The Annals of Statistics
2013, Vol. 41, No. 6, 2852–2876
DOI: 10.1214/13-AOS1170

CONFIDENCE SETS IN SPARSE REGRESSION

BY RICHARD NICKL AND SARA VAN DE GEER

University of Cambridge and ETH Zürich

Dedicated to the memory of Yuri I. Ingster

The problem of constructing confidence sets in the high-dimensional linear model with n response variables and p parameters, possibly $p \geq n$, is considered. Full honest adaptive inference is possible if the rate of sparse estimation does not exceed $n^{-1/4}$, otherwise sparse adaptive confidence sets exist only over strict subsets of the parameter spaces for which sparse estimators exist. Necessary and sufficient conditions for the existence of confidence sets that adapt to a fixed sparsity level of the parameter vector are given in terms of minimal ℓ^2-separation conditions on the parameter space. The design conditions cover common coherence assumptions used in models for sparsity, including (possibly correlated) sub-Gaussian designs.

1. Introduction.

Consider the linear model

$$Y = X\theta + \varepsilon, \tag{1}$$

where X is a $n \times p$ matrix, $\theta \in \mathbb{R}^p$, potentially $p > n$, and where ε is a $n \times 1$ vector consisting of i.i.d. Gaussian noise independent of X, with mean zero and known variance standardised to one. To develop the main ideas, let us assume for the moment that the matrix X consists of i.i.d. $N(0, 1)$ Gaussian entries (X_{ij}), reflecting a prototypical high-dimensional model, such as those encountered in compressive sensing; our main results hold for more general design assumptions that we introduce and discuss in detail below.

We denote by P_θ the law of (Y, X), by E_θ the corresponding expectation, and will omit the subscript θ when no confusion may arise. For the asymptotic analysis we shall let $\min(n, p)$ tend towards infinity, and the o, O-notation is to be understood accordingly. Let $B_0(k)$ be the ℓ^0-"ball" of radius k in $\mathbb{R}^p$, so all vectors in $\mathbb{R}^p$ with at most $k \leq p$ nonzero entries. As common in the literature on high-dimensional models, we shall consider p potentially greater than n but signals θ that are *sparse* in the sense that $\theta \in B_0(k)$ for some k significantly smaller than p. We parameterise k as

$$k \equiv k(\beta) \sim p^{1-\beta}, \qquad 0 < \beta < 1.$$

Received February 2013; revised July 2013.
MSC2010 subject classifications. Primary 62J05; secondary 62G15.
Key words and phrases. Composite testing problem, high-dimensional inference, detection boundary, quadratic risk estimation.

The parameter β measures the sparsity of the signal: if β is close to one, only very few of the p coefficients of θ are nonzero. If $\beta \in (0, 1/2]$, one speaks of the moderately sparse case and for $\beta \in (1/2, 1]$ of the highly sparse case. We include the case $\beta = 1$ where, by convention, $k \equiv \text{const} \times p^0 = \text{const}$.

A sparse adaptive estimator $\hat{\theta} \equiv \hat{\theta}_{np} = \hat{\theta}(Y, X)$ for θ achieves for every n, every $k \leq p$, some universal constant c and with high P_θ-probability, the risk bound

$$\|\hat{\theta} - \theta\|^2 \leq c \log p \times \frac{k}{n}, \tag{2}$$

uniformly for all $\theta \in B_0(k)$. Here $\|\cdot\| \equiv \|\cdot\|_2$ denotes the standard Euclidean norm on $\mathbb{R}^p$, with inner product $\langle \cdot, \cdot \rangle$. Such estimators exist (see Corollary 2 below, for example)—they attain the risk of an estimator that would know the positions of the k nonzero coefficients, with the typically mild penalty of $\log p$. The literature on such estimators is abundant; see, for instance, Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009), and the monograph Bühlmann and van de Geer (2011), where many further references can be found.

We are interested in the question of whether one can construct a confidence set for θ that takes inferential advantage of sparsity as in (2). Most of what follows applies as well to the related problem of constructing confidence sets for $X\theta$—we discuss this briefly at the end of the Introduction. A confidence set $C \equiv C_{np}$ is a random subset of $\mathbb{R}^p$—depending only on the sample Y, X and on a significance level $0 < \alpha < 1$—that we require to contain the true parameter θ with at least a prescribed probability $1 - \alpha$. Our positive results rely on the in many ways natural universal assumption $\theta \in B_0(k_1)$, with k_1 a minimal sparsity degree such that consistent estimation is possible. Formally,

$$k_1 \sim p^{1-\beta_1}, \qquad \beta_1 \in (0, 1); \qquad k_1 = o(n/\log p),$$

so that the risk bound in (2) converges to zero for $k = k_1$. Our statistical procedure should have coverage over signals that are at least k_1-sparse. Given $0 < \alpha < 1$, a level $1 - \alpha$ confidence set C should then be *honest* over $B_0(k_1)$,

$$\liminf_{\min(n,p)\to\infty} \inf_{\theta \in B_0(k_1)} P_\theta(\theta \in C) \geq 1 - \alpha. \tag{3}$$

Moreover, the Euclidean diameter $|C|_2$ of C should satisfy that for every $\alpha' > 0$ there exists a universal constant L such that for every $0 < k \leq k_1$,

$$\limsup_{\min(n,p)\to\infty} \sup_{\theta \in B_0(k)} P_\theta\left(|C|_2^2 > L \log p \times \frac{k}{n}\right) \leq \alpha'. \tag{4}$$

Such a confidence set would cover the true θ with prescribed probability and would shrink at an optimal rate for k-sparse signals without requiring knowledge of the position of the k nonzero coefficients.

A first attempt to construct such a confidence set, inspired by Li (1989), Beran and Dümbgen (1998), Baraud (2004) in nonparametric regression problems, is

based on estimating the accuracy of estimation in (2) directly via sample splitting. Heuristically the idea is to compute a sparse estimator $\tilde{\theta}$ based on the first subsample of (Y, X) and to construct a confidence set centred at $\tilde{\theta}$ based on the risk estimate

$$\frac{1}{n}(Y - X\tilde{\theta})^T(Y - X\tilde{\theta}) - 1$$

based on Y, X from the other subsample.

THEOREM 1. *Consider the model* (1) *with i.i.d. Gaussian design* $X_{ij} \sim N(0,1)$ *and assume* $k_1 = o(n/\log p)$. *There exists a confidence set* C *that is honest over* $B_0(k_1)$ *in the sense of* (3) *and which satisfies, for any* $k \leq k_1$, *and uniformly in* $\theta \in B_0(k)$,

$$|C|_2^2 = O_P\left(\log p \times \frac{k}{n} + n^{-1/2}\right).$$

In fact, we prove Theorem 1 for general correlated designs satisfying Condition 2 below. As a consequence, in such situations full adaptive inference is possible if the rate of sparse estimation in (2) is not desired to exceed $n^{-1/4}$.

One may next look for estimates of $\|\tilde{\theta} - \theta\|$ that have a better accuracy than just of order $n^{-1/4}$. In nonparametric estimation problems this has been shown to be possible; see Hoffmann and Lepski (2002), Juditsky and Lambert-Lacroix (2003), Cai and Low (2006), Robins and van der Vaart (2006), Bull and Nickl (2013). Translated to high-dimensional linear models, the accuracy of these methods can be seen to be of order $p^{1/4}/\sqrt{n}$, which for $p \geq n$ is of larger order of magnitude than $n^{-1/4}$ and hence of limited interest.

Indeed, our results below will show that the rate $n^{-1/4}$ is intrinsic to high-dimensional models: for $p \geq n$ a confidence set that simultaneously satisfies (3) and adapts at *any* rate $\sqrt{(k \log p)/n} = o(n^{-1/4})$ in (4) does not exist. Rather one then needs to remove certain 'critical regions' from the parameter space in order to construct confidence sets. This is so despite the existence of estimators satisfying (2); *the construction of general sparse confidence sets is thus a qualitatively different problem than that of sparse estimation.*

To formalise these ideas, we take the separation approach to adaptive confidence sets introduced in Giné and Nickl (2010), Hoffmann and Nickl (2011), Bull and Nickl (2013) in the framework of nonparametric function estimation. We shall attempt to make honest inference over maximal subsets of $B_0(k_1)$ where k_1 is given a priori as above, in a way that is adaptive over the submodel of sparse vectors θ that belong to $B_0(k_0)$,

$$k_0 \sim p^{1-\beta_0}, \qquad k_0 < k_1, \qquad \beta_0 > \beta_1.$$

By tracking constants in our proofs, we could include $\beta_0 = \beta_1$ too if $k_0 \leq ck_1$ for $c > 0$ a small constant without changing our findings. However, assuming $k_0 = o(k_1)$ results in a considerably cleaner mathematical exposition.

