pn)^d$. Then with probability at least $1 - \alpha$,

$$\rho_n^{MB}(\mathcal{A}^{si}) \leq C n^{-1/6} (\log^{5/6}(pn)) \log^{1/3}(1/\alpha),$$

where the constant $C$ depends only on $a, d, k_1$, and $k_2$.

Finally we shall derive explicit bounds on $\rho_n^{MB}(A)$ in the case where $A$ is the class of all $s$-sparsely convex sets: $A = \mathcal{A}^{sp}(s)$.

Corollary 4.3 (Multiplier Bootstrap for Sparsely Convex Sets). Let $\alpha \in (0, e^{-1})$ be a constant. Suppose that conditions $(M.1''$) and $(M.2)$ are satisfied. Then under $(E.1)$, we have with probability at least $1 - \alpha$, 

$$\rho_n^{MB}(\mathcal{A}^{sp}(s)) \leq C D_n^{(1)}(\alpha),$$

where the constant $C$ depends only on $b$ and $s$; while under $(E.2)$, we have with probability at least $1 - \alpha$,

$$\rho_n^{MB}(\mathcal{A}^{sp}(s)) \leq C \{ D_n^{(1)}(\alpha) + D_{n,q}^{(2)}(\alpha) \},$$

where the constant $C$ depends only on $b, s$, and $q$.

4.2. Empirical bootstrap. Here we consider the empirical bootstrap. For brevity, we shall focus here on the cases where $A$ is $\mathcal{A}^{re}$ or $\mathcal{A}^{si}$. Let $X_1^*, \ldots, X_n^*$ be i.i.d. draws from the empirical distribution of $X_1, \ldots, X_n$. Conditional
on $X^n = \{X_1, \ldots, X_n\}$, $X_1^*, \ldots, X_n^*$ are i.i.d. with mean $\hat{\mu}_n = \mathbb{E}[X_i]$. Consider the normalized sum:

$$S_n^X := (S_{n1}^X, \ldots, S_{np}^X)' := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i^* - \hat{\mu}_n).$$

We are interested in bounding

$$\rho_n^{EB}(A) := \sup_{A \in \mathcal{A}} \mathbb{P}(S_n^X \in A \mid X_i^n) - \mathbb{P}(S_n^Y \in A)$$

in the cases where $A = \mathcal{A}^{re}$ or $\mathcal{A}^{si}$. To state the bound, define

$$\hat{L}_n := \max_{1 \leq j \leq p} \sum_{i=1}^{n} |X_{ij} - \hat{\mu}_{nj}| / n,$$

which is an empirical analog of $L_n$, and for $\phi \geq 1$, define

$$\hat{M}_{n, X}(\phi) := n^{-1} \sum_{i=1}^{n} \max_{1 \leq j \leq p} |X_{ij} - \hat{\mu}_{nj}|^3 \left\{ \max_{1 \leq j \leq p} |X_{ij} - \hat{\mu}_{nj}| > \sqrt{n} / (4\phi \log p) \right\},$$

$$\hat{M}_{n, Y}(\phi) := \mathbb{E} \left[ \max_{1 \leq j \leq p} |S_{nj}^X|^3 \left\{ \max_{1 \leq j \leq p} |S_{nj}^X| > \sqrt{n} / (4\phi \log p) \right\} \mid X_i^n \right],$$

which are empirical analogs of $M_{n, X}(\phi)$ and $M_{n, Y}(\phi)$, respectively. Let

$$\hat{M}_n(\phi) := \hat{M}_{n, X}(\phi) + \hat{M}_{n, Y}(\phi).$$

Then we have the following theorem for the class of rectangles $A = \mathcal{A}^{re}$.

**Theorem 4.2 (Abstract Empirical Bootstrap Theorem).** For arbitrary positive constants $b$, $L_n$, and $\bar{M}_n$, the inequality

$$\rho_n^{EB}(\mathcal{A}^{re}) \leq \rho_n^{MB}(\mathcal{A}^{re}) + K_1 \left[ \left( \frac{T_n^2 \log^7 (pn)}{n} \right)^{1/6} + \frac{\bar{M}_n}{\hat{L}_n} \right]$$

holds on the event

$$\{\mathbb{E}[|X_{ij} - \hat{\mu}_{nj}|^2] \geq b \text{ for all } j = 1, \ldots, p \} \cap \{\hat{L}_n \leq L_n\} \cap \{\hat{M}_n \leq \bar{M}_n\},$$

where

$$\phi_n := K_2 \left( \frac{T_n^2 \log^4 p}{n} \right)^{-1/6}.$$

Here $K_1, K_2 > 0$ are constants that depend only on $b$.

As in the multiplier bootstrap, we shall derive explicit bounds on $\rho_n^{EB}(A)$ under suitable moment conditions. Here we only state the results for classes of simple convex sets $A = \mathcal{A}^{si}$ but note that the same result applies to the case of rectangles since a rectangle is a special case of a simple convex set.
Corollary 4.4 (Empirical Bootstrap for Simple Convex Sets). Let $\alpha \in (0, e^{-1})$ be a constant and suppose that $\log(1/\alpha) \leq K \log(pn)$ for some other constant $K$. Moreover, suppose that all the assumptions in Corollary 4.1 except for (E.1) and (E.2) are satisfied. Then under (E.1), we have with probability at least $1 - \alpha$,

$$\rho_n^{EB}(A^{si}) \leq CD_n^{(1)},$$

where the constant $C$ depends only on $a, b, d, s,$ and $K$; while under (E.2), we have with probability at least $1 - \alpha$,

$$\rho_n^{EB}(A^{si}) \leq C\{D_n^{(1)} + D_n^{(2)}(\alpha)\},$$

where the constant $C$ depends only on $a, b, d, s, q,$ and $K$.

When each $X_i$ obeys a log-concave distribution, then we have the following corollary.

Corollary 4.5 (Empirical Bootstrap for Simple Convex Sets with Log-concave Distributions). Let $\alpha \in (0, e^{-1})$ be a constant and suppose that $\log(1/\alpha) \leq K \log(pn)$ for some other constant $K$. Moreover, suppose that all the assumptions in Corollary 4.2 are satisfied. Then with probability at least $1 - \alpha$,

$$\rho_n^{EB}(A^{si}) \leq Cn^{-1/6} \log^{7/6}(pn),$$

where the constant $C$ depends only on $a, d, k_1, k_2,$ and $K$.

Remark 4.2 (Bootstrap CLTs in a.s. sense). Corollaries 4.1 and 4.4 lead to the following multiplier and empirical bootstrap CLTs in the a.s. sense. Suppose that all the assumptions in Corollary 4.1 except for (E.1) and (E.2) are satisfied. We allow $p = p_n \to \infty$ and $B_n \to \infty$ as $n \to \infty$ but assume that $a, b, d, q, s$ are all fixed. Then by applying Corollaries 4.1 and 4.4 with $\alpha = \alpha_n = n^{-1} \log^{-2} n$, together with the Borel-Cantelli lemma (note that $\sum_{n=1}^{\infty} n^{-1} \log^{-2} n < \infty$), we have with probability one

$$\rho_n^{MB}(A^{si}) = \rho_n^{EB}(A^{si}) \leq \begin{cases} O(D_n^{(1)}) & \text{under (E.1)} \\ O(D_n^{(1)} \lor D_n^{(2)}(\alpha_n)) & \text{under (E.2)} \end{cases},$$

and it is routine to verify that $D_n^{(1)} = o(1)$ if $B_n^2 \log^7(pn) = o(n)$, and $D_n^{(2)}(\alpha_n) = o(1)$ if $B_n^2(\log^3(pn)) \log^{1/4} n = o(n^{1 - 4/q})$.

5. Induction Lemma

In this section, we state a lemma that plays a key role in the proof of our high dimensional CLT for rectangles (Theorem 2.1). Define

$$\varrho_n := \sup_{y \in \mathbb{R}^p, \nu \in [0, 1]} \left| P \left( \sqrt{v} S_n^X + \sqrt{1 - v} S_n^Y \leq y \right) - P(S_n^Y \leq y) \right|,$$

and recall that $M_n(\phi) := M_n,X(\phi) + M_n,Y(\phi)$ for $\phi \geq 1$. The lemma below provides a bound on $\varrho_n$. 
Lemma 5.1 (Induction Lemma). Suppose that there exists some constant $b > 0$ such that $n^{-1} \sum_{i=1}^{n} E[X_{ij}^2] \geq b$ for all $j = 1, \ldots, p$. Then $\varrho_n$ satisfies the following inequality for all $\phi \geq 1$:

$$\varrho_n \lesssim \frac{\phi^2 \log^2 p}{n^{1/2}} \left( \phi L_n \varrho_n + L_n \log^{1/2} p + \phi M_n(\phi) \right) + \frac{\log^{1/2} p}{\phi}$$

up to a constant $K$ that depends only on $b$.

Lemma 5.1 has an immediate corollary. Indeed, define

$$\varrho'_n := \sup_{A \in \mathcal{A}_n, v \in [0,1]} \left| P(\sqrt{v} S_n^X + \sqrt{1-v} S_n^Y \in A) - P(S_n^Y \in A) \right|$$

where $\mathcal{A}_n$ is the class of all rectangles in $\mathbb{R}^p$. Then we have:

Corollary 5.1. Suppose that there exists some constant $b > 0$ such that $n^{-1} \sum_{i=1}^{n} E[X_{ij}^2] \geq b$ for all $j = 1, \ldots, p$. Then $\varrho'_n$ satisfies the following inequality for all $\phi \geq 1$:

$$\varrho'_n \lesssim \frac{\phi^2 \log^2 p}{n^{1/2}} \left( \phi L_n \varrho'_n + L_n \log^{1/2} p + \phi M_n(2\phi) \right) + \frac{\log^{1/2} p}{\phi}$$

up to a constant $K'$ that depends only on $b$.

**Appendix A. Anti-concentration inequalities**

One of the main ingredients of the proof of Lemma 5.1 (and the proofs of the other results indeed) is the following anti-concentration due to [28].

Lemma A.1 (Nazarov’s inequality, [28]). Let $Y = (Y_1, \ldots, Y_p)'$ be a centered Gaussian random vector in $\mathbb{R}^p$ such that $E[Y_{ij}^2] \geq b$ for all $j = 1, \ldots, p$ and some constant $b > 0$. Then for every $y \in \mathbb{R}^p$ and $a > 0$,

$$P(Y \leq y+a) - P(Y \leq y) \leq C a(\log p)^{1/2},$$

where $C$ is a constant depending only on $b$.

**Remark A.1.** This inequality is less sharp than the dimension-free anti-concentration bound $C a E[\max_{1 \leq j \leq p} Y_j]$ proved in [18] for the case of max rectangles. However, the former inequality allows for more general rectangles than the latter. The difference in sharpness for the case of max-rectangles arises due to dimension-dependence $(\log p)^{1/2}$, in particular the term $(\log p)^{1/2}$ can be much larger than $E[\max_{1 \leq j \leq p} Y_j]$. This also makes the anti-concentration bound in [18] more relevant for the study of suprema of Gaussian processes indexed by infinite classes. It is an interesting question whether one could establish a dimension-free anti-concentration bound similar to that in [18] for classes of rectangles other than max rectangles. ■

**Proof of Lemma A.1.** Let $\Sigma = E[YY']$; then $Y$ has the same distribution as $\Sigma^{1/2} Z$ where $Z$ is a standard Gaussian random vector. Write $\Sigma^{1/2} = \ldots$
$(\sigma_1, \ldots, \sigma_p)'$ where each $\sigma_j$ is a $p$-dimensional vector. Note that $\|\sigma_j\| = (E[Y_j^2])^{1/2} \geq b^{1/2}$. Then

$$P(Y \leq y + a) = P(\Sigma^{1/2} Z \leq y + a)$$

$$= P((\sigma_j/\|\sigma_j\|)' Z \leq (y_j + a)/\|\sigma_j\| \text{ for all } j = 1, \ldots, p),$$

and similarly

$$P(Y \leq y) = P((\sigma_j/\|\sigma_j\|)' Z \leq y_j/\|\sigma_j\| \text{ for all } j = 1, \ldots, p).$$

Since $Z$ is a standard Gaussian random vector, and $a/\|\sigma_j\| \leq a/b^{1/2}$ for all $j = 1, \ldots, p$, the assertion follows from Theorem 20 in [23], whose proof the authors credit to Nazarov [28].

We will use another anti-concentration inequality by [28] in the proofs for Section 3, which is an extension of Theorem 4 in [3].

**Lemma A.2.** Let $A$ be a $p \times p$ symmetric positive definite matrix, and let $A \sim N(0, A^{-1})$. Then there exists a universal constant $C > 0$ such that for every convex set $Q \subset \mathbb{R}^p$,

$$\limsup_{h \to 0} \frac{\gamma_A(Q^h \setminus Q)}{h} \leq C \sqrt{\|A\|_{HS}},$$

where $\|A\|_{HS}$ is the Hilbert-Schmidt norm of $A$.

**Proof.** See [28].

**Appendix B. Proof of Lemma 5**

We begin with stating the following variants of Chebyshev’s association inequality.

**Lemma B.1.** Let $\varphi_i : \mathbb{R} \to [0, \infty)$, $i = 1, 2$ be non-increasing functions, and let $\xi_i, i = 1, 2$ be independent real-valued random variables. Then

$$E[\varphi_1(\xi_1)] E[\varphi_2(\xi_2)] \leq E[\varphi_1(\xi_1) \varphi_2(\xi_1)], \quad (18)$$

$$E[\varphi_1(\xi_1)] E[\varphi_2(\xi_2)] \leq E[\varphi_1(\xi_1) \varphi_2(\xi_1)] + E[\varphi_1(\xi_2) \varphi_2(\xi_2)], \quad (19)$$

$$E[\varphi_1(\xi_1) \varphi_2(\xi_2)] \leq E[\varphi_1(\xi_1) \varphi_2(\xi_1)] + E[\varphi_1(\xi_2) \varphi_2(\xi_2)], \quad (20)$$

where we assume that all the expectations exist and are finite. Moreover, (20) holds without independence of $\xi_1$ and $\xi_2$.

**Proof of Lemma B.1.** The inequality (18) is Chebyshev’s association inequality; see Theorem 2.14 in [11]. Moreover, since $\xi_1$ and $\xi_2$ are independent, (19) follows from (20). In turn, (20) follows from

$$E[\varphi_1(\xi_1) \varphi_2(\xi_2)] \leq E[\varphi_1(\xi_1) \varphi_2(\xi_2)] + E[\varphi_2(\xi_1) \varphi_1(\xi_2)]$$

$$\leq E[\varphi_1(\xi_1) \varphi_2(\xi_1)] + E[\varphi_1(\xi_2) \varphi_2(\xi_2)].$$
where the first inequality follows from the fact that \( \varphi_2(\xi_1) \varphi_1(\xi_2) \geq 0 \), and the second inequality follows from rearranging the terms in the following inequality:

\[
E[(\varphi_1(\xi_1) - \varphi_1(\xi_2))(\varphi_2(\xi_1) - \varphi_2(\xi_2))] \geq 0,
\]

which follows from monotonicity of \( \varphi_1 \) and \( \varphi_2 \).

Proof of Lemma 5.1. The proof relies on a Slepian-Stein method developed in [15]. Here the notation \( \lesssim \) means that the left-hand side is bounded by the right hand side up to some constant depending only on \( b \).

We begin with preparing some notation. Let \( W_1, \ldots, W_n \) be a copy of \( Y_1, \ldots, Y_n \). Without loss of generality, we may assume that \( X_1, \ldots, X_n, Y_1, \ldots, Y_n \), and \( W_1, \ldots, W_n \) are independent. Consider

\[
S_n^W := \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i.
\]

Then \( P(S_n^Y \leq y) = P(S_n^W \leq y) \), so that

\[
\varrho_n = \sup_{s \in \mathbb{R}^p, v \in [0,1]} |P(\sqrt{v} S_n^X + \sqrt{1-v} S_n^Y \leq y) - P(S_n^W \leq y)|.
\]

Pick any \( y \in \mathbb{R}^p \) and \( v \in [0,1] \). Let \( \beta := \phi \log p \), and define the function

\[
F_\beta(w) := \frac{1}{\beta} \log \left( \sum_{j=1}^p \exp \left( \beta (w_j - y_j) \right) \right), \quad w \in \mathbb{R}^p.
\]

The function \( F_\beta(w) \) has the following property:

\[
0 \leq F_\beta(w) - \max_{1 \leq j \leq p} (w_j - y_j) \leq \beta^{-1} \log p = \phi^{-1}, \quad \text{for all } w \in \mathbb{R}^p. \quad (21)
\]

Consider a thrice continuously differentiable function \( g_0 : \mathbb{R} \to [0,1] \) whose derivatives up to the third order are all bounded such that \( g_0(t) = 1 \) for \( t \leq 0 \) and \( g_0(t) = 0 \) for \( t \geq 1 \). Define \( g(t) := g_0(\phi t), t \in \mathbb{R} \), and

\[
m(w) := g(F_\beta(w)), \quad w \in \mathbb{R}^p.
\]

For brevity of notation, we will use indices to denote partial derivatives of \( m \); for example, \( \partial_j \partial_k \partial_l m = m_{jkl} \). The function \( m(w) \) has the following property established in Lemmas A.5 and A.6 of [15]: for every \( j, k, l = 1, \ldots, p \), there exists a function \( U_{jkl}(w) \) such that

\[
|m_{jkl}(w)| \leq U_{jkl}(w), \quad (22)
\]

\[
\sum_{j,k,l=1}^p U_{jkl}(w) \leq (\phi^3 + \phi \beta + \phi \beta^2) \leq \phi \beta^2, \quad (23)
\]

\[
U_{jkl}(w) \lesssim U_{jkl}(w + \tilde{w}) \leq U_{jkl}(w), \quad (24)
\]

where the inequalities (22) and (23) hold for all \( w \in \mathbb{R}^p \), and the inequality (24) holds for all \( w, \tilde{w} \in \mathbb{R}^p \) with \( \max_{1 \leq j \leq p} |\tilde{w}_j| \beta \leq 1 \) (formally, [15] only
considered the case where $y = (0, \ldots, 0)'$ but the extension to $y \in \mathbb{R}^p$ is trivial). Moreover, define the functions
\[
h(w, t) := \mathbb{1}\left\{ -\theta^{-1} - t/\beta < \max_{1 \leq j \leq p} (w_j - y_j) \leq \theta^{-1} + t/\beta \right\}, \quad w \in \mathbb{R}^p, t > 0, \tag{25}
\]
\[
\omega(t) := \frac{1}{\sqrt{t} \wedge \sqrt{1-t}}, \quad t \in (0, 1).
\]

The proof consists of two steps. In the first step, we show that
\[
|E[I_n]| \lesssim \frac{\phi^2 \log^2 p}{n^{1/2}} \left( \phi L_n \rho_n + L_n \log^{1/2} p + \phi M_n(\phi) \right) \tag{26}
\]
where
\[
I_n := m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W).
\]
The second step combines this bound with Lemma A.1 to complete the proof.

**Step 1.** Define the Slepian interpolant
\[
Z(t) := \sum_{i=1}^n Z_i(t), \quad t \in [0, 1],
\]
where
\[
Z_i(t) := \frac{1}{\sqrt{n}} \left\{ \sqrt{t}(\sqrt{v}X_i + \sqrt{1-v}Y_i) + \sqrt{1-t}W_i \right\}.
\]
Note that $Z(1) = \sqrt{v}S_n^X + \sqrt{1-v}S_n^Y$ and $Z(0) = S_n^W$, and so
\[
I_n = m(\sqrt{v}S_n^X + \sqrt{1-v}S_n^Y) - m(S_n^W) = \int_0^1 \frac{dm(Z(t))}{dt} dt. \tag{27}
\]
Denote by $Z^{(i)}(t)$ the Stein leave-one-out term for $Z(t)$:
\[
Z^{(i)}(t) := Z(t) - Z_i(t).
\]
Finally, define
\[
\dot{Z}_i(t) := \frac{1}{\sqrt{n}} \left\{ \frac{1}{\sqrt{t}}(\sqrt{v}X_i + \sqrt{1-v}Y_i) - \frac{1}{\sqrt{1-t}}W_i \right\}.
\]
For brevity of notation, we omit the argument $t$; that is, we write $Z = Z(t)$, $Z_i = Z_i(t)$, $Z^{(i)} = Z^{(i)}(t)$, and $\dot{Z}_i = \dot{Z}_i(t)$.

Now, from (27) and Taylor’s theorem, we have
\[
E[I_n] = \frac{1}{2} \sum_{j=1}^p \sum_{i=1}^n \int_0^1 E[m_j(Z)\dot{Z}_{ij}] dt = \frac{1}{2} (I + II + III),
\]
where

\[ I := \sum_{j=1}^{p} \sum_{i=1}^{n} \int_0^1 E[m_j(Z^{(i)}) \hat{Z}_{ij}] dt, \]

\[ II := \sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 E[m_{jk}(Z^{(i)}) \hat{Z}_{ij} \hat{Z}_{ik}] dt, \]

\[ III := \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \int_0^1 (1 - \tau)E[m_{jkl}(Z^{(i)}) + \tau Z_{i}] \hat{Z}_{ij} \hat{Z}_{ik} \hat{Z}_{il}] d\tau dt. \]

By independence of \( Z^{(i)} \) from \( \hat{Z}_{ij} \) together with \( E[\hat{Z}_{ij}] = 0 \), we have \( I = 0 \).

Also, by independence of \( Z^{(i)} \) from \( \hat{Z}_{ij} \hat{Z}_{ik} \) together with

\[ E[\hat{Z}_{ij} \hat{Z}_{ik}] = \frac{1}{n} E[(\sqrt{v}X_{ij} + \sqrt{1-v}Y_{ij})(\sqrt{v}X_{ik} + \sqrt{1-v}Y_{ik}) - W_{ij}W_{ik}] \]

\[ = \frac{1}{n} E[vX_{ij}X_{ik} + (1-v)Y_{ij}Y_{ik} - W_{ij}W_{ik}] = 0, \]

we have \( II = 0 \). Therefore, it suffices to bound \( III \).

To this end, let

\[ \chi_i = \{ \max_{1 \leq j \leq p} |X_{ij}| \vee |Y_{ij}| \vee |W_{ij}| \leq \sqrt{n}/(4\beta) \}, \]

and decompose \( III \) as \( III = III_1 + III_2 \), where

\[ III_1 := \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \int_0^1 (1 - \tau)E[\chi_i m_{jkl}(Z^{(i)}) + \tau Z_{i}] \hat{Z}_{ij} \hat{Z}_{ik} \hat{Z}_{il}] d\tau dt, \]

\[ III_2 := \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \int_0^1 (1 - \tau)E[(1 - \chi_i) m_{jkl}(Z^{(i)}) + \tau Z_{i}] \hat{Z}_{ij} \hat{Z}_{ik} \hat{Z}_{il}] d\tau dt. \]

We shall bound \( III_1 \) and \( III_2 \) separately. For \( III_2 \), we have

\[ |III_2| \leq \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_0^1 \int_0^1 E[(1 - \chi_i) U_{jkl}(Z^{(i)}) + \tau Z_{i}] |\hat{Z}_{ij} \hat{Z}_{ik} \hat{Z}_{il}] d\tau dt \]

\[ \lesssim \phi \beta^2 \sum_{i=1}^{n} \int_0^1 \int_0^1 E[(1 - \chi_i) \max_{1 \leq j,k,l \leq p} |\hat{Z}_{ij} \hat{Z}_{ik} \hat{Z}_{il}] d\tau dt \]

\[ \lesssim \phi \beta^2 \frac{1}{\eta^{3/2}} \sum_{i=1}^{n} \int_0^1 \int_0^1 \omega(t) E[(1 - \chi_i) \max_{1 \leq j \leq p} |X_{ij}|^3 \vee |Y_{ij}|^3 \vee |W_{ij}|^3] dt, \quad (28) \]

where the first and the second inequalities follow from (22) and (23), respectively. Moreover, by letting \( \mathcal{T} = \sqrt{n}/(4\beta) \) and using the union bound, we have