We shall remove those $\theta \in B_0(k_1)$ that are too close in Euclidean distance to $B_0(k_0)$, and consider

$$\tilde{B}_0(k_1, \rho) = \{\theta \in B_0(k_1) : \|\theta - B_0(k_0)\| \geq \rho\}, \tag{5}$$

where $\rho = \rho_{np}$ is a separation sequence and where $\|\theta - Z\| = \inf_{z \in Z} \|\theta - z\|$ for any $Z \subset \mathbb{R}^p$. Thus, if $\theta \notin B_0(k_0)$, we remove the k_0 coefficients θ_j with largest modulus $|\theta_j|$ from θ, and require a lower bound on the ℓ^2-norm of the remaining subvector. In other words, if $|\theta_{(1)}| \leq \cdots \leq |\theta_{(j)}| \leq \cdots \leq |\theta_{(p)}|$ are any order statistics of $\{|\theta_j| : j = 1, \ldots, p\}$, then

$$\|\theta - B_0(k_0)\|^2 = \sum_{j=1}^{p-k_0} \theta_{(j)}^2$$

needs to exceed ρ^2. Defining the new model

$$\Theta(\rho) = B_0(k_0) \cup \tilde{B}_0(k_1, \rho),$$

we now require, instead of (3) and (4), the weaker coverage property

$$\liminf_{\min(n,p)\to\infty} \inf_{\theta \in \Theta(\rho_{np})} P_\theta(\theta \in C_{np}) \geq 1 - \alpha \tag{6}$$

for any $0 < \alpha < 1$, as well as, for some finite constant $L > 0$,

$$\limsup_{\min(n,p)\to\infty} \sup_{\theta \in B_0(k_0)} P_\theta\left(|C_{np}|_2^2 > L \log p \times \frac{k_0}{n}\right) \leq \alpha' \tag{7}$$

and

$$\limsup_{\min(n,p)\to\infty} \sup_{\theta \in \tilde{B}_0(k_1, \rho_{np})} P_\theta\left(|C_{np}|_2^2 > L \log p \times \frac{k_1}{n}\right) \leq \alpha' \tag{8}$$

and search for minimal assumptions on the separation sequence ρ_{np}. Note that any confidence set C that satisfies (3) and (4) also satisfies the above three conditions for any $\rho \geq 0$, so if one can prove the necessity of a lower bound on the sequence ρ_{np}, then one disproves in particular the existence of adaptive confidence sets in the stronger sense of (3) and (4).

The following result describes our findings under the conditions of Theorem 1, but now requiring adaptation to $B_0(k_0)$ at estimation rate $\sqrt{(k_0 \log p)/n}$ faster than $n^{-1/4}$ or, what is the same, assuming

$$k_0 = o(\sqrt{n}/\log p).$$

When specialising to the high-dimensional case $p \geq n$ this automatically forces $\beta_0 > 1/2$. We require coverage over moderately sparse alternatives ($\beta_1 \leq 1/2$); the cases $\beta_1 > 1/2$, $p \leq n$ as well more general design assumptions will be considered below.

THEOREM 2. *Consider the model* (1) *with i.i.d. Gaussian design* $X_{ij} \sim N(0,1)$ *and* $p \geq n$. *For* $0 < \beta_1 \leq 1/2 < \beta_0 \leq 1$ *and* $k_0 < k_1$ *as above assume*

$$k_0 = o(\sqrt{n}/\log p), \qquad k_1 = o(n/\log p).$$

An honest adaptive confidence set C_{np} *over* $\Theta(\rho_{np})$ *in the sense of* (6), (7), (8) *exist if and only if* ρ_{np} *exceeds, up to a multiplicative universal constant,* $n^{-1/4}$, *which is the minimax rate of testing between the composite hypotheses*

$$H_0 : \theta \in B_0(k_0) \quad \textit{vs.} \quad H_1 : \theta \in \tilde{B}_0(k_1, \rho_{np}). \tag{9}$$

The question arises whether insisting on exact rate adaptation in (7) is crucial in Theorem 2 or whether some mild 'penalty' for adaptation (beyond $\log p$) could be paid to avoid separation conditions ($\rho > 0$). The proof of Theorem 2 implies that requiring $|C|_2^2$ in (7) to shrink at any rate $o(n^{-1/2})$ that is possibly slower than $(k_0 \log p)/n$ but still $o((k_1 \log p)/n)$ does not alter the conclusion of necessity of separation at rate $\rho \simeq n^{-1/4}$ in Theorem 2. In particular, for $p \geq n$, Theorem 1 cannot be improved if one wants adaptive confidence sets that are honest over all of $B_0(k_1)$.

Theorem 2 and our further results below show that sparse $o(n^{-1/4})$-adaptive confidence sets exist precisely over those parameter subspaces of $B_0(k_1)$ for which the degree of sparsity is asymptotically detectable. Sparse adaptive confidence sets solve the composite testing problem (9) in a minimax way, either implicitly or explicitly. Theorem 2 reiterates the findings in Hoffmann and Nickl (2011) and Bull and Nickl (2013) that adaptive confidence sets exist over parameter spaces for which the structural property one wishes to adapt to–in the present case, sparsity—can be detected from the sample.

The paper Ingster, Tsybakov and Verzelen (2010), where the testing problem (9) is considered with simple $H_0 : \theta = 0$, is instrumental for our lower bound results. Our upper bounds show that a minimax test for the composite problem (9) exists without requiring stronger separation conditions than those already needed in the case of $H_0 : \theta = 0$, and under general correlated design assumptions. In the setting of Theorem 2 the tests are based on rejecting H_0 if T_n defined by

$$t_n(\theta') = \frac{1}{\sqrt{2n}} \sum_{i=1}^{n} \big[(Y_i - (X\theta')_i)^2 - 1 \big], \qquad T_n = \inf_{\theta' \in B_0(k_0)} |t_n(\theta')| \tag{10}$$

exceeds a critical value. In practice, the computation of T_n requires a convex relaxation of the minimisation problem as is standard in the construction of sparse estimators. The proofs that such minimum tests are minimax optimal are based on ratio empirical process techniques, particularly Lemmas 2 and 3 below, which are of independent interest.

Our results give weakest possible conditions on the regions of the parameter space that have to be removed from consideration in order to obtain sparse adaptive

confidence sets for θ. Another separation condition that may come to mind would be a lower bound γ_{np} on the smallest nonzero entry of $\theta \in B_0(k_1)$. Then

$$\|\theta - B_0(k_0)\|^2 \geq (k_1 - k_0)\gamma_{np}^2$$

and if one considers, for example, moderately sparse $\beta_1 < 1/2$, and $p \geq n, k_0 = o(k_1)$, the lower bound required on γ_{np} for Theorem 2 to apply is in fact $o(n^{-1/2})$. A sparse estimator will not be able to detect nonzero coefficients of such size, rather one needs tailor-made procedures as presented here, and similar in spirit to results in sparse signal detection [Ingster, Tsybakov and Verzelen (2010), Arias-Castro, Candès and Plan (2011)].

Our results concern confidence sets for the parameter vector θ itself in the Euclidean norm $\|\cdot\|$. Often, instead of on θ, inference on $Z\theta$ is of interest, where Z is a $m \times p$ prediction vector. If

$$\|Z\theta\| \geq c\|\theta\| \qquad \forall \theta \in B_0(k_1)$$

with high probability, including the important case $Z = X$ under the usual coherence assumptions on the design matrix X, then any honest confidence set for $Z\theta$ can be used to solve the testing problem (13) below, so that lower bounds for sparse confidence sets for θ carry over to lower bounds for sparse confidence sets for $Z\theta$. In contrast, for regular fixed linear functionals of θ such as low-dimensional projections, the situation may be different: for instance, in the recent papers of Zhang and Zhang (2011), van de Geer, Bühlmann and Ritov (2013) and Javanmard and Montanari (2013) one-dimensional confidence intervals for a fixed element θ_j in the vector θ are constructed.

2. Main results. A heuristic summary of our findings for all parameters simultaneously is as follows: if the rate of estimation in the submodel $B_0(k_0)$ of $B_0(k_1)$ one wishes to adapt to is faster than

$$\rho \simeq \min\left(n^{-1/4}, \frac{p^{1/4}}{\sqrt{n}}, \sqrt{\frac{k_1 \log p}{n}}\right), \tag{11}$$

then separation is necessary for adaptive confidence sets to exist at precisely this rate ρ. For $p \geq n$ this simply reduces to requiring that the rate of adaptive estimation in $B_0(k_0)$ beats $n^{-1/4}$—the natural condition expected in view of Theorem 1, which proves existence of honest adaptive confidence sets when the estimation rate is $O(n^{-1/4})$.

We consider the following conditions on the design matrix X.

CONDITION 1. Consider the model (1) with independent and identically distributed (X_{ij}) satisfying $EX_{ij} = 0$, $EX_{ij}^2 = 1$ $\forall i, j$.

(a) For some $h_0 > 0$,

$$\max_{1 \leq j \leq l \leq p} E\big(\exp(hX_{1j}X_{1l})\big) = O(1) \qquad \forall |h| \leq h_0.$$

(b) $|X_{ij}| \le b$ for some $b > 0$ and all i, j.

Let next $\hat{\Sigma} := X^T X/n$ denote the Gram matrix and let $\Sigma := E\hat{\Sigma}$. We will sometimes write $\|X\theta\|_n^2 := \theta^T \hat{\Sigma}\theta$ to expedite notation.

CONDITION 2. In the model (1) assume the following:

(a) The matrix X has independent rows, and for each $i \in \{1, \dots, n\}$ and each $u \in \mathbb{R}^p$ with $u^T \Sigma u \le 1$, the random variable $(Xu)_i$ is sub-Gaussian with constants σ_0 and κ_0:

$$\kappa_0^2\big(E\exp\big[|(Xu)_i|^2/\kappa_0^2\big] - 1\big) \le \sigma_0^2 \qquad \forall u^T \Sigma u \le 1.$$

(b) The smallest eigenvalue $\Lambda_{\min}^2 \equiv \Lambda_{\min,p}^2$ of Σ satisfies $\inf_p \Lambda_{\min,p}^2 > 0$.

Condition 1(a) could be replaced by a fixed design assumption as in Remark 4.1 in Ingster, Tsybakov and Verzelen (2010). Condition 1(b) clearly implies Condition 1(a); it also implies Condition 2 with $\Sigma = I$ and universal constants κ_0, σ_0: we have $(Xu)_i = \sum_{m=1}^p X_{im}u_m$ with mean zero and independent summands bounded in absolute value by $b|u_m|$, so that by Hoeffding's inequality $(Xu)_i$ is sub-Gaussian,

$$P\big(|(Xu)_i| \ge t\big) \le 2e^{-t^2/2b\|u\|_2^2}$$

and Condition 2 follows, integrating tail probabilities.

2.1. *Adaptation to sparse signals when $p \ge n$.* We first give a version of Theorem 2 for general (not necessarily Gaussian) design matrices. The proofs imply that part (B) actually holds also for $p \le n$ and for $0 < \beta_1 < \beta_0 \le 1$.

THEOREM 3 (Moderately sparse case). *Let $p \ge n$, $0 < \beta_1 \le 1/2 < \beta_0 \le 1$ and let $k_0 \sim p^{1-\beta_0} < k_1 \sim p^{1-\beta_1}$ such that $k_0 = o(\sqrt{n}/\log p)$.*

(A) Lower bound. *Assume Condition 1(a) and that $\log^3 p = o(n)$. Suppose for some separation sequence $\rho_{np} \ge 0$ and some $0 < \alpha, \alpha' < 1/3$, the confidence set C_{np} is both honest over $\Theta(\rho_{np})$ and adapts to sparsity in the sense of (7), (8). Then necessarily*

$$\liminf_{n,p} \frac{\rho_{np}}{n^{-1/4}} > 0.$$

(B) Upper bound. *Assume Condition 2 and $k_1 = o(n/\log p)$. Then for every $0 < \alpha, \alpha' < 1$ there exists a sequence $\rho_{np} \ge 0$ satisfying*

$$\limsup_{n,p} \frac{\rho_{np}}{n^{-1/4}} < \infty$$

and a level α-confidence set C_{np} that is honest over $\Theta(\rho_{np})$ and that adapts to sparsity in the sense of (7), (8).

We next consider restricting the maximal parameter space itself to highly sparse $\theta \in B_0(k_1)$, $\beta_1 > 1/2$. If the rate of estimation in $B_0(k_1)$ accelerates beyond $n^{-1/4}$, then one can take advantage of this fact, although separation of $B_0(k_0)$ and $B_0(k_1)$ is still necessary to obtain sparse adaptive confidence sets. The following result holds also for $p \le n$.

THEOREM 4 (Highly sparse case). *Let* $1/2 < \beta_1 < \beta_0 \le 1$ *and let* $k_0 \sim p^{1-\beta_0} < k_1 \sim p^{1-\beta_1}$ *such that* $k_0 = o(\sqrt{n}/\log p)$.

(A) Lower bound. *Assume Condition* 1(a) *and that* $\log^3 p = o(n)$. *Suppose for some separation sequence* $\rho_{np} \ge 0$ *and some* $0 < \alpha, \alpha' < 1/3$, *the confidence set* C_{np} *is both honest over* $\Theta(\rho_{np})$ *and adapts to sparsity in the sense of* (7), (8). *Then necessarily*

$$\liminf_{n,p} \frac{\rho_{np}}{\min(\sqrt{\log p \times (k_1/n)}, n^{-1/4})} > 0.$$

(B) Upper bound. *Assume Condition* 2 *and that* $k_1 = o(n/\log p)$. *Then for every* $0 < \alpha', \alpha < 1$ *there exists a sequence* $\rho_{np} \ge 0$ *satisfying*

$$\limsup_{n,p} \frac{\rho_{np}}{\min(\sqrt{\log p \times (k_1/n)}, n^{-1/4})} < \infty$$

and a level α*-confidence set* C_{np} *that is honest over* $\Theta(\rho_{np})$ *and that adapts to sparsity in the sense of* (7), (8).

2.2. *The case* $p \le n$*—approaching nonparametric models*. The case of highly sparse alternatives and $p \le n$ was already considered in Theorem 4, explaining the presence of $\sqrt{(k_1 \log p)/n}$ in (11). We thus now restrict to $0 < \beta_1 \le 1/2$ and, moreover, to highlight the main ideas, also to $\beta_0 > 1/2$ corresponding to the most interesting highly sparse null-models. We now require from any confidence set C_n the conditions (6), (7), (8) with the infimum/supremum there intersected with

$$B_r(M) = \left\{ \theta \in \mathbb{R}^p : \|\theta\|_r^r = \sum_{j=1}^p |\theta_j|^r \le M^r \right\}.$$

Let us denote the new conditions by (6′), (7′), (8′).

THEOREM 5. *Assume* $p \le n$, *let* $0 < \beta_1 \le 1/2 < \beta_0 \le 1$, $0 < M < \infty$, *and let* $k_0 \sim p^{1-\beta_0} < k_1 \sim p^{1-\beta_1}$.

(A) Lower bound. *Assume Condition* 1(a), *and suppose for some separation sequence* $\rho_{np} \ge 0$ *and some* $0 < \alpha, \alpha' < 1/3$, *the confidence set* C_{np} *is both honest over* $\Theta(\rho_{np}) \cap B_r(M)$ *and adapts to sparsity in the sense of* (7′), (8′). *If* $r = 2$ *or if* $r = 1$ *and* $p = O(n^{2/3})$, *then necessarily*

$$\liminf_{n,p} \frac{\rho_{np}}{p^{1/4} n^{-1/2}} > 0.$$

(B) Lower bound. *Assume Condition* 1(b) *holds and either* $r = 1$, $k_1 = o(n/\log p)$ *or* $r = 2$, $\beta_0 = 1$, $k_1 = o(\sqrt{n/\log p})$. *Then for every* $0 < \alpha, \alpha' < 1$ *there exists a sequence* $\rho_{np} \geq 0$ *satisfying*

$$\limsup_{n,p} \frac{\rho_{np}}{p^{1/4} n^{-1/2}} < \infty$$

and a level α*-confidence set* $C \equiv C(n, p, b, M)$ *that is honest over* $\Theta(\rho_{np}) \cap \{\theta : \|\theta\|_r \leq M\}$ *and that adapts to sparsity in the sense of* (7′), (8′).

The rate ρ in the previous theorems is related to the results in Bull and Nickl (2013) and approaches, for $p = \text{const}$, the parametric theory, where the separation rate equals, quite naturally, $1/\sqrt{n}$. This is in line with the findings in Pötscher (2009), Pötscher and Schneider (2011) in the $p \leq n$ setting, who point out that a class of specific but common sparse estimators cannot reliably be used for the construction of confidence sets.

3. Proofs. All lower bounds are proved in Section 3.1. The proofs of existence of confidence sets are given in Section 3.2. Theorem 1 is proved at the end, after some auxiliary results that are required throughout.

3.1. *Proof of Theorems* 2 (*necessity*), 3(A), 4(A), 5(A). The necessity part of Theorem 2 follows from Theorem 3(A) since any i.i.d. Gaussian matrix satisfies Condition 1(a), and since its assumptions imply the growth condition $\log^3 p = o(n)$. Except for the ℓ^r-norm restrictions of Theorem 5 discussed at the end of the proof, Theorems 3(A) and 5(A) can be joined into a single statement with separation sequence $\min(p^{1/4} n^{-1/2}, n^{-1/4})$, valid for every p. We thus have to consider, for all values of p, two cases: the moderately sparse case $\beta_1 < 1/2$ with separation lower bound $\min(p^{1/4} n^{-1/2}, n^{-1/4})$, and the highly sparse case $\beta_1 > 1/2$ with separation lower bound $\min((\log p \times (k_1/n))^{1/2}, n^{-1/4})$. Depending on the case considered, denote thus by $\rho^* = \rho^*_{np}$ either $\min((\log p \times (k_1/n))^{1/2}, n^{-1/4})$ or $\min(p^{1/4} n^{-1/2}, n^{-1/4})$.