\[ 1 - \chi_i \leq 1 \left\{ \max_{1 \leq j \leq p} |X_{ij}| > \mathcal{T} \right\} + 1 \left\{ \max_{1 \leq j \leq p} |Y_{ij}| > \mathcal{T} \right\} + 1 \left\{ \max_{1 \leq j \leq p} |W_{ij}| > \mathcal{T} \right\}. \]
Hence, using the inequality
\[
\max_{1 \leq j \leq p} |X_{ij}|^3 \vee |Y_{ij}|^3 \vee |W_{ij}|^3 \leq \max_{1 \leq j \leq p} |X_{ij}|^3 + \max_{1 \leq j \leq p} |Y_{ij}|^3 + \max_{1 \leq j \leq p} |W_{ij}|^3
\]
together with the inequality \((20)\) in Lemma B.1, we conclude that the integral in \((28)\) is bounded from above up to a universal constant by
\[
E \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \left\{ \max_{1 \leq j \leq p} |X_{ij}| > T \right\} \right] + E \left[ \max_{1 \leq j \leq p} |Y_{ij}|^3 \left\{ \max_{1 \leq j \leq p} |Y_{ij}| > T \right\} \right]
\]
since \(W_i\)'s have the same distribution as that of \(Y_i\)’s. Therefore,
\[
|III_2| \lesssim (M_n, X(\phi) + M_n, Y(\phi)) \phi \beta^2 / n^{1/2} = M_n(\phi) \phi \beta^2 / n^{1/2}.
\]
To bound \(III_1\), recall the definition of \(h(w, t)\) in \((25)\). Note that \(m_{jkl}(Z^{(i)} + \tau Z_i) = 0\) for all \(\tau \in [0, 1]\) whenever \(h(Z^{(i)}, 2) = 0\) and \(\chi_i = 1\). Indeed if \(\chi_i = 1\), then \(\max_{1 \leq j \leq p} |Z_{ij}| \leq 3/(4\beta) \leq 1/\beta\), and so when \(h(Z^{(i)}, 2) = 0\) and \(\chi_i = 1\), we have \(h(Z^{(i)} + \tau Z_i, 0) = 0\), which in turn implies that either \(F_\beta(Z^{(i)} + \tau Z_i) \leq 0\) or \(F_\beta(Z^{(i)} + \tau Z_i) \geq \phi^{-1}\) because of \((21)\); in both cases, the assertion follows from the definitions of \(m\) and \(g\). Hence
\[
|III_1| \leq \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} \int_{0}^{1} E[\chi_i |m_{jkl}(Z^{(i)} + \tau Z_i) \dot{Z}_{ij} Z_{ik} Z_{il}|] d\tau dt
\]
\[
\lesssim \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} \int_{0}^{1} E[\chi_i h(Z^{(i)}, 2) U_{jkl}(Z^{(i)} + \tau Z_i) |\dot{Z}_{ij} Z_{ik} Z_{il}|] d\tau dt
\]
\[
\lesssim \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} \int_{0}^{1} E[\chi_i h(Z^{(i)}, 2) U_{jkl}(Z^{(i)}) |\dot{Z}_{ij} Z_{ik} Z_{il}|] d\tau dt
\]
\[
\lesssim \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} E[h(Z^{(i)}, 2) U_{jkl}(Z^{(i)})] E[|\dot{Z}_{ij} Z_{ik} Z_{il}|] dt,
\]
where the second inequality follows from \((22)\), the third inequality from \((24)\), and the fourth inequality from the independence of \(Z^{(i)}\) from \(\dot{Z}_{ij} Z_{ik} Z_{il}\). Then we split the integral in \((29)\) by inserting \(\chi_i + (1 - \chi_i)\) under the first expectation sign. We have
\[
\sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} E[(1 - \chi_i) h(Z^{(i)}, 2) U_{jkl}(Z^{(i)})] E[|\dot{Z}_{ij} Z_{ik} Z_{il}|] dt
\]
\[
\lesssim \phi \beta^2 \sum_{i=1}^{n} \int_{0}^{1} E[1 - \chi_i] E \left[ \max_{1 \leq j,k,l \leq p} |\dot{Z}_{ij} Z_{ik} Z_{il}| \right] dt \lesssim M_n(\phi) \phi \beta^2 / n^{1/2},
\]
where the last inequality follows from the argument similar to that used to bound \(III_2\) with applying \((18)\) and \((19)\) instead of \((20)\) in Lemma B.1. Moreover, since \(h(Z^{(i)}, 2) = 0\) whenever \(h(Z, 4) = 0\) and \(\chi_i = 1\) (which
follows from the same argument as before), we have

\[
\sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} E[\chi_i h(Z^{(i)}, 2) U_{jkl}(Z^{(i)})] E[|\bar{Z}_{ij} Z_{ik} Z_{il}|] dt \\
\lesssim \sum_{j,k,l=1}^{p} \sum_{i=1}^{n} \int_{0}^{1} E[h(Z, 4) U_{jkl}(Z)] E[|\bar{Z}_{ij} Z_{ik} Z_{il}|] dt \\
= \sum_{j,k,l=1}^{p} \int_{0}^{1} E[h(Z, 4) U_{jkl}(Z)] \sum_{i=1}^{n} E[|\bar{Z}_{ij} Z_{ik} Z_{il}|] dt \\
\lesssim \phi \beta^2 \int_{0}^{1} E[h(Z, 4)] \max_{1 \leq j,k,l \leq p} \sum_{i=1}^{n} E[|\bar{Z}_{ij} Z_{ik} Z_{il}|] dt. \quad (30)
\]

To bound (30), observe that

\[
|\bar{Z}_{ij} Z_{ik} Z_{il}| \lesssim \frac{\omega(t)}{n^{3/2}} (|X_{ij}|^3 + |Y_{ij}|^3 + |W_{ij}|^3),
\]

which, together with the facts that \(E[|W_{ij}|^3] = E[|Y_{ij}|^3]\) and \(E[|Y_{ij}|^3] \lesssim (E[|Y_{ij}|^2])^{3/2} = (E[|X_{ij}|^2])^{3/2} \leq E[|X_{ij}|^3]\), implies that

\[
\max_{1 \leq j,k,l \leq p} \sum_{i=1}^{n} E[|\bar{Z}_{ij} Z_{ik} Z_{il}|] \lesssim \frac{\omega(t)}{n^{3/2}} \max_{1 \leq j,k,l \leq p} \sum_{i=1}^{n} (E[|X_{ij}|^3] + E[|Y_{ij}|^3]) \lesssim \frac{\omega(t)}{n^{1/2}} L_n.
\]

Meanwhile, observe that

\[
E[h(Z, 4)] = P(\bar{V}_n \leq \bar{T}) - P(\bar{V}_n \leq \bar{I}),
\]

where

\[
\bar{V}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\sqrt{tvX_i} + \sqrt{1-t}Y_i + \sqrt{\bar{I}}W_i)
\]

and \(\bar{I} = y - \phi^{-1} - 4\beta^{-1}, \bar{T} = y + \phi^{-1} + 4\beta^{-1}\); here the notation \(\overset{d}{=}\) denotes equality in distribution, and \(\bar{I}\) and \(\bar{T}\) are vectors in \(\mathbb{R}^p\) (recall the rules of summation of vectors and scalars defined in Section 1.1). Now by the definition of \(\varrho_n\),

\[
P(\bar{V}_n \leq \bar{T}) \leq P(S_n^Y \leq \bar{T}) + \varrho_n, \quad P(\bar{V}_n \leq \bar{I}) \geq P(S_n^Y \leq \bar{I}) - \varrho_n,
\]

and by Lemma A.1,

\[
P(S_n^Y \leq \bar{T}) - P(S_n^Y \leq \bar{I}) \lesssim \phi^{-1} \log^{1/2} p
\]

since \(\beta^{-1} \lesssim \phi^{-1}\) and \(E[(S_n^Y)^2] = E[(S_n^Y)^2] = n^{-1} \sum_{i=1}^{n} E[X_{ij}^2] \geq b\) for all \(j = 1, \ldots, p\). Hence

\[
E[h(Z, 4)] \lesssim \varrho_n + \phi^{-1} \log^{1/2} p.
\]
By these bounds, together with the fact that \( \int_0^1 \omega(t)dt \lesssim 1 \), we conclude that

\[
(30) \lesssim \frac{\phi \beta^2 L_n}{n^{1/2}} (\varrho_n + \phi^{-1} \log^{1/2} p) \lesssim \frac{\phi^2 \log^2 p}{n^{1/2}} (\phi L_n \varrho_n + L_n \log^{1/2} p),
\]

where we have used \( \beta = \phi \log p \). The desired assertion (26) then follows.

**Step 2.** We are now in position to finish the proof. Let

\[
V_n := \sqrt{\psi S_n^X} + \sqrt{1 - \psi} S_n^Y.
\]

Then we have

\[
P(V_n \leq y - \phi^{-1}) \leq P(F_\beta(V_n) \leq 0) \leq E[m(V_n)]
\]

\[
\leq P(F_\beta(S_n^W) \leq \phi^{-1} + (E[m(V_n)] - E[m(S_n^W)]))
\]

\[
\leq P(S_n^W \leq y + \phi^{-1}) + E[I_n]
\]

\[
\leq P(S_n^W \leq y - \phi^{-1}) + C \phi^{-1} \log^{1/2} p + E[I_n],
\]

where the first three lines follow from the properties of \( F_\beta(w) \) and \( g(t) \) (recall that \( m(w) = g(F_\beta(w)) \)), and the last inequality follows from Lemma A.1. Here the constant \( C \) depends only on \( b \). Likewise we have

\[
P(V_n \leq y - \phi^{-1}) \geq P(S_n^W \leq y - \phi^{-1}) - C \phi^{-1} \log^{1/2} p + E[I_n].
\]

The conclusion of the lemma follows from combining these inequalities with the bound on \( |E[I_n]| \) derived in Step 1.

**Proof of Corollary 5.1.** Pick any rectangle

\[
A = \{w \in \mathbb{R}^p : w_j \in [a_j, b_j] \text{ for all } j = 1, \ldots, p\}.
\]

For \( i = 1, \ldots, n \), consider the random vectors \( \tilde{X}_i \) and \( \tilde{Y}_i \) in \( \mathbb{R}^{2p} \) defined by \( \tilde{X}_{ij} = X_{ij} \) and \( \tilde{Y}_{ij} = Y_{ij} \) for \( j = 1, \ldots, p \), and \( \tilde{X}_{ij} = -X_{i,j-p} \) and \( \tilde{Y}_{ij} = -Y_{i,j-p} \) for \( j = p+1, \ldots, 2p \). Then

\[
P(S_n^X \in A) = P(S_n^\tilde{X} \leq y), \quad P(S_n^Y \in A) = P(S_n^\tilde{Y} \leq y),
\]

where the vector \( y \in \mathbb{R}^{2p} \) is defined by \( y_j = b_j \) for \( j = 1, \ldots, p \) and \( y_j = -a_{j-p} \) for \( j = p+1, \ldots, 2p \), and \( S_n^\tilde{X} \) and \( S_n^\tilde{Y} \) are defined as \( S_n^X \) and \( S_n^Y \) with \( X_i \)'s and \( Y_i \)'s replaced by \( \tilde{X}_i \)'s and \( \tilde{Y}_i \)'s. Hence the corollary follows from applying Lemma 5.1 to \( \tilde{X}_1, \ldots, \tilde{X}_n \) and \( \tilde{Y}_1, \ldots, \tilde{Y}_n \).

**Appendix C. Proofs for Section 2**

**Proof of Theorem 2.1.** The proof relies on Lemma 5.1 and its Corollary 5.1. Let \( K' \) denote a constant from the conclusion of Corollary 5.1. This constant depends only on \( b \). Set \( K_2 := 1/(K' \vee 1) \) in (7), so that

\[
\phi_n = \frac{1}{K' \vee 1} \left( \frac{\log^4 p}{n} \right)^{-1/6}.
\]
Without loss of generality, we may assume that $\phi_n \geq 2$; otherwise, the assertion of the theorem holds trivially by setting $K_1 = 2(K' \lor 1)$.

Then applying Corollary 5.1 with $\phi = \phi_n/2$, we have

$$
\varrho'_n \leq \frac{\varrho'_n}{8(K' \lor 1)^2} + \frac{3(K' \lor 1)^2 T_n^{1/3} \log^{7/6} p}{n^{1/6}} + \frac{M_n(\phi_n)}{8(K' \lor 1)^2 L_n}.
$$

Since $8(K' \lor 1)^2 > 1$, solving this inequality for $\varrho'_n$ and observing that $\rho_n(A^{re}) \leq \varrho'_n$ leads to the desired assertion.

Before proving Corollary 2.1, we shall verify the following elementary inequality.

**Lemma C.1.** Let $\xi$ be a non-negative random variable such that $P(\xi > x) \leq Ae^{-x/B}$ for all $x \geq 0$ and for some constants $A, B > 0$. Then for every $t \geq 0$, $E[\xi^3 \mathbb{1}\{\xi > t\}] \leq 6A(t + B)^3 e^{-t/B}$.

**Proof of Lemma C.1.** Observe that

$$
E[\xi^3 \mathbb{1}\{\xi > t\}] = 3 \int_0^t P(\xi > t)x^2 dx + 3 \int_t^\infty P(\xi > x)x^2 dx
$$

$$
= P(\xi > t)t^3 + 3 \int_t^\infty P(\xi > x)x^2 dx.
$$

Since $P(\xi > x) \leq Ae^{-x/B}$, using integration by parts, we have

$$
\int_t^\infty P(\xi > s)x^2 dx \leq A(Bt^2 + 2B^2t + 2B^3)e^{-t/B},
$$

which leads to

$$
E[\xi^3 \mathbb{1}\{\xi > t\}] \leq A(t^3 + 3Bt^2 + 6B^2t + 6B^3)e^{-t/B} \leq 6A(t + B)^3 e^{-t/B}.
$$

**Proof of Corollary 2.1.** The proof relies on application of Theorem 2.1. Without loss of generality, we may assume that

$$
\{D_n^{(1)}\}^6 = \frac{B_n^2 \log^7(pn)}{n} \leq c := \min\{(c_1/2)^3, (K_2/2)^6\},
$$

where $K_2$ appears in (7) and $c_1 > 0$ is a constant that depends only on $b$ ($c_1$ will be defined later), since otherwise we can make the assertions trivial by setting $C$ large enough.

Now by Theorem 2.1, we have

$$
\rho_n(A^{re}) \leq K_1 \left[ \left( \frac{B_n^2 \log^7 p}{n} \right)^{1/6} + \frac{M_n(\phi_n) + M_n(\phi_n)}{L_n} \right],
$$
where $\phi_n = K_2\{n^{-1}\log^4 p\}^{-1/6}$, and $L_n$ is any constant such that $L_n \geq L_n$. Recall that

$$L_n = \max_{1 \leq j \leq p} \sum_{i=1}^{n} E[|X_{ij}|^3]/n,$$

$$M_{n,X}(\phi_n) = n^{-1} \sum_{i=1}^{n} E \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \left\{ \max_{1 \leq j \leq p} |X_{ij}| > \sqrt{n}/(4\phi p) \right\} \right],$$

and $M_{n,Y}(\phi_n)$ is defined similarly with $X_{ij}$’s replaced by $Y_{ij}$’s.