The main idea of the proof follows the mechanism introduced in Hoffmann and Nickl (2011). Suppose by way of contradiction that C is a confidence set as in the relevant theorems, for some sequence $\rho = \rho_{np}$ such that

$$\liminf_{n,p} \frac{\rho}{\rho^*} = 0.$$

By passing to a subsequence, we may replace the lim inf by a proper limit, and we shall in what follows only argue along this subsequence $n_k \equiv n$. We claim that we can then find a further sequence $\bar{\rho}_{np} \equiv \bar{\rho}$, $\rho^*_{np} \geq \bar{\rho}_{np} \geq \rho_{np}$, s.t.

$$\sqrt{\log p \times \frac{k_0}{n}} = o(\bar{\rho}), \qquad \bar{\rho} = o(\rho^*), \tag{12}$$

that is, $\bar{\rho}$ can be taken to be squeezed between the rate of adaptive estimation in the submodel $B_0(k_0)$ and the separation rate ρ^* that we want to establish as a lower bound. To check that this is indeed possible, we need to verify that $(\log p \times (k_0/n))^{1/2}$ is of smaller order than any of the three terms

$$\sqrt{\log p \times \frac{k_1}{n}}, \qquad p^{1/4}n^{-1/2}, \qquad n^{-1/4}$$

appearing in ρ^*. This is obvious for the first in view of the definition of k_0, k_1 ($\beta_1 < \beta_0$); follows for the second from $\beta_0 > 1/2$; and follows for the third from our assumption $k_0 = o(\sqrt{n}/\log p)$ [automatically verified in Theorem 5(A) as $p \le n$, $\beta_0 > 1/2$].

For such a sequence $\bar{\rho}$ consider testing

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \in \tilde{B}_0(k_1, \bar{\rho}).$$

Using the confidence set C, we can test H_0 by $\Psi = 1\{C \cap H_1 \neq \varnothing\}$—we reject H_0 if C contains any of the alternatives. The type two errors satisfy

$$\sup_{\theta \in H_1} E_\theta(1 - \Psi) = \sup_{\theta \in H_1} P_\theta(C \cap H_1 = \varnothing) \le \sup_{\theta \in H_1} P_\theta(\theta \notin C) \le \alpha + o(1)$$

by coverage of C over $H_1 \subset \Theta(\rho)$ (recall $\bar{\rho} \ge \rho$). For the type one errors we have, again by coverage, since $0 \in B_0(k_0)$ for any k_0, using adaptivity (7) and (12), that

$$E_0\Psi = P_0(C \cap H_1 \neq \varnothing) \le P_0\big(0 \in C, |C|_2 \ge \bar{\rho}\big) + \alpha + o(1) = \alpha' + \alpha + o(1).$$

We conclude from $\min(\alpha', \alpha) < 1/3$ that

$$E_0\Psi + \sup_{\theta \in H_1} E_\theta(1 - \Psi) \le \alpha' + 2\alpha + o(1) < 1 + o(1). \tag{13}$$

On the other hand, we now show

$$\liminf_{n,p} \inf_{\Psi} \Big(E_0\Psi + \sup_{\theta \in H_1} E_\theta(1 - \Psi) \Big) \ge 1, \tag{14}$$

a contradiction, so that

$$\liminf_{n,p} \frac{\rho}{\rho^*} > 0$$

necessarily must be true. Our argument proceeds by deriving (14) from Theorem 4.1 in Ingster, Tsybakov and Verzelen (2010). Let $0 < c < 1$, $b = \frac{\bar{\rho}}{c\sqrt{k_1}}$, $h = \frac{ck_1}{p}$, and note that

$$b^2 ph = \frac{\bar{\rho}^2}{c} \ge \bar{\rho}^2, \qquad b^2 k_0 = o\big(b^2 ph\big) \tag{15}$$

using that $k_0 = o(k_1)$. Consider a product prior π on θ with marginal coefficients $\theta_j = b\varepsilon_j$, $j = 1, \ldots, p$, where the ε_j are i.i.d. with $P(\varepsilon_j = 0) = 1 - h$, $P(\varepsilon_j =$

$1) = P(\varepsilon_j = -1) = h/2$. We show that this prior asymptotically concentrates on our alternative space $H_1 = \tilde{B}_0(k_1, \bar{\rho})$. Let $Z_j = \varepsilon_j^2$ and denote by $Z_{(j)}$ the corresponding order statistics (counting ties in any order, for instance, ranking numerically by dimension), then for any $\delta > 0$ and n large enough, using (15),

$$\begin{aligned}\pi\big(\|\theta - B_0(k_0)\|^2 < (1+\delta)\bar{\rho}^2\big) &= P\left(b^2 \sum_{j=1}^{p-k_0} Z_{(j)} < (1+\delta)\bar{\rho}^2\right) \\ &\le P\left(b^2 \sum_{j=1}^{p} Z_{(j)} < (1+\delta)\bar{\rho}^2 - b^2 k_0\right) \\ &\le P\left(b^2 \sum_{j=1}^{p} \varepsilon_j^2 < \bar{\rho}^2\right) = \pi\big(\|\theta\|^2 < \bar{\rho}^2\big),\end{aligned}$$

which by the proof of Lemma 5.1 in Ingster, Tsybakov and Verzelen (2010) converges to 0 as $\min(n, p) \to \infty$. Moreover, that lemma also contains the proof that $\pi(\theta \in B_0(k_1)) \to 1$ (identifying k there with our k_1), which thus implies $\pi(\tilde{B}_0(k_1, \bar{\rho})) \to 1$ as $\min(n, p) \to \infty$. The testing lower bound based on this prior, derived in Theorem 4.1 in Ingster, Tsybakov and Verzelen (2010) (cf. particularly page 1487), then implies (14), which is the desired contradiction. Finally, for Theorem 5, note that the above implies immediately that $\theta \sim \pi$ asymptotically concentrates on any fixed ℓ^2-ball. Moreover, $E_\pi \|\theta\|_1 = bph = o(1)$ under the hypotheses of Theorem 5 when $p = O(n^{2/3})$, and likewise $\mathrm{Var}_\pi(\|\theta\|_1) = b^2 ph$, so we conclude as in the proof of Lemma 5.1 in Ingster, Tsybakov and Verzelen (2010) that the prior asymptotically concentrates on any fixed ℓ^1-ball in this situation.

3.2. *Proofs of upper bounds*: *Theorems* 2 (*sufficiency*), 3(B), 4(B), 5(B). We first note that sufficiency in Theorem 2 follows from Theorem 3(B) as i.i.d. Gaussian design satisfies Condition 2. The main idea, which is the same for all theorems, follows Hoffmann and Nickl (2011), Bull and Nickl (2013) to solve the composite testing problem

$$H_0 : \theta \in B_0(k_0) \quad \text{vs.} \quad H_1 : \theta \in \tilde{B}_0(k_1, \rho) \tag{16}$$

under the pa rameter constellations of k_0, k_1, ρ, p, n relevant in Theorems 3(B), 4(B), 5(B) [and in the last case with both hypotheses intersected with $B_r(M)$, suppressed in the notation in what follows]. Once a minimax test Ψ is available for which type one and type two errors

$$\sup_{\theta \in H_0} E_\theta \Psi_n + \sup_{\theta \in H_1} E_\theta (1 - \Psi_n) \le \gamma \tag{17}$$

can be controlled, for n large enough, at any level $\gamma > 0$, one takes $\tilde{\theta}$ to be the estimator from (30) below with λ chosen as in Lemma 4, and constructs the confidence

set

$$C_n = \begin{cases} \left\{\theta : \|\theta - \tilde{\theta}\|_2 \le L' \sqrt{\log p \frac{k_0}{n}}\right\}, & \text{if } \Psi_n = 0, \\ \left\{\theta : \|\theta - \tilde{\theta}\|_2 \le L' \sqrt{\log p \frac{k_1}{n}}\right\}, & \text{if } \Psi_n = 1. \end{cases}$$

Assuming (17), we now prove that C_n is honest for $B_0(k_0) \cup \tilde{B}_0(k_1, \rho_{np})$ if we choose the constant L' large enough: for $\theta \in B_0(k_0)$ we have from Corollary 2 below, for L' large,

$$\inf_{\theta \in B_0(k_0)} P_\theta\{\theta \in C_n\} \ge 1 - \sup_{\theta \in B_0(k_0)} P_\theta\left\{\|\tilde{\theta} - \theta\|_2 > L' \sqrt{\log p \frac{k_0}{n}}\right\} \to 1$$

as $n \to \infty$. When $\theta \in \tilde{B}_0(k_1, \rho_{np})$, we have that $P_\theta\{\theta \in C_n\}$ exceeds

$$1 - \sup_{\theta \in B_0(k_1)} P_\theta\left\{\|\tilde{\theta} - \theta\|_2 > L' \sqrt{\log p \frac{k_1}{n}}\right\} - \sup_{\theta \in \tilde{B}_0(k_1, \rho_{np})} P_\theta\{\Psi_n = 0\}.$$

The first subtracted term converges to zero for L' large enough, as before. The second subtracted term can be made less than $\gamma = \alpha$, using (17). This proves that C_n is honest. We now turn to sparse adaptivity of C_n: by the definition of C_n we always have $|C_n| \le L' \sqrt{\log p \times k_1/n}$, so the case $\theta \in \tilde{B}_0(k_1, \rho_{np})$ is proved. If $\theta \in B_0(k_0)$, then

$$P_\theta\left\{|C_n| > L' \sqrt{\log p \frac{k_0}{n}}\right\} = P_\theta\{\Psi_n = 1\} \le \alpha'$$

by the bound on the type one errors of the test, completing the reduction of the proof to (17).