It remains to choose a suitable constant $\mathcal{L}_n$ such that $\mathcal{L}_n \geq L_n$ and bound $M_{n,X}(\phi_n)$ and $M_{n,Y}(\phi_n)$. To this end, we consider cases (E.1) and (E.2) separately. In what follows, the notation $\lesssim$ means that the left hand side is bounded by the right hand side up to a positive constant that depends only on $b$ under case (E.1), and on $b$ and $q$ under case (E.2).

**Case (E.1).** By condition (M.2), we have $L_n \leq B_n =: \mathcal{L}_n$. Observe that (E.1) implies that $\|X_{ij}\|_{\psi_1} \leq B_n$ for all $i$ and $j$. Hence by Lemma 2.2.2 in [40], we have for some universal constant $C_1 > 0$, $\|\max_{1 \leq j \leq p} X_{ij}\|_{\psi_1} \leq C_1 B_n \log p$, which, together with Markov’s inequality, implies that for every $t > 0$,

$$P \left( \max_{1 \leq j \leq p} |X_{ij}| > t \right) \leq 2 \exp \left( -\frac{t}{C_1 B_n \log p} \right).$$

Applying Lemma C.1, we have

$$M_{n,X}(\phi_n) \lesssim (\sqrt{n}/(\phi_n \log p) + B_n \log p)^3 \exp \left( -\frac{\sqrt{n}}{4C_1 \phi_n B_n \log^2 p} \right).$$

Here

$$\frac{\sqrt{n}}{4C_1 \phi_n B_n \log^2 p} = \frac{c_1 n^{1/3}}{B_n^{2/3} \log^{4/3} p} \left( c_1 := \frac{1}{4K_2 C_1} \right) \geq c_1 e^{-1/3} \log(pn) \geq 2 \log(pn). \quad \text{(by (31))}$$

Moreover, by (31) and $\phi^{-1}_n = K_2^{-1} \{n^{-1} B_n^2 \log^4 p\}^{1/6} \leq c^{1/6}/K_2 \leq 1$, we have $(\sqrt{n}/(\phi_n \log p) + B_n \log p)^3 \lesssim n^{2/3}$, which implies that

$$M_{n,X}(\phi_n) \lesssim n^{2/3} \exp(-2 \log(pn)) \leq n^{-1/2}.$$ 

For $M_{n,Y}(\phi_n)$, since $E[Y_{ij}^2] = E[X_{ij}^2] \leq C_1 B_n^2$ and hence $\|Y_{ij}\|_{\psi_1} \lesssim B_n$ for all $i$ and $j$ (as each $Y_{ij}$ is Gaussian), we also have $M_{n,Y}(\phi_n) \lesssim n^{-1/2}$. The conclusion of the corollary in this case follows from the fact that $n^{-1/2} B_n^{-1} \leq D_n^{(1)}$.

**Case (E.2).** Without loss of generality, in addition to (31), we may assume that

$$D_n^{(2)} 3/2 = \frac{B_n \log^{3/2} p}{n^{1/2 - 1/q}} \leq (K_2/2)^{3/2}. \quad (32)$$
We begin with noting that
\[ L_n \leq B_n \leq \left\{ B_n + \frac{B_n^2}{n^{1/2-2/q} \log^{1/2} p} \right\} =: \overline{T}_n. \]

As the map \( x \mapsto x^{1/6} \) is sub-linear, \( \{n^{-1} \overline{T}_n^2 \log^p p\}^{1/6} \leq D_n^{(1)} + D_n^{(2)} \leq K_2, \)
so that by (31) and (32), \( \phi^{-1}_n = K_2^{-1} \{n^{-1} \overline{T}_n^2 \log^4 p\}^{1/6} \leq \epsilon^{1/6}/K_2 \leq 1. \)

Note that for any real-valued random variable \( Z \) and any \( t > 0, E[|Z|^3(|Z| > t)] \leq E[|Z|^3(|Z|/t)^{-3}] \leq t^{-3}E[|Z|^3]. \) Hence
\[ M_{n,X}(\phi_n) \lesssim \frac{B_n^2 \phi_n^{-3} \log^{q-3} p}{n^{q/2-3/2}}. \]

Here using the bound \( \overline{T}_n^{-1} \leq B_n^{-2} n^{1/2-2/q} \log^{1/2} p, \) we have that \( \phi_n \lesssim n^{1/3-2/(3q)} B_n^{-2/3} \log^{1/2} p, \) so that
\[ M_{n,X}(\phi_n) \lesssim \frac{B_n^{3+2}(\log p)^{q/2-3/2}}{n^{q+6/(1+6)-2/q}}, \]

which implies that \( M_{n,X}(\phi_n)/\overline{T}_n \lesssim D_n^{(2)}. \) Meanwhile, as in the previous case, we have \( M_{n,Y}(\phi_n) \lesssim n^{-1/2}, \) which leads to the desired conclusion in this case. ■

**Appendix D. Proofs for Section 3**

**Proof of Corollary 3.1.** Here \( C \) denotes a positive constant that depends only on \( a, b, \) and \( d \) if (E.1') is satisfied, and on \( a, b, d, \) and \( q \) if (E.2') is satisfied; the value of \( C \) may change from place to place. Pick any \( A \in A^m. \) Let \( A_n^m \) be an approximating \( m \)-generated convex set as in (C.1). By assumption, \( A_n^m \subset A \subset A_n^{m,\epsilon}, \) so that by letting
\[ \overline{p} := |P(S_n^X \notin A^m) - P(S_n^X \notin A^m, m, \epsilon) - P(S_n^Y \notin A^m, m, \epsilon)|, \]
we have
\[ P(S_n^X \in A) \leq P(S_n^X \in A^m, m, \epsilon) \leq P(S_n^Y \in A^m, m, \epsilon) + \overline{p} \]
\[ \leq P(S_n^Y \in A^m) + C\epsilon \log^{1/2} p + \overline{p} \leq P(S_n^Y \in A) + C\epsilon \log^{1/2} p + \overline{p}. \]

Likewise we have \( P(S_n^X \in A) \geq P(S_n^Y \in A) - C\epsilon \log^{1/2} p - \overline{p}, \) by which we conclude
\[ |P(S_n^X \in A) - P(S_n^Y \in A)| \leq C\epsilon \log^{1/2} p + \overline{p}. \]

Recalling that \( \epsilon = a/n \) and \( B_n \geq 1, \) we have \( \epsilon \log^{1/2} p \leq CD_n^{(1)}, \) hence the assertions of the corollary follow if we prove
\[ \overline{p} \leq \begin{cases} CD_n^{(1)} & \text{if (E.1') is satisfied,} \\ C\{D_n^{(1)} + D_n^{(2)}\} & \text{if (E.2') is satisfied.} \end{cases} \]

However, this follows from application of Corollary 2.1 to \( \tilde{X}_1, \ldots, \tilde{X}_n \) instead of \( X_1, \ldots, X_n. \) ■
Proof of Corollary 3.2. Since $X_i$ is a centered random vector with a log-concave distribution in $\mathbb{R}^p$, Borell’s inequality [see 10, Lemma 3.1] implies that $\|v'X_i\|_{q_1} \leq c(E[(v'X_i)^2])^{1/2}$ for all $v \in \mathbb{R}^p$ for some universal constant $c > 0$ [see 26, Appendix III]; hence if the maximal eigenvalue of each $E[X_iX_i']$ is bounded from below by a constant $k$, then every simple convex set $A \in \mathcal{A}^s$ obeys conditions (M.2′) and (E.1′) with $B_n$ replaced by a constant that depends only on $c$ and $k_2$. Besides if the minimal eigenvalue of each $E[X_iX_i']$ is bounded from below by a constant $k_1$, then every simple convex set $A \in \mathcal{A}^s$ obeys condition (M.1′) with $b$ replaced by a positive constant that depends only on $k_1$. Hence the conclusion of the corollary follows from application of Corollary 3.1.

Proof of Corollary 3.3. Here $C$ denotes a positive constant that depends only on $b$ and $s$ if condition (E.1) is satisfied, and on $b, s,$ and $q$ if condition (E.2) is satisfied; the value of $C$ may change from place to place. Without loss of generality, we may assume that $B_2^d \leq n$ since otherwise the assertions are trivial. We begin with preparing some notation. Let $R = pn^{5/2}$ and $\varepsilon = n^{-1}$, and let $\mathcal{A}^p_1(s)$ denote the subclass of $\mathcal{A}^p(s)$ consisting of every set $A$ in $\mathcal{A}^p(s)$ satisfying $\max_{1 \leq j \leq p} |w_j| \leq R$ for every $w \in A$ and containing a ball with radius $\varepsilon$ and center at, say, $w_A$. Let $\mathcal{A}^p_2(s) = \mathcal{A}^p(s) \setminus \mathcal{A}^p_1(s)$.

We divide the rest of the proof into five steps. In Steps 1-4, we verify conditions (C), (M.1′), (M.2′), (E.1′) (if (E.1) is satisfied), and condition (E.2′) (if (E.2) is satisfied) for all $A \in \mathcal{A}^p_1(s)$. An application of Corollary 3.1 then shows that the assertions (12) and (13) hold with $\rho_n(\mathcal{A}^p(s))$ replaced by $\rho_n(\mathcal{A}^p_1(s))$. Step 5 shows that the same assertions also hold with $\rho_n(\mathcal{A}^p(s))$ replaced by $\rho_n(\mathcal{A}^p_2(s))$. Since $\rho_n(\mathcal{A}^p(s)) = \rho_n(\mathcal{A}^p_1(s)) \cup \rho_n(\mathcal{A}^p_2(s))$, this will complete the proof. Step 1 relies on the following lemma, whose proof is given after the proof of this corollary.

Lemma D.1. Let $A$ be an $s$-sparsely convex set with a sparse representation $A = \cap_{q=1}^{Q} A_q$ for some $Q \leq p^s$. Assume that $A$ contains the origin, that $\sup_{w \in A} \|w\| \leq R$, and that all sets $A_q$ satisfy $-A_q \subset \mu A_q$ for some $\mu \geq 1$. Then for any $\gamma > e/8$, there exists $e_0 = e_0(\gamma) > 0$ such that for any $0 < \epsilon < e_0$, the set $A$ admits an approximation with precision $Re$ by an $m$-generated convex set $A^m$ where

$$m \leq Q \left( \gamma \sqrt{\frac{\mu + 1}{\epsilon}} \log \frac{1}{\epsilon} \right) R^2. $$

Moreover, the set $A^m$ can be chosen to satisfy

$$\|v\|_0 \leq s \text{ for all } v \in V(A^m).$$

(33)

Therefore, since $Q \leq p^s$, if $R \leq (pn)^{d_0}$ and $\mu \leq (pn)^{d_0}$ for some constant $d_0 \geq 1$, then there exists an absolute integer $n_0$ such that the set $A$ satisfies condition (C) for all $n \geq n_0$ with $a = 1$ and $d$ depending only on $s$ and $d_0$, and the approximating $m$-generated convex set $A^m$ satisfying (33).
Step 1. For Steps 1-4, pick any $s$-sparsely convex set $A \in \mathcal{A}_1^{sp}(s)$ with a sparse representation $A = \bigcap_{q=1}^{Q} A_q$ for some $Q \leq p^s$. Here we verify condition (C) for this set $A$. Consider the set $B := A - w_A := \{ w \in \mathbb{R}^p : w + w_A \in A \}$. The set $B$ contains a ball with radius $\varepsilon$ and center at the origin, satisfies the inequality $\|w\| \leq 2p^{1/2}R$ for all $w \in B$, and has a sparse representation $B = \bigcap_{q=1}^{Q} B_q$ where $B_q = A_q - w_A$. Clearly, each $B_q$ satisfies $-B_q \subset \mu B_q$ with $\mu = 2p^{1/2}R/\varepsilon = 2p^{3/2}n^{7/2}$. Therefore, applying Lemma D.1 to the set $B$ and noting that $A = B + w_A$ and $Q \leq p^s$, we see that there exists an absolute integer $n_0$ such that the set $A$ satisfies condition (C) for all $n \geq n_0$ with $a = 1$ and $d$ depending only on $s$, and an approximating $m$-generated convex set $A^m$ such that $\|v\|_0 \leq s$ for all $v \in \mathcal{V}(A^m)$.

Step 2. Here we verify condition (M.1'). Since we have $\|v\|_0 \leq s$ for all $v \in \mathcal{V}(A^m)$, condition (M.1') follows immediately from (M.1'').

Step 3. We shall verify condition (M.2'). For $v \in \mathcal{V}(A^m)$, let $J(v)$ be the set consisting of positions of non-zero elements of $v$, so that $\text{Card}(J(v)) \leq s$. Using the inequality $(\sum_{j \in J(v)} |a_j|^2)^{1/k} \leq s^1/k \sum_{j \in J(v)} |a_j|^2 + k$ for $a = (a_1, \ldots, a_p) \in \mathbb{R}^p$ (which follows from Hölder’s inequality), we have for $k = 1$ or 2,

$$\frac{1}{n} \sum_{i=1}^{n} E[|v' X_i|^{2+k}] \leq \frac{1}{n} \sum_{i=1}^{n} E\left[ \left( \sum_{j \in J(v)} |X_{ij}| \right)^{2+k} \right] \leq s^{1+k} \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in J(v)} |X_{ij}|^{2+k} \leq s^{2+k} B_n^k \leq (B'_n)^k,$$

where $B'_n = s^3 B_n$, which leads to condition (M.2') with $B_n$ replaced by $s^3 B_n$.