3.2.1. *Proof of Theorem* 3(B). Throughout this subsection we impose the assumptions from Theorem 3—in fact, without the restriction $p \ge n$—and with $\rho_{np} \ge L_0 n^{-1/4}$ for some L_0 large enough that we will choose below. By the arguments from the previous subsection, it suffices to solve the testing problem (17) with this choice of ρ, for any $\gamma > 0$. Define $t_n(\theta')$, T_n as in (10) and the test $\Psi_n = 1\{T_n \ge u_\gamma\}$ where u_γ is a suitable fixed quantile constant such that, for every $\theta \in B_0(k_0)$, the type one error $E_\theta \Psi_n$ is bounded by

$$P_\theta(T_n \ge u_\gamma) \le P_\theta\big(|t_n(\theta)| \ge u_\gamma\big) = P_\theta\left(\frac{1}{\sqrt{2n}} \sum_{i=1}^n (\varepsilon_i^2 - 1) \ge u_\gamma\right) \le \gamma. \tag{18}$$

For the type two errors $\theta \in H_1$, let θ^* be a minimiser in T_n (if the infimum is not attained, the argument below requires obvious modifications). Then

$$\begin{aligned}\sqrt{2n}t_n(\theta^*) &= \sum_{i=1}^n [(Y_i - (X\theta^*)_i)^2 - 1] \\ &= \sum_{i=1}^n [(Y_i - (X\theta)_i + (X\theta)_i - (X\theta^*)_i)^2 - 1] \\ &= \sum_{i=1}^n [(Y_i - (X\theta)_i)^2 - 1] + 2\langle Y - X\theta, X(\theta - \theta^*)\rangle + \|X(\theta - \theta^*)\|^2,\end{aligned}$$

so the type two errors $E_\theta(1 - \Psi_n)$ are controlled by

$$\begin{aligned}&P_\theta\left(\left|\sum_{i=1}^n [(Y_i - (X\theta)_i)^2 - 1] + 2\langle Y - X\theta, X(\theta - \theta^*)\rangle \right.\right. \\ &\qquad \left.\left. + \|X(\theta - \theta^*)\|^2\right| < \sqrt{2n}u_\gamma\right) \\ &\le P_\theta\left(\left|\sum_{i=1}^n (\varepsilon_i^2 - 1)\right| > \frac{\|X(\theta - \theta^*)\|^2}{2} - \sqrt{n}u_\gamma\right) \\ &\quad + P_\theta\left(|2\langle \varepsilon, X(\theta - \theta^*)\rangle| > \frac{\|X(\theta - \theta^*)\|^2}{2} - \sqrt{n}u_\gamma\right).\end{aligned} \tag{19}$$

Since $\theta^* \in B_0(k_0)$, $\theta \in \tilde{B}_0(k_1, \rho)$ and $k_0 + k_1 = o(n/\log p)$, we have, from Corollary 1 below with $t = (k_0 + k_1)\log p$ that, for n large enough and with probability at least $1 - 4e^{-(k_0+k_1)\log p} \to 1$,

$$\|X(\theta - \theta^*)\|^2 \ge \inf_{\theta' \in H_0} \|X(\theta - \theta')\|^2 \ge c(\Lambda_{\min})n\rho_{np}^2 \ge L'\sqrt{n} \tag{20}$$

for every $L' > 0$ (choosing L_0 large enough). We thus restrict to this event. The probability in the last but one line of (19) is then bounded by

$$P_\theta\left(\left|\sum_{i=1}^n (\varepsilon_i^2 - 1)\right| > \sqrt{n}(L' - u_\gamma)\right)$$

for n large enough, which can be made as small as desired by choosing $L' \ge 4u_\gamma$, as in (18). Likewise, the last probability in the display (19) is bounded, for n large enough, by

$$P_\theta\left(|2\langle \varepsilon, X(\theta - \theta^*)\rangle| > \frac{\|X(\theta - \theta^*)\|^2}{4}\right) \le P_\theta\left(\sup_{\theta' \in H_0} \frac{2|\langle \varepsilon, X(\theta - \theta')|\rangle}{\|X(\theta - \theta')\|^2} > \frac{1}{4}\right),$$

which converges to zero for large enough separation constant L_0, uniformly in $\tilde{B}_0(k_1, \rho)$, proved in Lemma 2 below [using the lower bound (20) for $\|X(\theta - \theta')\|$ and that $\sqrt{k_0 \log p/n} = o(n^{-1/4})$].

3.2.2. *Proof of Theorem* 4(B). Throughout this subsection we impose the assumptions from Theorem 4(B), with ρ_{np} exceeding $L_0\sqrt{(k_1/n)\log p}$ for some L_0 large enough that we will choose below (the $n^{-1/4}$-regime was treated already in Theorem 3(B), whose proof holds for all p). By the arguments from the beginning of Section 3.2, it suffices to solve the testing problem (17) with this choice of ρ, for any level $\gamma > 0$. Let $\tilde{\theta}$ be the estimator from (30) below with λ chosen as in Corollary 2 below, and define the test statistic

$$T_n = \inf_{\theta \in B_0(k_0)} \|\tilde{\theta} - \theta\|^2, \qquad \Psi_n = 1\left\{T_n \geq D \log p \frac{k_1}{n}\right\}$$

for D to be chosen. The type one errors satisfy, uniformly in $\theta \in H_0$, for D large enough,

$$E_\theta \Psi_n \leq P_\theta\left(\|\tilde{\theta} - \theta\|^2 \geq D \log p \frac{k_1}{n}\right) \to 0$$

as $\min(p, n) \to \infty$, by Corollary 2. Likewise, we bound $E_\theta(1 - \Psi_n)$ under $\theta \in \tilde{B}_0(k_1, \rho)$, for some $\theta^* \in B_0(k_0)$, by the triangle inequality,

$$\begin{aligned} P_\theta\left(\|\tilde{\theta} - \theta^*\|_2^2 < C \log p \frac{k_1}{n}\right) &\leq P_\theta\left(\|\tilde{\theta} - \theta\| > \|\theta^* - \theta\| - \sqrt{C \log p \frac{k_1}{n}}\right) \\ &\leq P_\theta\left(\|\tilde{\theta} - \theta\|^2 \geq (L_0 - C) \log p \frac{k_1}{n}\right) \to 0 \end{aligned}$$

for L_0 large enough, again by Corollary 2 below.

3.2.3. *Proof of Theorem* 5(B). Throughout this subsection we impose the assumptions from Theorem 5(B), with $\rho_{np} \geq L_0 p^{1/4}/\sqrt{n}$ for some L_0 large enough that we will choose below. By the arguments from the beginning of Section 3.2, it suffices to solve the testing problem (17) [with both hypotheses there intersected with $B_r(M)$] for this choice of ρ and any level $\gamma > 0$. For $\theta' \in \mathbb{R}^p$ we define the U-statistic

$$U_n(\theta') = \frac{2}{n(n-1)} \sum_{i<k} \sum_{j=1}^{p} (Y_i X_{ij} - \theta'_j)(Y_k X_{kj} - \theta'_j),$$

which equals $\|n^{-1} X^T Y - \theta'\|^2$ with diagonal terms ($i = k$) removed. Then

$$\begin{aligned} &\frac{1}{n} E_\theta X^T Y = E_\theta\left(\frac{1}{n} X^T X\right)\theta = \theta, \\ &E_\theta Y_1 X_{1j} = \theta_j, \qquad E_\theta U_n(\theta') = \|\theta - \theta'\|^2 \end{aligned} \tag{21}$$

and we define the test statistic and test as

$$T_n = \inf_{\theta' \in B_0(k_0)} |U_n(\theta')|, \qquad \Psi_n = 1\left\{T_n \geq u_\gamma \frac{\sqrt{p}}{n}\right\}$$

for u_γ quantile constants specified below. For the type one errors we have, uniformly in H_0, by Chebyshev's inequality

$$E_\theta \Psi_n = P_\theta\left(T_n \geq u_\gamma \frac{\sqrt{p}}{n}\right) \leq P_\theta\left(|U_n(\theta)| \geq u_\gamma \frac{\sqrt{p}}{n}\right) \leq \frac{\mathrm{Var}(U_n(\theta))}{u_\gamma^2}\frac{n^2}{p}. \tag{22}$$

Under P_θ the U-statistic $U_n(\theta)$ is fully centered [cf. (21)], and by standard U-statistic arguments the variance can be bounded by $\mathrm{Var}_\theta(U_n(\theta)) \leq Dp/n^2$ for some constant D depending only on M and $\max_{1\leq j\leq p} EX_{1j}^4 \leq b^4$; see, for instance, display (6.6) in Ingster, Tsybakov and Verzelen (2010) and the arguments preceding it. We can thus choose $u_\gamma = u_\gamma(M, b)$ to control the type one errors in (22).