Step 4. We shall verify condition (E.1') when (E.1) is satisfied, and (E.2') when (E.2) is satisfied. When (E.1) is satisfied, we have $\|X_{ij}\|_{\psi_1} \leq B_n$, so that $\|v' X_i\|_{\psi_1} \leq \sum_{j \in J(v)} \|X_{ij}\|_{\psi_1} \leq s B_n$ showing that the vectors $\tilde{X}_i$, $i = 1, \ldots, n$, satisfy (E.1') with $B_n$ replaced by $s B_n$.

When (E.2) is satisfied,

$$E \left[ \max_{v \in \mathcal{V}(A^m)} |v' X_i|^q \right] \leq s^q E \left[ \max_{1 \leq j \leq p} |X_{ij}|^q \right],$$

showing that the vectors $\tilde{X}_i$, $i = 1, \ldots, n$, satisfy (E.2') with $B_n$ replaced by $s B_n$.

Combining Steps 1-4 and applying Corollary 3.1 shows that the assertions (12) and (13) hold with $\rho_n(\mathcal{A}^{sp}(s))$ replaced by $\rho_n(\mathcal{A}^{sp}_1(s))$.

Step 5. Here we show that the assertions (12) and (13) hold with $\rho_n(\mathcal{A}^{sp}(s))$ replaced by $\rho_n(\mathcal{A}^{sp}_2(s))$. Fix any $s$-sparsely convex set $A \in \mathcal{A}_2^{sp}(s)$ with a sparse representation $A = \bigcap_{q=1}^{Q} A_q$ for some $Q \leq p^s$. Let $A^R := \{ w \in \mathbb{R}^p : w_A \in A^R \}$.

Combining Steps 1-4 and applying Corollary 3.1 shows that the assertions (12) and (13) hold with $\rho_n(\mathcal{A}^{sp}(s))$ replaced by $\rho_n(\mathcal{A}^{sp}_2(s))$. Fix any $s$-sparsely convex set $A \in \mathcal{A}_2^{sp}(s)$ with a sparse representation $A = \bigcap_{q=1}^{Q} A_q$ for some $Q \leq p^s$. Let $A^R := \{ w \in \mathbb{R}^p : w_A \in A^R \}$.
$\mathbb{R}^p : \max_{1 \leq j \leq p} |w_j| > R$. Then $A = \overline{A} \cup (A \cap A^R)$ for some $s$-sparsely convex set $\overline{A} \subset \mathbb{R}^p$ such that $\max_{1 \leq j \leq p} |w_j| \leq R$ for all $w \in \overline{A}$.

Now observe that by Markov’s inequality,
\[
P(\max_{i,j} |X_{ij}| > pn^2) \leq \frac{E[\max_{i,j} |X_{ij}|]}{pn^2} \leq \frac{E[\sum_{i,j} |X_{ij}|]}{pn^2} \leq \max_{i,j} E[|X_{ij}|]/n \leq CB_n/n \leq C/n^{1/2},
\]
where $\max_{i,j}$ stands for $\max_{1 \leq i \leq n, 1 \leq j \leq p}$. Hence
\[
P(S_n^X \in A^R) \leq C/n^{1/2}.
\]

It is easy to verify that the same inequality also holds with $S_n^X$ replaced by $S_n^Y$, and we have
\[
|P(S_n^X \in A) - P(S_n^Y \in A)| = |P(S_n^X \in \overline{A}) - P(S_n^Y \in \overline{A})| + C/n^{1/2}.
\]
Therefore it suffices to only consider the case where $A \in A_2^{sp}(s)$ is such that $\max_{1 \leq j \leq p} |w_j| \leq R$ for all $w \in A$.

Next, we consider the two cases separately. First, suppose that at least one $A_q$ does not contain a ball with radius $\varepsilon$. Then the set $\cap_{n \geq s^p - 1} \{w \in \mathbb{R}^p : w'v \leq S_{A_q}(v) - \varepsilon\}$ is empty, and so under condition (M.2), Lemma A.2 implies that $P(S_n^X \in A_q) \leq C\varepsilon = C/n$ (since the Hilbert-Schmidt norm is equal to the square-root of the sum of squares of the eigenvalues of the matrix, under our condition (M.1′)), the constant $C$ in the bound $C\varepsilon$ above depends only on $b$ and $s$). In addition, under conditions (M.1′′) and (M.2), the Berry-Esseen theorem [see 21, Theorem 1.3] implies that
\[
|P(S_n^X \in A_q) - P(S_n^Y \in A_q)| \leq CB_n/n^{1/2}.
\]
Since $A \subset A_q$, both $P(S_n^X \in A)$ and $P(S_n^Y \in A)$ are bounded from above by the quantities on the right hand sides of (12) and (13) depending on whether (E.1) or (E.2) is satisfied, and so is their difference. This completes the proof in this case.

Second, suppose that each $A_q$ contains a ball with radius $\varepsilon$ (possibly depending on $q$). Then applying Lemma D.1 to each $A_q$ separately shows that for $n \geq n_0$ and $m \leq (pn)^d$ with $d$ depending only on $s$, we can construct an $m$-generated convex sets $A_q^m$ such that
\[
A_q^m \subset A_q \subset A_q^{m,1/n}
\]
and $\|v\| \leq s$ for all $v \in V(A_q^m)$. The set $A^0 = \cap_{q=1}^Q A_q^{m,1/n}$ trivially satisfies condition (C) with $a = 0$ and $d$ depending only on $s$. In addition, it follows from the same arguments as those used in Steps 2-4 that the set $A^0$ satisfies conditions (M.1′), (M.2′), (E.1′) (if (E.1) is satisfied), and (E.2′) (if (E.2) is satisfied). Therefore, by applying Corollary 3.1, we conclude that $|P(S_n^X \in A^0) - P(S_n^Y \in A^0)|$ is bounded from above by the quantities on the right hand sides of (10) and (11) depending on whether (E.1) or (E.2) is satisfied. Also, observe that $A \subset A^0$ and that $\cap_{q=1}^Q A_q^{m,-\varepsilon}$ is empty because $\cap_{q=1}^Q A_q^{m} \subset A$.
and $A$ contains no balls with radius $\varepsilon$. This implies that $P(S_n^Y \in A^0) \leq C(\log p)^{1/2}/n$ by Lemma A.1 and condition (M.1'). Since $A \subseteq A^0$, both $P(S_n^X \in A)$ and $P(S_n^Y \in A)$ are bounded from above by the quantities on the right hand sides of (12) and (13) depending on whether (E.1) or (E.2) is satisfied, and so is their difference. This completes the proof in this case. ■

Here we prove Lemma D.1 used in the proof of Corollary 3.3.

**Proof of Lemma D.1.** For convex sets $P_1$ and $P_2$ containing the origin and such that $P_1 \subseteq P_2$, define

$$d_{BM}(P_1, P_2) := \inf\{\epsilon > 0 : P_2 \subset (1 + \epsilon)P_1\}.$$

It is immediate to verify that the function $d_{BM}$ has the following useful property: for any convex sets $P_1$, $P_2$, $P_3$, and $P_4$ containing the origin and such that $P_1 \subseteq P_2$ and $P_3 \subseteq P_4$,

$$d_{BM}(P_1 \setminus P_3, P_2 \setminus P_4) \leq d_{BM}(P_1 \cap P_2) \leq d_{BM}(P_3 \cap P_4). \quad (34)$$

Let $A = \cap_{q=1}^Q A_q$ be a sparse representation of $A$ as appeared in the statement of the lemma. Fix any $A_q$. By assumption, the indicator function $w \mapsto I(w \in A_q)$ depends only on $s_q \leq s$ elements of its argument $w = (w_1, \ldots, w_p)$. Since $A$ contains the origin, $A_q$ contains the origin as well. Therefore, applying Corollary 1.5 in [4] as if $A_q$ was a set in $\mathbb{R}^{s_q}$ shows that one can construct a polytope $P_q \subset \mathbb{R}^p$ with at most $(\gamma((\mu + 1)/\epsilon)^{1/2} \log(1/\epsilon))^{s_q}$ vertices such that

$$P_q \subset A_q \subset (1 + \epsilon)P_q$$

and such that for all $v \in \mathcal{V}(P_q)$, non-zero elements of $v$ correspond to some of the main components of $A_q$. Since we need at most $s_q$ vertices to form a face of the polytope $P_q$, the polytope $P_q$ has

$$m_q \leq \left(\gamma \sqrt{\frac{\mu + 1}{\epsilon} \log \frac{1}{\epsilon}}\right)^{s_q} \leq \left(\gamma \sqrt{\frac{\mu + 1}{\epsilon} \log \frac{1}{\epsilon}}\right)^{s_q} \quad (35)$$

faces. Now observe that $P_q$ is an $m_q$-generated convex set. Thus, we have constructed an $m_q$-generated convex set $P_q$ such that $P_q \subset A_q \subset (1 + \epsilon)P_q$ and all vectors in $\mathcal{V}(P_q)$ having at most $s$ non-zero elements. Hence

$$d_{BM}(P_q, A_q) \leq \epsilon.$$

Next, it follows from (34) that

$$d_{BM}(\cap_{q=1}^Q P_q, \cap_{q=1}^Q A_q) \leq \epsilon.$$

Therefore, defining $A^m = \cap_{q=1}^Q P_q$, we obtain from $A = \cap_{q=1}^Q A_q$ that

$$A^m \subset A \subset (1 + \epsilon)A^m \subset A^{m \cdot R},$$

where the last assertion follows from the assumption that $\sup_{w \in A} \|w\| \leq R$. Since $A^m$ is an $m$-generated convex set with $m \leq \sum_{q=1}^Q m_q$, the first
claim of the lemma now follows from (35). The second claim (33) holds by construction of $A^n$, and the final claim is trivial.

APPENDIX E. PROOFS FOR SECTION 4

E.1. Maximal inequalities. Here we collect some useful maximal inequalities that will be used in the proofs for Section 4.

**Lemma E.1.** Let $X_1, \ldots, X_n$ be independent centered random vectors in $\mathbb{R}^p$ with $p \geq 2$. Define $Z := \max_{1 \leq j \leq p} \left| \sum_{i=1}^n X_{ij} \right|$, $M := \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \left| X_{ij} \right|$ and $\sigma^2 := \max_{1 \leq j \leq p} \sum_{i=1}^n E[X_{ij}^2]$. Then

$$E[Z] \leq K(\sigma \sqrt{\log p} + \sqrt{E[M^2]} \log p),$$

where $K$ is a universal constant.

**Proof.** See Lemma 8 in [18].

**Lemma E.2.** Assume the setting of Lemma E.1. (i) For every $\eta > 0$, $\beta \in (0, 1]$ and $t > 0$,

$$P\{Z \geq (1 + \eta)E[Z] + t\} \leq \exp\{-t^2/(3\sigma^2)\} + 3 \exp\{-(t/(K\|M\|\beta))^{\beta}\},$$

where $K = K(\eta, \beta)$ is a constant depending only on $\eta, \beta$. (ii) For every $\eta > 0$, $s \geq 1$ and $t > 0$,

$$P\{Z \geq (1 + \eta)E[Z] + t\} \leq \exp\{-t^2/(3\sigma^2)\} + K'E[M^s]/t^s,$$

where $K' = K'(\eta, s)$ is a constant depending only on $\eta, s$.

**Proof.** See Theorem 4 in [1] for case (i) and Theorem 2 in [2] for case (ii). See also [20].

**Lemma E.3.** Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^p$ with $p \geq 2$ such that $X_{ij} \geq 0$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Define $Z := \max_{1 \leq j \leq p} \sum_{i=1}^n X_{ij}$ and $M := \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} X_{ij}$. Then

$$E[Z] \leq K \left( \max_{1 \leq j \leq p} E[\sum_{i=1}^n X_{ij}] + E[M] \log p \right),$$

where $K$ is a universal constant.

**Proof.** See Lemma 9 in [18].

**Lemma E.4.** Assume the setting of Lemma E.3. (i) For every $\eta > 0$, $\beta \in (0, 1]$ and $t > 0$,

$$P\{Z \geq (1 + \eta)E[Z] + t\} \leq 3 \exp\{-t/(K\|M\|\beta)^{\beta}\},$$

where $K = K(\eta, \beta)$ is a constant depending only on $\eta, \beta$. (ii) For every $\eta > 0$, $s \geq 1$ and $t > 0$,

$$P\{Z \geq (1 + \eta)E[Z] + t\} \leq K'E[M^s]/t^s,$$

where $K' = K'(\eta, s)$ is a constant depending only on $\eta, s$. 
The proof of Lemma E.4 relies on the following lemma, which follows from Theorem 10 in [25].

**Lemma E.5.** Assume the setting of Lemma E.3. Suppose that there exists a constant $B$ such that $M \leq B$. Then for every $\eta, t > 0$,

$$P\left\{ Z \geq (1 + \eta)E[Z] + B \left( \frac{2}{3} + \frac{1}{\eta} \right) t \right\} \leq e^{-t}.$$  

**Proof of Lemma E.5.** By homogeneity, we may assume that $B = 1$. Then by Theorem 10 in [25], for every $\lambda > 0$,

$$\log E[\exp(\lambda(Z - E[Z]))] \leq \varphi(\lambda)E[Z],$$

where $\varphi(\lambda) = e^\lambda - \lambda - 1$. Hence by Markov’s inequality, with $a = E[Z]$,

$$P\{ Z - E[Z] \geq t \} \leq e^{-\lambda t + a\varphi(\lambda)}.$$

The right hand side is minimized at $\lambda = \log(1+t/a)$, at which $-\lambda t + a\varphi(\lambda) = -aq(t/a)$ where $q(t) = (1 + t)\log(1 + t) - t$. It is routine to verify that $q(t) \geq t^2/(2(1 + t/3))$, so that

$$P\{ Z - E[Z] \geq t \} \leq e^{-\frac{t^2}{2(a + t/3)}}.$$

Solving $t^2/(2(a + t/3)) = s$ gives $t = s/3 + \sqrt{s^2/9 + 2as} \leq 2s/3 + \sqrt{2as}$. Therefore, we have

$$P\{ Z \geq E[Z] + \sqrt{2as} + 2s/3 \} \leq e^{-s}.$$

The conclusion follows from the inequality $\sqrt{2as} \leq \eta a + \eta^{-1}s$. □

**Proof of Lemma E.4.** The proof is a modification of that of Theorem 4 in [1] (or Theorem 2 in [2]). We begin with noting that we may assume that $(1 + \eta)8E[M] \leq t/4$, since otherwise we can make the lemma trivial by setting $K$ or $K'$ large enough. Take

$$\rho = 8E[M], \quad Y_{ij} = \begin{cases} X_{ij}, & \text{if } \max_{1 \leq j \leq p} X_{ij} \leq \rho, \\ 0, & \text{otherwise} \end{cases}$$

Define

$$W_1 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} Y_{ij}, \quad W_2 = \max_{1 \leq j \leq p} \sum_{i=1}^{n} (X_{ij} - Y_{ij}).$$

Then

$$P\{ Z \geq (1 + \eta)E[Z] + t \} \leq P\{ W_1 \geq (1 + \eta)E[Z] + 3t/4 \} + P(W_2 \geq t/4) \leq P\{ W_1 \geq (1 + \eta)E[W_1] - (1 + \eta)E[W_2] + 3t/4 \} + P(W_2 \geq t/4).$$

Observe that

$$P\left\{ \max_{1 \leq m \leq n} \max_{1 \leq j \leq p} \sum_{i=1}^{m} (X_{ij} - Y_{ij}) > 0 \right\} \leq P(M > \rho) \leq 1/8,$$
so that by the Hoffmann-Jørgensen inequality [see 24, Proposition 6.8], we have
\[ \mathbb{E}[W_2] \leq 8\mathbb{E}[M] \leq t/(4(1 + \eta)). \]

Hence
\[ \mathbb{P}\{Z \geq (1 + \eta)\mathbb{E}[Z] + t\} \leq \mathbb{P}\{W_1 \geq (1 + \eta)\mathbb{E}[W_1] + t/2\} + \mathbb{P}(W_2 \geq t/4). \]

By Lemma E.5, the first term on the right hand side is bounded by \( e^{-ct/\rho} \) where \( c \) depends only on \( \eta \). We bound the second term separately in cases (i) and (ii). Below \( C_1, C_2, \ldots \) are constants that depend only on \( \eta, \beta, s \).

Case (i). By Theorem 6.21 in [24] (note that a version of their theorem applies to nonnegative random vectors) and the fact that \( \mathbb{E}[W_2^2] \leq 8\mathbb{E}[M] \),
\[ \|W_2\|_{\psi, \beta} \leq C_1(\mathbb{E}[W_2] + \|M\|_{\psi, \beta}) \leq C_2\|M\|_{\psi, \beta}, \]
which implies that \( \mathbb{P}(W_2 \geq t/4) \leq 2\exp\{-t/(C_5\|M\|_{\psi, \beta})^\beta\}. \) Since \( \rho \leq C_4\|M\|_{\psi, \beta} \), we conclude that
\[ e^{-ct/\rho} + \mathbb{P}(W_2 \geq t/4) \leq 3\exp\{-t/(C_5\|M\|_{\psi, \beta})^\beta\}. \]

Case (ii). By Theorem 6.20 in [24] (note that a version of their theorem applies to nonnegative random vectors) and the fact that \( \mathbb{E}[W_2^2] \leq 8\mathbb{E}[M] \),
\[ (\mathbb{E}[W_2^s])^{1/s} \leq C_6(\mathbb{E}[W_2] + (\mathbb{E}[M^s])^{1/s}) \leq C_7(\mathbb{E}[M^s])^{1/s}. \]

The conclusion follows from Markov’s inequality together with the simple fact that \( e^{-t/\rho} \to 0 \) as \( t \to \infty \).

\textbf{Proof of Theorem 4.1.} In this proof, \( C \) is a positive constant that depends only on \( a, b, \) and \( d \) but its value may change at each appearance. Fix any \( A \in \mathcal{A}^a \). Let \( A^m \) be an approximating \( m \)-generated convex set as in (C).

By assumption, \( A^m \subset A \subset A^{m,e} \). Let
\[ \bar{\rho} := \max \left\{ |\mathbb{P}(S_n^X \in A^m \mid X_1^n) - \mathbb{P}(S_n^Y \in A^m)|, \right. \]
\[ \left. |\mathbb{P}(S_n^X \in A^{m,e} \mid X_1^n) - \mathbb{P}(S_n^Y \in A^{m,e})| \right\}. \]

As in the proof of Corollary 3.1, we have
\[ |\mathbb{P}(S_n^X \in A \mid X_1^n) - \mathbb{P}(S_n^Y \in A)| \]
\[ \leq C\epsilon \log^{1/2}(pm) + \bar{\rho} \leq Cn^{-1} \log^{1/2}(pm) + \bar{\rho}, \]
so that the problem reduces to proving that under (M.1), the inequality
\[ \rho_n^{MB}(A^{re}) \leq C\Delta_n^{1/3} \log^{2/3} p \tag{36} \]
holds on the event \( \Delta_{n,r} \leq \bar{\Delta}_n, \) where \( \Delta_{n,r} := \max_{1 \leq j,k \leq p} |\Sigma_{jk}^X - \Sigma_{jk}^Y| \) with \( \Sigma_{jk}^X \) and \( \Sigma_{jk}^Y \) denoting the \((j,k)\)-th elements of \( \Sigma^X \) and \( \Sigma^Y \), respectively.

To this end, we first show that
\[ \rho_n^{MB} := \sup_{y \in \mathbb{R}^p} |\mathbb{P}(S_n^X \leq y \mid X_1^n) - \mathbb{P}(S_n^Y \leq y)| \leq C\Delta_n^{1/3} \log^{2/3} p. \tag{37} \]
To show (37), fix any $y = (y_1, \ldots, y_p)' \in \mathbb{R}^p$. As in the proof of Lemma 5.1, for $\beta > 0$, define

$$F_\beta(w) := \frac{1}{\beta} \log \left( \sum_{j=1}^p \exp(\beta(w_j - y_j)) \right), \ w \in \mathbb{R}^p.$$ 

Note that conditional on $X_1^n$, $S_{n}^{eX}$ is a centered Gaussian random vector with covariance matrix $\Sigma_{n}^{eX}$. Then a small modification of the proof of Theorem 1 in [18] implies that for every $g \in C^2(\mathbb{R})$ with $\|g'\|_{\infty} \vee \|g''\|_{\infty} < \infty$, we have

$$|E[g(F_{\beta}(S_{n}^{eX})) | X_1^n] - E[g(F_{\beta}(S_{n}^Y))]| \leq (\|g''\|_{\infty}/2 + \beta \|g'\|_{\infty})\Delta_{n;r}. $$

Hence, as in Step 2 of the proof of Lemma 5.1, we obtain with $\phi = \beta/\log p$ that

$$|P(S_{n}^{eX} \leq s - \phi^{-1} | X_1^n) - P(S_{n}^Y \leq s - \phi^{-1})| \leq C \left\{ \phi^{-1}(\log p)^{1/2} + (\phi^2 + \beta \phi)\Delta_{n;r} \right\}.$$

Substituting $\beta = \phi \log p$, optimizing the resulting expression with respect to $\phi$, and noting that $y \in \mathbb{R}^p$ is arbitrary lead to (37). Finally (36) follows from the fact that the inequality $Q_{n}^{MB} \leq C\Delta_{n}^{1/3} \log^{2/3} p$ holds on the event $\Delta_{n;r} \leq \hat{\Delta}_n$, and applying the same argument as that used in the proof of Corollary 5.1.

**Proof of Corollary 4.1.** In this proof, $c$ and $C$ are positive constants that depend only on $a, b, d$, and $s$ under (E.1), and on $a, b, d, s,$ and $q$ under (E.2); their values may vary from place to place. For brevity of notation, we implicitly assume here that $i$ is varying over $\{1, \ldots, n\}$, and $j$ and $k$ are varying over $\{1, \ldots, p\}$. Finally, without loss of generality, we will assume that

$$B_n^2(\log^5(pn)) \log^2(1/\alpha) \leq n$$

(38)
since otherwise the assertions are trivial.

We shall apply Theorem 4.1 to prove the corollary. Since $n^{-1/2} \log^{1/2}(pn) \leq CD_n^{(1)}(\alpha)$, it suffices to construct an appropriate $\hat{\Delta}_n$ such that $P(\Delta_n > \hat{\Delta}_n) \leq \alpha$ and to bound $\Delta_{n}^{1/3} \log^{2/3} (pn)$.

We begin with observing that under condition (C'), $\Delta_n \leq C\Delta_{n;r}$ where $\Delta_{n;r} = \max_{1 \leq i, k \leq p} |\Sigma_{ik}^{eX} - \Sigma_{ik}^Y|$. As $E[X_i X_i'] = E[Y_i Y_i']$ for all $i$, we have

$$\Sigma_{ik}^{eX} - \Sigma_{ik}^Y = n^{-1} \sum_{i=1}^n (X_i X_i' - E[X_i X_i']) - \hat{\mu}_n^{X'} \hat{\mu}_n^{X}' \hat{\mu}_n^{X'},$$

by which we have $\Delta_{n;r} \leq \Delta_{n;r}^{(1)} + (\Delta_{n;r}^{(2)})^2$, where

$$\Delta_{n;r}^{(1)} := \max_{1 \leq i, k \leq p} n^{-1} \sum_{i=1}^n |X_{ij} X_{ik} - E[X_{ij} X_{ik}]|, \Delta_{n;r}^{(2)} := \max_{1 \leq i, j \leq p} |\hat{\mu}_n^{X_{ij}}|.$$
The desired assertions then follow from the bounds on $\Delta_{n,r}^{(1)}$ and $\Delta_{n,r}^{(2)}$ derived separately for (E.1) and (E.2) cases below.

**Case (E.1).** Observe that by Hölder’s inequality and (M.2),

$$\sigma_n^2 := \max_{j,k} \sum_{i=1}^{n} E \left[ (X_{ij} X_{ik} - E[X_{ij} X_{ik}])^2 \right] \leq \max_{j,k} \sum_{i=1}^{n} E[|X_{ij} X_{ik}|^2] \leq nB_n^2.$$

In addition, by (E.1),

$$\|X_{ij} X_{ik}\|_{\psi_1/2} = \|X_{ij}\|_{\psi_1/2} = \|X_{ij}\|_{\psi_1} \leq CB_n^2 \log^2(pn),$$

so that for $M_n := \max_{j,k} |X_{ij} X_{ik} - E[X_{ij} X_{ik}]|$, we have

$$\|M_n\|_{\psi_1/2} \leq C\{\|\max_{i,j,k} |X_{ij} X_{ik}|\|_{\psi_1/2} + \max_{i,j,k} E[|X_{ij} X_{ik}|]\} \leq C\{B_n^2 \log^2(pn) + B_n^2\} \leq CB_n^2 \log^2(pn),$$

which also implies that $(E[M_n^2])^{1/2} \leq CB_n^2 \log^2(pn)$. Hence by Lemma E.1, we have

$$E[\Delta_{n,1}] \leq Cn^{-1}\{\sqrt{\sigma_n^2 \log p} + \sqrt{E[M_n^2] \log p}\} \leq C\{(n^{-1}B_n^2 \log p)^{1/2} + n^{-1}B_n^2 \log^3(pn)\} \leq C\{n^{-1}B_n^2 \log(pn)^{1/2}\},$$

where the last inequality follows from (38). Applying Lemma E.2 (i) with $\beta = 1/2$ and $\eta = 1$, we conclude that for every $t > 0$,

$$P\left\{\Delta_{n,r}^{(1)} > C\{n^{-1}B_n^2 \log(pn)^{1/2} + t\}\right\} \leq \exp\{-nt^2/(3B_n^2)\} + 3\exp\{-c\sqrt{nt}/(B_n \log(pn))\}.$$

Choosing $t = C\{n^{-1}B_n^2 \log(pn)^{1/2}\}^{1/2}$ for sufficiently large $C > 0$, the right hand side is bounded by

$$\alpha/4 + 3\exp\{-C^{1/2}n^{1/4}(\log^{1/2}(1/\alpha))/(B_n^{1/2} \log^{3/4}(pn))\} \leq \alpha/2,$$

where the last inequality follows from (38). Therefore

$$P\{E[\Delta_{n,r}^{(1)}]^2]^{1/3} > CD_{n,r}^{(1)}(\alpha)\} \leq \alpha/2.$$

It is routine to verify that the same inequality holds with $\Delta_{n,r}^{(1)}$ replaced by $\Delta_{n,r}^{(2)}$. This leads to the conclusion of the corollary under (E.1).