We now turn to the type two errors and assume $\theta \in \tilde{B}_0(k_1, \rho)$: let θ^* be a minimiser in T_n, then $U_n(\theta^*)$ has Hoeffding decomposition $U_n(\theta^*) = U_n(\theta) + 2L_n(\theta^*) + \|\theta^* - \theta\|^2$ with the linear term

$$L_n(\theta') = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^p (\theta_j - Y_i X_{ij})(\theta_j - \theta_j').$$

We can thus bound the type two errors $E_\theta(1 - \Psi_n)$ as follows:

$$\begin{aligned} P_\theta\left(T_n < u_\gamma \frac{\sqrt{p}}{n}\right) &\leq P_\theta\left(|U_n(\theta)| + 2|L_n(\theta^*)| \geq \|\theta - \theta^*\|^2 - u_\gamma \frac{\sqrt{p}}{n}\right) \\ &\leq P_\theta\left(|U_n(\theta)| \geq \frac{\|\theta - \theta^*\|^2}{2} - u_\gamma \frac{\sqrt{p}}{2n}\right) \\ &\quad + P_\theta\left(|L_n(\theta^*)| \geq \frac{\|\theta - \theta^*\|^2}{4} - u_\gamma \frac{\sqrt{p}}{4n}\right). \end{aligned}$$

By hypothesis on ρ_{np} we can find L_0 large enough such that $\|\theta - \theta^*\|^2 \geq \inf_{\theta' \in H_0} \|\theta - \theta'\|^2 \geq L\sqrt{p}/n$ for any $L > 0$, so that the first probability in the previous display can be bounded by $P_\theta(|U_n(\theta)| > u_\gamma \sqrt{p}/n)$, which involves a fully centered U-statistic and can thus be dealt with as in the case of type one errors. The critical term is the linear term, which, by the above estimate on $\|\theta - \theta^*\|$, is less than or equal to

$$P_\theta\left(|L_n(\theta^*)| \geq \frac{\|\theta - \theta^*\|^2}{8}\right) \leq P_\theta\left(\sup_{\theta' \in H_0} \frac{|L_n(\theta')|}{\|\theta - \theta'\|^2} > \frac{1}{8}\right).$$

The process $L_n(\theta')$ can be written as

$$\begin{aligned} \langle \theta - n^{-1}X^T Y, \theta - \theta' \rangle &= \langle \theta - n^{-1}X^T X\theta, \theta - \theta' \rangle - \langle n^{-1}X^T \varepsilon, \theta - \theta' \rangle \\ &= \frac{1}{n}\langle (E_\theta X^T X - X^T X)\theta, \theta - \theta' \rangle - \frac{1}{n}\langle \varepsilon, X(\theta - \theta') \rangle \\ &\equiv L_n^{(1)}(\theta') + L_n^{(2)}(\theta') \end{aligned}$$

and we can thus bound the last probability by

$$P_\theta\Big(\sup_{\theta'\in H_0}\frac{|L_n^{(1)}(\theta')|}{\|\theta-\theta'\|^2}>\frac{1}{16}\Big)+P_\theta\Big(\sup_{\theta'\in H_0}\frac{|L_n^{(2)}(\theta')|}{\|\theta-\theta'\|^2}>\frac{1}{16}\Big). \tag{23}$$

To show that the probability involving the second process approaches zero, it suffices to show that

$$P_\theta\Big(\sup_{\theta'\in H_0}\frac{|\varepsilon^T X(\theta-\theta')/n|}{\|X(\theta-\theta')\|^2/n}>\frac{1}{16\Lambda}\Big) \tag{24}$$

converges to zero, using that $\sup_{v\in B_0(k_1)}\|Xv\|_2^2/(n\|v\|_2^2)\le\Lambda$ for some $0<\Lambda<\infty$, on events of probability approaching one, by Lemma 1 [noting $k_0+k_1=o(n/\log p)$]. By Lemma 2 this last probability approaches zero as $\min(n,p)\to\infty$, for L_0 large enough, noting that the lower bound on R_t there is satisfied for our separation sequence ρ_{np}, by Corollary 1 and since $(k_0/n)\log p=o(p^{1/2}/n)$ in view of $\beta_0>1/2$. Likewise, using the preceding arguments with Lemma 3 instead of Lemma 2, the probability involving the first process also converges to zero, which completes the proof.

3.3. *Remaining proofs.*

LEMMA 1. *Assume Condition* 2(a) *and denote by* P *the law of* X. *Let* $\theta\in B_0(k_1)$ *and* $k\in\{1,\dots,p\}$. *Then for some constants* σ *and* κ *depending only on* σ_0 *and* κ_0, $C_{k,k_1,p}\equiv(k+k_1+1)\log(25p)$ *and for all* $t>0$,

$$P\Big(\sup_{\theta'\in B_0(k),(\theta'-\theta)^T\Sigma(\theta'-\theta)\neq0}\Big|\frac{(\theta'-\theta)^T\hat\Sigma(\theta'-\theta)}{(\theta'-\theta)^T\Sigma(\theta'-\theta)}-1\Big| \ge 4\sigma\sqrt{\frac{t+C_{k,k_1,p}}{n}}+4\kappa\frac{t+C_{k,k_1,p}}{n}\Big)\le 4\exp[-t].$$

COROLLARY 1. *Let* X *satisfy Conditions* 2(a) *and* 2(b). *Let* σ, κ, θ, k, k_1, $C_{k,k_1,p}$ *be defined as in Lemma* 1. *Suppose that* k, k_1 *and* $t>0$ *are such that*

$$\Big(\frac{8C_{k,k_1,p}}{n}\vee\frac{8t}{n}\Big)\le\Big(\frac{1}{4(\sigma\vee\kappa)}\wedge1\Big).$$

Then for all $\theta\in B_0(k_1)$,

$$P_\theta\Big((\theta'-\theta)^T\hat\Sigma(\theta'-\theta)\ge\frac{1}{2}\|\theta'-\theta\|^2\Lambda_{\min}^2\ \forall\theta'\in B_0(k)\Big)\ge1-4\exp[-t].$$

PROOF OF LEMMA 1. The vector $\theta'-\theta$ has at most $k+k_1$ nonzero entries; in the lemma we may thus replace $\theta'-\theta$ by a fixed vector in $B_0(k+k_1)$ and take the supremum over all $k+k_1$-sparse nonzero vectors. In abuse of notation let

us still write θ' for any such vector, and fix a set $S \subset \{1, \dots, p\}$ with cardinality $|S| = k + k_1$. Let $\mathbb{R}^p_S := \{\theta \in \mathbb{R}^p : \theta_j = 0 \ \forall j \notin S\}$. We will show, for $\bar{C}(t,n) \equiv (t + 2(k+k_1)\log 5)/n$, that

$$P\Big(\sup_{\theta' \in \mathbb{R}^p_S, (\theta')^T \Sigma \theta' \neq 0} \left| \frac{(\theta')^T \hat{\Sigma} \theta'}{(\theta')^T \Sigma \theta'} - 1 \right| \geq 4\sigma \sqrt{\bar{C}(t,n)} + 4\kappa \bar{C}(t,n)\Big) \leq 4\exp[-t].$$

Since there are $\binom{p}{(k+k_1)} \leq p^{(k+k_1)}$ sets S of cardinality $k + k_1$, the result then follows from the union bound. To establish the inequality in the last display, it suffices to show

$$P\Big(\sup_{\theta' \in \mathcal{B}_S} |(\theta')^T \Phi \theta'| \geq 4\sigma \sqrt{\bar{C}(t,n)} + 4\kappa \bar{C}(t,n)\Big) \leq 4e^{-t}, \tag{25}$$

where $\mathcal{B}_S := \{(\theta' \in \mathbb{R}^p_S : (\theta')^T \Sigma \theta' \leq 1\}$ and $\Phi := \hat{\Sigma} - \Sigma$.