**Case (E.2).** Define $\sigma_n^2$ and $M_n$ by the same expressions as those in the previous case; then $\sigma_n^2 \leq nB_n^2$. For $M_n$, we have

$$E[M_n^2] \leq C\{E[\max_{i,j,k} |X_{ij} X_{ik}|^{q/2}] + \max_{i,j,k} E[|X_{ij} X_{ik}|]^{q/2}\} \leq C\{E[\max_{i,j,k} |X_{ij} X_{ik}|^{q/2}]\} = CE[\max_{i,j} |X_{ij}|^q] \leq CnB_n^q,$$
which also implies that \( (E[M_n^2])^{1/2} \leq Cn^{2/q}B_n^2 \). Hence by Lemma E.1, we have

\[
E[\Delta_n^{(1)}] \leq Cn^{-1}\{\sqrt{\sigma_k^2\log p} + \sqrt{E[M_n^2]\log p}\} \\
\leq C\{(n^{-1}B_n^2\log p)^{1/2} + n^{-1+2/q}B_n^2\log p\}.
\]

Applying Lemma E.2 (ii) with \( s = q/2 \) and \( \eta = 1 \), we have for every \( t > 0 \),

\[
P\left\{ \Delta_n^{(1)} > C\{(n^{-1}B_n^2\log p)^{1/2} + n^{-1+2/q}B_n^2\log p\} + t \right\} \\
\leq \exp\{-nt^2/(3B_n^2)\} + ct^{-q/2}n^{1-q/2}B_n^q.
\]

Choosing

\[
t = C\{(n^{-1}B_n^2(\log(pn))\log^2(1/\alpha))^1/2 + n^{1-q/2}\alpha^{-2/q}B_n^2\}
\]

for sufficiently large \( C > 0 \), we conclude that

\[
P\{\Delta_n^{(1)}\log^2 p\}^{1/3} > C\{D_n^{(1)}(\alpha) + D_n^{(2)}(\alpha)\} \leq \alpha/2.
\]

It is routine to verify that the same inequality holds with \( \Delta_n^{(1)} \) replaced by \( \{\Delta_n^{(2)}\}^2 \). This leads to the conclusion of the corollary under (E.2). 

**Proof of Corollary 4.2.** Here \( C \) is understood to be a positive constant that depends only on \( a, d, k_1 \) and \( k_2 \); the value of \( C \) may change from place to place. To prove this corollary, we apply Theorem 4.1, to which we have to verify condition (M.1') and derive a suitable bound on \( \Delta_n \). Condition (M.1') follows from the fact that the minimum eigenvalue of \( E[X_i'X_i'] \) is bounded from below by \( k_1 \). By log-concavity of the distributions of \( X_i \), we have \( \|v'X_i\|_{\psi_1} \leq C(E[(v'X_i)^2])^{1/2} \leq C \) for all \( v \in \mathbb{R}^p \) with \( \|v\| = 1 \) (see the proof of Corollary 3.2). For each \( 1 \leq i \leq n \), let \( \tilde{X}_i \) be a random vector whose elements are given by \( v'X_i \), \( v \in \cup_{A \in A^d} V(A^m(A)) \); for each \( 1 \leq i \leq n \), the dimension of \( \tilde{X}_i \), denoted by \( \tilde{p} \), is at most \( (pn)^d \), and \( \|\tilde{X}_i\|_{\psi_1} \leq C \) for all \( 1 \leq j \leq \tilde{p} \). Then \( \Delta_n \) coincides with \( \Delta_{n,r} \) with \( X_i \) replaced by \( \tilde{X}_i \), that is,

\[
\Delta_n = \max_{1 \leq j, k \leq \tilde{p}} \left| n^{-1} \sum_{i=1}^{n} (\tilde{X}_{ij}\tilde{X}_{ik} - E[\tilde{X}_{ij}\tilde{X}_{ik}]) - E_n[\tilde{X}_{ij}]E_n[\tilde{X}_{ik}] \right|.
\]

Noting that \( \log \tilde{p} \leq d\log(pn) \), by the same argument as that used in the proof of Corollary 4.1 case (E.1), we can find a constant \( \tilde{\Delta}_n \) such that

\[
P(\Delta_n > \tilde{\Delta}_n) \leq \alpha \quad \text{and} \quad (\tilde{\Delta}_n\log^2 p)^{1/3} \leq C\{n^{-1}(\log^5(pn))\log^2(1/\alpha)\}^{1/6}.
\]

Here without loss of generality we assume that \( (\log^5(pn))\log^2(1/\alpha) \leq n \). The desired assertion then follows.
Proof of Corollary 4.3. In this proof, let $C$ be a positive constant depending only on $b, s, q$ ($C$ depends on $q$ only in the case where (E.2) is satisfied); the value of $C$ may change from place to place. Moreover, without loss of generality, we will assume that $B_n^2(\log^2(pn)) \log^2(1/\alpha) \leq n$ since otherwise the assertions are trivial.

Let $\Delta_{n,r} := \max_{1 \leq j,k \leq p} |\Sigma_{jk}^X - \Sigma_{jk}^Y|$, and

$$\Delta_n = \left\{ \begin{array}{ll} \left( \frac{B_n^2(\log^2(pn))\log^2(1/\alpha)}{n} \right)^{1/2} & \text{if (E.1) is satisfied} \\
\frac{B_n^2(\log^2(pn))\log^2(1/\alpha)}{n} + \frac{B_n^2\log p}{\alpha^2/q^{1-\alpha/2}} & \text{if (E.2) is satisfied} \end{array} \right.$$ 

Then by the proof of Corollary 4.1, in either case where (E.1) or (E.2) is satisfied, there exists a positive constant $C_1$ depending only on $b, s, q$ ($C_1$ depends on $q$ only in the case where (E.2) is satisfied) such that

$$P(\Delta_{n,r} > C_1\Delta_n) \leq \alpha/2.$$ 

We may further assume that $C_1s\Delta_n \leq b/2$, since otherwise the assertions are trivial.

Let $R = pn^{5/2}$ and $\varepsilon = n^{-1}$, and define the subclasses $A_{1p}^X(s)$ and $A_{2p}^Y(s)$ as in the proof of Corollary 3.3. Moreover, as in the proof of Corollary 3.3, for any $A \in A_{2p}^X(s)$, we can verify conditions (C), (M.1'), (M.2'), (E.1') (if (E.1) is satisfied), and condition (E.2') (if (E.2) is satisfied). Therefore, the bounds (14) and (15) with $\rho_n^{MB}(A_{1p}^X(s))$ replaced by $\rho_n^{MB}(A_{1p}^X(s))$ follow from Corollary 4.1. Since $\rho_n^{MB}(A_{2p}^X(s)) = \rho_n^{MB}(A_{1p}^X(s)) \vee \rho_n^{MB}(A_{2p}^Y(s))$, it remains to bound $\rho_n^{MB}(A_{2p}^Y(s))$.

As in the proof of Corollary 3.3, fix any $s$-sparsely convex set $A \in A_{2p}^X(s)$ with sparse representation $A = \cap_{q=1}^{Q} A_q$ with $Q \leq p^s$, and let $A^R = \{ w \in A : \max_{1 \leq j \leq p} |w_j| > R \}$. Then $A = \overline{A} \cup (A \cap A^R)$ for some $s$-sparsely convex set $\overline{A}$ with $\max_{1 \leq j \leq p} |w_j| \leq R$ for all $w \in \overline{A}$. It is routine to verify that

$$P(S_1^{X_1} \in A^R) \leq C/n^{1/2}.$$ 

Moreover, conditional on $X_1$, $S_{nj}^{X_1}$ is Gaussian with mean zero and variance $\mathbb{E}_n[(X_{ij} - \hat{\mu}_{nj})^2] = \Sigma_{jj}^{X_1}$, so that

$$P(S_1^{X_1} \in A^R | X_1^n) \leq P(\max_{1 \leq j \leq p} |S_{nj}^{X_1}| > R | X_1^n) \leq \frac{C(\log p)^{1/2}}{R} \max_{1 \leq j \leq p} (\Sigma_{jj}^{X_1})^{1/2},$$

which is bounded by $C/n^{1/2}$ on the event $\Delta_{n,r} \leq C_1\Delta_n$. Hence on the event $\Delta_{n,r} \leq C_1\Delta_n$,

$$\left| P(S_1^{X_1} \in A | X_1^n) - P(S_1^{Y_1} \in A) \right| \leq \left| P(S_1^{X_1} \in \overline{A} | X_1^n) - P(S_1^{Y_1} \in \overline{A}) \right| + C/n^{1/2},$$

so that it suffices to only consider the case where $A \in A_{2p}^X(s)$ is such that $\max_{1 \leq j \leq p} |w_j| \leq R$ for all $w \in A$. 


As in the proof of Corollary 3.3, we separately consider two cases. First, suppose that at least one of \( A_q \) does not contain a ball of radius \( \varepsilon \); then \( \cap_{v \in \mathbb{R}^p : u'v \leq S_{n,q}(v) - \varepsilon} \) is empty, so that by condition (M.1') and Lemma A.2, \( P(S_{n,q} \in A_q) \leq C\varepsilon \). Moreover, since \( S_{n,X}^c \) is Gaussian conditional on \( X_1^n \), by condition (M.1') and Lemma A.2, we have, on the event \( \Delta_n,r \leq C_1\Delta_n \), \( P(S_{n,X}^c \in A_q \mid X_1^n) \leq C\varepsilon \). Since \( A \subset A_q \), we conclude that on the event \( \Delta_n,r \leq C_1\Delta_n \), \( |P(S_{n,X}^c \in A \mid X_1^n) - P(S_{n,X}^c \in A)| \leq C\varepsilon = C/n \).

Second, suppose that each \( A_q \) contains a ball with radius \( \varepsilon \). Then by applying Lemma D.1 to each \( A_q \), for \( n \geq n_0 \) and \( m \leq (pm)^d \) with \( d \) depending only on \( s \), we can construct an \( m \)-generated convex set \( A_q^m \) such that \( A_q \subset A_q^m \) with \( \|v\|_0 \leq s \) for all \( v \in V(A_q^m) \). Let \( A_0 = \cap_{q=1}^Q A_q^m \); then \( A \subset A_0 \) and \( \cap_{q=1}^Q A_q^m \) is empty. By the latter fact, together with condition (M.1') and Lemma A.1, we have \( P(S_{n}^Y \in A^0) \leq C(\log p)^{1/2}/n \). Moreover, since \( S_{n,X}^c \) is Gaussian conditional on \( X_1^n \), by condition (M.1') and Lemma A.1, the inequality \( P(S_{n,X}^c \in A^0 \mid X_1^n) \leq C(\log p)^{1/2}/n \) holds on the event \( \Delta_n,r \leq C_1\Delta_n \). Since \( A \subset A^0 \), we conclude that on the event \( \Delta_n,r \leq C_1\Delta_n \), \( |P(S_{n,X}^c \in A \mid X_1^n) - P(S_{n}^Y \in A)| \leq C(\log p)^{1/2}/n \). The final conclusion follows from the fact that \( P(\Delta_n,r > C_1\Delta_n) \leq \alpha/2 \).

**Proof of Theorem 4.2.** By the triangle inequality, \( \rho_n^{EB} \leq \rho_n^{MB} + \varrho_n^{EB} \), where

\[
\varrho_n^{EB} := \sup_{A \in \mathcal{A}_n} \left| P(S_n^{X^*} \in A \mid X_1^n) - P(S_n^{X} \in A \mid X_1^n) \right|.
\]

Also conditional on \( X_1^n \), \( X_1^n - \hat{\mu}_n^X, \ldots, X_p^n - \hat{\mu}_n^X \) are i.i.d. with zero mean and covariance matrix \( \Sigma_{n,X} \). In addition, conditional on \( X_1^n \), \( S_n^{X^*} \equiv \sum_{i=1}^n Y_i^*/\sqrt{n} \), where \( Y_1^*, \ldots, Y_p^* \) are i.i.d. centered Gaussian random vectors with the same covariance matrix \( \Sigma_{n,X} \). Hence the conclusion of the theorem follows from applying Theorem 2.1 conditional on \( X_1^n \) (with \( L_n \) and \( M_n(\phi_n) \) in Theorem 2.1 substituted by \( \hat{L}_n \) and \( \hat{M}_n(\phi_n) \)) to bound \( \varrho_n^{EB} \) on the event \( \{E_n[(X_{ij} - \hat{\mu}_n)^2] \geq b \text{ for all } 1 \leq j \leq p\} \cap \{\hat{L}_n \leq \hat{L}_n\} \cap \{\hat{M}_n(\phi_n) \leq \hat{M}_n\} \).

**Proof of Corollary 4.4.** Here \( c, C \) are constants depending only on \( b, q, K \); their values may change from place to place. We first note that, for sufficiently small \( c > 0 \), we may assume that

\[
B_n^2 \log^7(pm) \leq cn, \quad (39)
\]

since otherwise we can make the assertion of the lemma trivial by setting \( C \) sufficiently large.

Moreover, by the same argument as that used in the proof of Corollary 4.1, the problem reduces to the case of rectangles \( \mathcal{A} = \mathcal{A}^e \); that is, it suffices to prove the bounds (16) and (17) with \( \rho_n^{EB}(\mathcal{A}^e) \) replaced by \( \rho_n^{EB}(\mathcal{A}^e) \) and condition (M.1') replaced by (M.1). For the latter problem, we will apply Theorem 4.2.
Case (E.1). With (39) in mind, by the proof of Corollary 4.1, we see that $P(\Delta_n > b/2) \leq \alpha/6$, so that with probability larger than $1 - \alpha/6$, $b/2 \leq \mathbb{E}_n[(X_{ij} - \hat{\mu}_{ij}^X)^2] \leq CB_n$ for all $j = 1, \ldots, p$. We turn to bounding $\hat{L}_n$. Using the inequality $|a - b|^3 \leq 4(|a|^3 + |b|^3)$ together with Jensen’s inequality, we have

$$
\hat{L}_n \leq 4 \left( \max_{1 \leq j \leq p} \mathbb{E}_n[|X_{ij}|^3] + \max_{1 \leq j \leq p} |\hat{\mu}_{ij}^X|^3 \right) \leq 8 \max_{1 \leq j \leq p} \mathbb{E}_n[|X_{ij}|^3].
$$

By Lemma E.3,

$$
\mathbb{E}[\max_{1 \leq j \leq p} |X_{ij}|^3] \leq C \{ \mathbb{L}_n + n^{-1} \mathbb{E}[\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{ij}|^3] \log p \}
\leq C \{ B_n + n^{-1} B_n^3 \log^4(pn) \}.
$$

Note that $||X_{ij}||_{\psi_1}^3 \leq ||X_{ij}||_{\psi_1}^3 \leq B_n^3$, so that applying Lemma E.4 (i) with $\beta = 1/3$, we have for every $t > 0$,

$$
P(\hat{L}_n \geq C \{ B_n + n^{-1} B_n^3 \log^4(pn) + n^{-1} B_n^3 \} ) \leq 3e^{-t}.
$$

Taking $t = \log(18/\alpha) \leq C \log(pn)$, we conclude that, with $T_n = CB_n$ (recall (39)), $P(\hat{L}_n > T_n) \leq \alpha/6$.

Consider to bound $\hat{M}_{n,X}(\phi_n)$. Observe that

$$
\max_{1 \leq j \leq p} |X_{ij} - \hat{\mu}_{ij}^X| \leq 2 \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |X_{ij}|,
$$

so that

$$
P(\hat{M}_{n,X}(\phi_n) > 0) \leq P(\max_{i,j} |X_{ij}| > \sqrt{n}/(8\phi_n \log p)).
$$

Since $||X_{ij}||_{\psi_1} \leq B_n$, the right hand side is bounded by

$$
2(pn) \exp\{-\sqrt{n}/(8B_n \phi_n \log p)\}.
$$

Observe that

$$
B_n \phi_n \log p \leq Cn^{-1/6} B_n^{2/3} \log^{1/3} p,
$$

so that using (39), we conclude that $P(\hat{M}_{n,Y}(\phi_n) > 0) \leq \alpha/6$. For $\hat{M}_{n,Y}(\phi_n)$, since with probability larger than $1 - \alpha/6$, $\mathbb{E}_n[(X_{ij} - \hat{\mu}_{ij}^X)^2] \leq CB_n$ for all $j = 1, \ldots, p$, on that event, conditional on $X_1, \ldots, X_n$, $||S_{nij}^X||_{\psi_2} \leq CB_n^{1/2}$ for all $j = 1, \ldots, p$. Hence, using the same argument used in bounding $\hat{M}_{n,X}(\phi_n)$, we conclude that

$$
P(\hat{M}_{n,Y}(\phi_n) > 0) \leq \alpha/6 + \alpha/6 = \alpha/3,
$$

which implies that

$$
P(\hat{M}(\phi_n) = 0) \geq 1 - (\alpha/6 + \alpha/3) = 1 - \alpha/2.
$$

Taking these together, by Theorem 4.2, with probability larger than $1 - (\alpha/6 + \alpha/6 + \alpha/2) = 1 - 5\alpha/6$, we have

$$
\rho_n^{EB} \leq \rho_n^{MB} + C\{n^{-1} B_n^2 \log^7(pn)\}^{1/6}.
$$

The final conclusion follows from Corollary 4.1.
Case (E.2). In this case, in addition to (39), we may assume that
\[
\frac{B_n^4 \log^3(pn)}{a^{2/q} n^{1-2/q}} \leq c \leq 1, \tag{40}
\]
since otherwise we can make the assertion of the lemma trivial by setting \( C \) sufficiently large. Then as in the previous case, by the proof of Corollary 4.1, with probability larger than \( 1 - \alpha/6, b/2 \leq E_n[(X_{ij} - \hat{\mu}_{ij})^2] \leq CB_n \) for all \( j = 1, \ldots, p \).

To bound \( \hat{L}_n \), recall that \( \hat{L}_n \leq 8 \max_{1 \leq j \leq p} E_n[|X_{ij}|^3] \), and by Lemma E.3,
\[
E[\max_{1 \leq j \leq p} E_n[|X_{ij}|^3]] \leq C(B_n + B_n^3 n^{-1+3/q} \log p).
\]
Hence by applying Lemma E.4 (ii) with \( s = q/3 \), we have for every \( t > 0 \),
\[
P\{\hat{L}_n \geq C(B_n + B_n^3 n^{-1+3/q} \log p) + n^{-1}\} \leq C t^{-q/3} E[\max_{i,j} |X_{ij}|^q] \leq C t^{-q/3} n B_n^q.
\]
Solving \( C t^{-q/3} n B_n^q = \alpha/6 \), we conclude that \( P(\hat{L}_n \geq \mathcal{T}_n) \leq \alpha/6 \) where \( \mathcal{T}_n = C(B_n + B_n^3 n^{-1+3/q} \alpha^{-3/q} \log p) \).

We turn to bounding \( \tilde{M}_{n,X}(\phi_n) \). As in the previous case,
\[
P\{\tilde{M}_{n,X}(\phi_n) > 0\} \leq P\{\max_{i,j} |X_{ij}| > \sqrt{n}/(8\phi_n \log p)\}.
\]
Since the right hand side is nondecreasing in \( \phi_n \), and \( \phi_n \leq c B_n^{-1} n^{1/2-1/q} \alpha^{1/q} (\log p)^{-1} \), we have (by choosing the constant \( C \) in \( \mathcal{T}_n \) large enough)
\[
P\{\max_{i,j} |X_{ij}| > \sqrt{n}/(8\phi_n \log p)\} \leq n \max_j P\{\max_{i} |X_{ij}| > C B_n n^{1/q} \alpha^{-1/q}\} \leq \alpha/6.
\]
For \( \tilde{M}_{n,Y}(\phi_n) \), we make use of the argument in the previous case, and conclude that
\[
P\{\tilde{M}_{n,Y}(\phi_n) > 0\} \leq \alpha/2.
\]
The rest of the proof is the same as in the previous case. Note that
\[
\left( \frac{E_n \log^7(pn)}{n} \right)^{1/6} \leq C \left[ \left( \frac{B_n^2 \log^7(pn)}{n} \right)^{1/6} + \left( \frac{B_n^2 \log^3(pn)}{a^{2/q} n^{1-2/q}} \right)^{1/2} \right],
\]
and because of (40), the second term inside the bracket on the right hand side is at most
\[
\left( \frac{B_n^2 \log^3(pn)}{a^{2/q} n^{1-2/q}} \right)^{1/3}.
\]
\[\blacksquare\]

Proof of Corollary 4.5. The proof is analogous to that of Corollary 4.2 \[\blacksquare\]
References

[1] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* 13 1000-1034.

[2] Adamczak, R. (2010). A few remarks on the operator norm of random Toeplitz matrices. *J. Theoret. Probab.* 23 85-108.

[3] Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* 10 411-420.

[4] Barvinok, A. (2013). Thrifty approximations of convex bodies by polytopes. *Int. Math. Res. Not.* 00 1-16.

[5] Bentkus, V. (1985). Lower bounds for the rate of convergence in the central limit theorem in Banach spaces. *Lithuanian Mathematical Journal* 25 312-320.

[6] Bentkus, V. (1986). Dependence of the Berry-Esseen estimate on the dimension (in Russian). *Litovsk. Mat. Sb.* 26 205-210.

[7] Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Stat. Plann. Infer.* 113 385-402.

[8] Bhattacharya, R. (1975). On the errors of normal approximation. *Ann. Probab.* 3 815-828.

[9] Bolthausen, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete* 66 379-386.

[10] Borell, C. (1974). Convex measures on locally convex spaces. *Ark. Mat.* 12 239-252.

[11] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

[12] Chatterjee, S. (2005a). A simple invariance theorem. arXiv:math/0508213.

[13] Chatterjee, S. (2006). A generalization of Lindeberg’s principle. *Ann. Probab.* 34 2061-2076.

[14] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* 4 257-283.

[15] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* 41 2786-2819.

[16] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Supplemental Material to “Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors”. *Ann. Statist.* 41 2786-2819.

[17] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014a). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* 42 1564-1597.
Chernozhukov, V., Chetverikov, D. and Kato, K. (2014b). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. To appear in Probab. Theory Related Fields.

Dudley, R.M. (1999). Uniform Central Limit Theorems. Cambridge University Press.

Einmahl, U. and Li, D. (2008). Characterization of LIL behavior in Banach space. Trans. Amer. Math. Soc. 360 6677-6693.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. Ann. Probab. 19 724-739.

Goldstein, L., and Rinott, Y. (1996) Multivariate normal approximations by Stein’s method and size bias couplings. J. Appl. Probab. 33 1-17.

Klivans, A., O’Donnell, R., and Servedio, R. (2008). Learning geometric concepts via Gaussian surface area. 49th Annual IEEE Symposium on Foundations of Computer Science.

Ledoux, M. and Talagrand, M. (1991). Probability in Banach Spaces. Springer.

Massart, P. (2000). About the constants in Talagrand’s concentration inequalities for empirical processes. Ann. Probab. 28 863-884.

Milman, V.D. and Schechtman, G. (1986). Asymptotic Theory of Finite Dimensional Normed Spaces. Lecture Notes in Mathematics 1200 Springer.

Nagaev, S. (1976). An estimate of the remainder term in the multidimensional central limit theorem. Proc. Third Japan-USSR Symp. Probab. Theory. 550 419-438.

Nazarov, F. (2003). On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In: Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics Volume 1807, Springer, pp. 169-187.

Panchenko, D. (2013). The Sherrington-Kirkpatrick Model. Springer.

Praestgaard, J. and Wellner, J.A. (1993). Exchangeably weighted bootstraps of the general empirical processes. Ann. Probab. 21 2053-2086.

Reinert, G. and Rollin, A. (2009). Multivariate normal approximation with Stein’s method of exchangeable pairs under a general linearity condition. Ann. Probab. 37 2150-2173.

Rollin, A. (2011). Stein’s method in high dimensions with applications, Ann. Inst. H. Poincaré Probab. Statist. 49, 529-549.

Sazonov, V. (1968). On the multi-dimensional central limit theorem. Sankhya Ser. A 30 181-204.

Sazonov, V. (1981). Normal Approximations: Some Recent Advances. Lecture Notes in Mathematics Volume 879, Springer.

Senatov, V. (1980). Several estimates of the rate of convergence in the multidimensional CLT. Dokl. Acad. Nauk SSSR 254 809-812.

Sweeting, T. (1977). Speed of convergence for the multidimensional central limit theorem. Ann. Probab. 5 28-41.
[37] Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41** 463-501.

[38] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.

[39] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians.* Springer.

[40] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer.

(V. Chernozhukov) DEPARTMENT OF ECONOMICS AND OPERATIONS RESEARCH CENTER, MIT, 50 MEMORIAL DRIVE, CAMBRIDGE, MA 02142, USA.

E-mail address: vchern@mit.edu

(D. Chetverikov) DEPARTMENT OF ECONOMICS, UCLA, BUNCHE HALL, 8283, 315 PORTOLA PLAZA, LOS ANGELES, CA 90095, USA.

E-mail address: chetverikov@econ.ucla.edu

(K. Kato) GRADUATE SCHOOL OF ECONOMICS, UNIVERSITY OF TOKYO, 7-3-1 HONGO BUNKYO-KU, TOKYO 113-0033, JAPAN.

E-mail address: kkato@e.u-tokyo.ac.jp