We use the notation $\|Xu\|^2_\Sigma := u^T \Sigma u$, $u \in \mathbb{R}^p$, and we let for $0 < \delta < 1$, $\{X\theta^l_S\}_{l=1}^{N(\delta)}$ be a minimal δ-covering of $(\{X\theta' : \theta' \in \mathcal{B}_S\}, \|\cdot\|_\Sigma)$. Thus, for every $\theta' \in \mathcal{B}_S$ there is a $\theta^l = \theta^l_S(\theta')$ such that $\|X(\theta' - \theta^l)\|_\Sigma \leq \delta$. Note that $\{\theta^l_S\} \subset \mathbb{R}^p_S$. Following an idea of Loh and Wainwright (2012), we then have

$$\sup_{\theta' \in \mathcal{B}_S} |(\theta' - \theta^l_S(\theta'))^T \Phi (\theta' - \theta^l_S(\theta'))| \leq \delta^2 \sup_{\vartheta \in \mathcal{B}_S} \vartheta^T \Phi \vartheta$$

and also that $\sup_{\theta' \in \mathcal{B}_S} |(\theta' - \theta^l_S(\theta'))^T \Phi \theta| \leq \delta \sup_{\vartheta \in \mathcal{B}_S} |\vartheta^T \Phi \vartheta|$. This implies with $\delta = 1/3$ that

$$\sup_{\theta' \in \mathcal{B}_S} |(\theta') \Phi \theta'| \leq (9/2) \max_{l=1,\dots,N(1/3)} |(\theta^l_S) \Phi (\theta^l_S)|.$$

Condition 2(a) ensures that for some constants σ and κ depending only on σ_0 and κ_0, for any u with $\|Xu\|_\Sigma \leq 1$, and any $t > 0$, it holds that

$$P\left(|u^T \Phi u| \geq \sigma \sqrt{\frac{t}{n}} + \kappa \frac{t}{n}\right) \leq 2\exp[-t].$$

This follows from the fact that the $((Xu)_i)$ are sub-Gaussian, hence, the squares $((Xu)_i^2)$ are sub-exponential. Bernstein's inequality can therefore be used [e.g., Bühlmann and van de Geer (2011), Lemma 14.9]. Finally, the covering number of a ball in $k + k_1$-dimensional space is well known. Apply, for example, Lemma 14.27 in Bühlmann and van de Geer (2011): $N(\delta) \leq ((2+\delta)/\delta)^{k+k_1}$. If we take $\delta = 1/3$, this gives $N(1/3) \leq 9^{k+k_1}$. The union bound then proves (25). □

3.3.1. *A ratio-bound for $\theta' \mapsto \varepsilon^T X(\theta - \theta')$.*

LEMMA 2. *Suppose that $\varepsilon \sim N(0, I)$ is independent of X. Let $\delta > 0$. Then for any $t \geq \max(1/\delta, 1)$, and for $R_t = tC_0\sqrt{k_0 \log p/n}$ where C_0 is a universal*

constant, we have for some universal constants C_1 and C_2,

$$P\Bigg(\sup_{\theta'\in B_0(k_0),\|X(\theta-\theta')\|_n>R_t}\frac{|\varepsilon^T X(\theta-\theta')|/n}{\|X(\theta-\theta')\|_n^2}\geq\delta\Big|X\Bigg)\leq C_1\exp\Bigg[-\frac{t^2\delta^2k_0\log p}{C_2}\Bigg].$$

PROOF. Let $\mathcal{G}_R(\theta):=\{\theta':\|X(\theta-\theta')\|_n\leq R,\theta'\in B_0(k_0)\}$. Then, using the bound $\log\binom{p}{k_0}\leq k_0\log p$ and, for example, Lemma 14.27 in Bühlmann and van de Geer (2011), we have, for $H(u,B,\|\cdot\|)=\log N(u,B,\|\cdot\|)$ the logarithm of the usual u-covering number of a subset B of a normed space

$$H\big(u,\{X(\theta-\theta'):\theta'\in\mathcal{G}_R(\theta)\},\|\cdot\|_n\big)\leq(k_0+1)\log\Big(\frac{2R+u}{u}\Big)+k_0\log p,\qquad u>0.$$

Indeed, if we fix the locations of the zeros, say, $\theta'\in B_0'(k_0):=\{\vartheta:\vartheta_j=0\ \forall j>k_0\}$, then $\{X\theta':\theta'\in B_0'(k_0)\}$ is a k_0-dimensional linear space, so

$$H\big(u,\{X\theta':\theta'\in B_0'(k_0),\|X\theta'\|_n\leq R\},\|\cdot\|_n\big)\leq k_0\log\Big(\frac{2R+u}{u}\Big),\qquad u>0.$$

Furthermore, the vector $X\theta$ is fixed, so that $\mathcal{G}_R(\theta)$ is a subset of a ball with radius R in the (k_0+1)-dimensional linear space spanned by $\{X_j\}_{j=1}^{k_0}$, $X\theta$.

By Dudley's bound [see Dudley (1967) or more recent references such as van der Vaart and Wellner (1996), van de Geer (2000)], applied to the (conditional on X) Gaussian process $\theta'\mapsto\varepsilon^TX(\theta-\theta')$, and using $\int_0^c\sqrt{\log(c/x)}\,dx=c\int_0^1\sqrt{\log(1/x)}\,dx=cA$, where A is the constant $A=\int_0^1\sqrt{\log(1/x)}\,dx$, we obtain

$$E\Big[\sup_{\theta'\in\mathcal{G}_R(\theta)}|\varepsilon^TX(\theta-\theta')||X\Big]\leq C'\int_0^R\sqrt{nH\big(u,\mathcal{G}_R(\theta),\|\cdot\|_n\big)}\,du\leq C\sqrt{2k_0\log p}\sqrt{n}R$$

for some universal constants $C\geq1$ and C'. By the Borell–Sudakov–Cirelson Gaussian concentration inequality [e.g., Boucheron, Lugosi and Massart (2013)], we therefore have for all $u>0$,

$$P\Bigg(\sup_{\theta'\in\mathcal{G}_R(\theta)}|\varepsilon^TX(\theta-\theta')|/n\geq CR\sqrt{\frac{2k_0\log p}{n}}+R\sqrt{\frac{2u}{n}}\Big|X\Bigg)\leq\exp[-u].$$

Substituting $u=v^2k_0\log p$ gives that for all $v>0$,

$$P\Bigg(\sup_{\theta'\in\mathcal{G}_R(\theta)}|\varepsilon^TX(\theta-\theta')|/n\geq(C+v)R\sqrt{\frac{2k_0\log p}{n}}\Big|X\Bigg)\leq\exp\big[-v^2k_0\log p\big],$$

which implies that for all $v \geq 1$,

$$P\Big(\sup_{\theta' \in \mathcal{G}_R(\theta)} |\varepsilon^T X(\theta - \theta')|/n \geq 2vCR\sqrt{\frac{2k_0 \log p}{n}}\Big|X\Big) \leq \exp[-v^2 k_0 \log p].$$

Now insert the peeling device [see Alexander (1985), the terminology coming from van de Geer (2000), Section 5.3]. Let $R_t := 8Ct\sqrt{2k_0 \log p/n}$. We then have

$$\begin{aligned}
&P\Big(\sup_{\theta' \in B_0(k_0), \|X(\theta-\theta')\|_n > R_t} \frac{|\varepsilon^T X(\theta - \theta')|/n}{\|X(\theta - \theta')\|_n^2} \geq \delta\Big|X\Big) \\
&\quad \leq \sum_{s=1}^{\infty} P\Big(\sup_{\theta' \in \mathcal{G}_{2^s R_t}(\theta)} |\varepsilon^T X(\theta - \theta')|/n \geq \delta 2^{2(s-1)} R_t^2\Big|X\Big) \\
&\quad = \sum_{s=1}^{\infty} P\Big(\sup_{\theta' \in \mathcal{G}_{2^s R_t}(\theta)} |\varepsilon^T X(\theta - \theta')|/n \geq 2^s R_t \times 2C(2^s t\delta)\sqrt{\frac{2k_0 \log p}{n}}\Big|X\Big) \\
&\quad \leq \sum_{s=1}^{\infty} \exp[-2^{2s} t^2 \delta^2 k_0 \log p] \leq C_1 \exp\Big[-\frac{t^2 \delta^2 k_0 \log p}{C_2}\Big]
\end{aligned}$$

for some universal constants C_1 and C_2, completing the proof. □

3.3.2. *A ratio-bound for* $\theta' \mapsto L_n^{(1)}(\theta') \equiv \langle (E_\theta X^T X - X^T X)\theta, \theta - \theta' \rangle$.

LEMMA 3. *We have, for every* $\delta > 0$, $R_t = tD_1\sqrt{k_0 \log p/n}$, $t \geq 1$, *some positive constants* D_1, D_2, D_3, D_4, D_5 *depending on* δ, *that*

$$\sup_{\theta \in B_r(M)} P_\theta\Big(\sup_{\theta' \in B_0(k_0): \|\theta - \theta'\| > R_t} \frac{|L_n^{(1)}(\theta')|}{\|\theta - \theta'\|^2} > \delta\Big) \leq B(t, p, n),$$

where $B(t, p, n) = D_2 e^{-D_3 t^2 \delta^2 k_0 \log p}$ *under the assumptions of Theorem* 5(B), $r = 1$, *and* $B(t, p, n) = D_4 e^{-D_5 t\delta\sqrt{n \log p}/k_1}$ *under the assumptions of Theorem* 5(B), $r = 2$.

PROOF. The process in question is of the form

$$L_n^{(1)} : \theta' \mapsto \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(Z_{ij} - EZ_{ij})(\theta_j - \theta_j'), \qquad Z_{ij} = \sum_{m=1}^{p} \theta_m X_{im} X_{ij}. \tag{26}$$

Since the X_{ij} are uniformly bounded by b, we conclude that the summands in i of this process are uniformly bounded by

$$2b^2 \sum_{j=1}^{p} |\theta_j - \theta_j'| \sum_{m=1}^{p} |\theta_m| \tag{27}$$

and the weak variances $n\operatorname{Var}_\theta(L_n^{(1)}(\theta'))$ equal, for δ_{mj} the Kronecker delta,

$$
\begin{aligned}
&E\sum_{j,l}(Z_{ij}-EZ_{ij})(Z_{il}-EZ_{il})(\theta_j-\theta'_j)(\theta_l-\theta'_l)\\
&\quad=E\sum_{j,l,m,m'}(X_{im}X_{ij}-\delta_{mj})(X_{im'}X_{il}-\delta_{m'l})\theta_m\theta_{m'}(\theta_j-\theta'_j)(\theta_l-\theta'_l)\qquad(28)\\
&\quad=\sum_{j,l,m,m'}D_{mjm'l}\theta_m\theta_{m'}(\theta_j-\theta'_j)(\theta_l-\theta'_l)\le c\|\theta\|_2^2\|\theta-\theta'\|_2^2,
\end{aligned}
$$

where we have used, by the design assumptions, that $D_{mjm'l}\le 1$ whenever the indices m,j,m',l match exactly to two distinct values, $D_{mjm'l}\le EX_{11}^4$ if $m=l=j=m'$, and $D_{mjm'l}=0$ in all other cases, as well as the Cauchy–Schwarz inequality. So $L_n^{(1)}$ is a uniformly bounded empirical process $\{(P_n-P)(f_{\theta'})\}_{\theta'\in H_0}$ given by

$$
\frac{1}{n}\sum_{i=1}^n\bigl(f_{\theta'}(Z_i)-Ef_{\theta'}(Z_i)\bigr),\qquad f_{\theta'}(Z_i)=\sum_{j=1}^p\sum_{m=1}^p\theta_mX_{im}X_{ij}(\theta_j-\theta'_j)
$$

with variables $Z_i=(X_{i1},\dots,X_{ip})^T\in\mathbb{R}^p$. Define $\mathcal{F}_s\equiv\{f=f_{\theta'}:\theta'\in H_0,\|\theta'-\theta\|^2\le 2^{s+1}\}$. We know $R_t<\|\theta-\theta'\|\le\sqrt{C}$ so the first probability in (23) can be bounded, for $c'>0$ a small constant, by

$$
\begin{aligned}
&P_\theta\Bigl(\max_{s\in\mathbb{Z}:\,c'R_t^2\le 2^s\le C}\ \sup_{\theta'\in H_0,2^s<\|\theta-\theta'\|^2\le 2^{s+1}}\frac{|L_n^{(1)}(\theta')|}{\|\theta-\theta'\|^2}>\delta\Bigr)\\
&\quad\le\sum_{s\in\mathbb{Z}:\,c'R_t^2\le 2^s\le C}P_\theta\Bigl(\sup_{\theta'\in H_0,\|\theta-\theta'\|^2\le 2^{s+1}}|L_n^{(1)}(\theta')|>2^s\delta\Bigr),\\
&\sum_{s\in\mathbb{Z}:\,c'R_t^2\le 2^s\le C}P_\theta\bigl(\|P_n-P\|_{\mathcal{F}_s}-E\|P_n-P\|_{\mathcal{F}_s}>2^s\delta-E\|P_n-P\|_{\mathcal{F}_s}\bigr).
\end{aligned}
$$

Moreover, $\mathcal{F}_s$ varies in a linear space of measurable functions of dimension k_0, so we have, from $\log\binom{p}{k_0}\le k_0\log p$ and from Theorem 2.6.7 and Lemma 2.6.15 in van der Vaart and Wellner (1996), that

$$
H\bigl(u,\mathcal{F}_s,L^2(Q)\bigr)\lesssim k_0\log(AU/u)+k_0\log p,\qquad 0<u<UA
$$

for some universal constant A and envelope bound U of $\mathcal{F}_s$. Using (27), if θ,θ' are bounded in ℓ^1 by M, we can take U a large enough fixed constant depending on M,b only, and if k_0 is constant, we can take $U=\max(k_1\sqrt{2^s},1)$ since $\|\theta-\theta'\|_1\le\sqrt{k_1}\|\theta-\theta'\|_2$. A standard moment bound for empirical processes under a uniform entropy condition [e.g., Proposition 3 in Giné and Nickl (2009)] then gives, using (28),

$$
E\|P_n-P\|_{\mathcal{F}_s}\lesssim\sqrt{\frac{2^sk_0}{n}\log p}+\frac{Uk_0\log p}{n},\qquad(29)
$$

which is, under the maintained hypotheses, of smaller order than $2^s\delta$ precisely for those s such that $R_t^2 \simeq (k_0/n)\log p \lesssim 2^s$. The last sum of probabilities can thus be bounded, for D_1 large enough and c_0 some positive constant, by

$$\sum_{s\in\mathbb{Z}:\, c'R_t^2\le 2^s\le C} P_\theta\big(n\|P_n - P\|_{\mathcal{F}_s} - nE\|P_n - P\|_{\mathcal{F}_s} > c_0 n 2^s\delta\big),$$

to which we can apply Talagrand's inequality [Talagrand (1996)] [as at the end of the proof of Proposition 1 in Bull and Nickl (2013)], to obtain the bound

$$\sum_{s\in\mathbb{Z}:\, c'R_t^2\le 2^s\le C} \exp\left\{-\delta^2\frac{c_0^2 n^2 (2^s)^2}{n2^{s+1} + nUE\|P_n - P\|_{\mathcal{F}_s} + Uc_0 n 2^s\delta}\right\}.$$

Using (29), this gives the desired bound $D_2 e^{-D_3 t^2\delta^2 k_0\log p}$ when the envelope U is constant, and the bound $B(t,p,n) = D_4 e^{-D_5 t\delta(n\log p)^{1/2}/k_1}$ when the envelope is $U = \max(k_1\sqrt{2^s}, 1)$ (with k_0 constant), completing the proof. □

3.3.3. *Tail inequalities for sparse estimators.* Recall that $S_\vartheta := \{j : \vartheta_j \neq 0\}$. Let $k_\vartheta := |S_\vartheta|$. For $\lambda > 0$, take the estimator

$$\tilde\theta := \arg\min_\vartheta\big\{\|Y - X\vartheta\|_2^2/n + \lambda^2 k_\vartheta\big\}. \tag{30}$$

LEMMA 4. *Let $\varepsilon \sim \mathcal{N}(0, I)$ be independent of X. Take $\lambda^2 = C_3\log p/n$, where C_3 is an appropriate universal constant. Let $t \ge 1$ be arbitrary and $R_t := \sqrt{t/n}$. Then for some universal constants C_4 and C_5,*

$$\sup_{\theta\in B_0(k_0)} P_\theta\big(\|X(\tilde\theta - \theta)\|_n^2 + \lambda^2 k_{\tilde\theta} > 2\lambda^2 k_0 + R_t^2 | X\big) \le C_4\exp\left[-\frac{nR_t^2}{C_5}\right].$$

PROOF. The result follows from an oracle inequality for least squares estimators with general penalties as given in van de Geer (2001). For completeness, we present a full proof. Define

$$\tau^2(\vartheta;\theta) := \|X(\vartheta - \theta)\|_n^2 + \lambda^2 k_\vartheta \quad\text{and}\quad \mathcal{G}_R(\theta) := \big\{\vartheta : \tau^2(\vartheta) \le R\big\}.$$

If $\tau^2(\tilde\theta;\theta) \le 2\lambda^2 k_\theta$, we are done. So suppose $\tau^2(\tilde\theta;\theta) > 2\lambda^2 k_\theta$. We then have $(2/n)\varepsilon^T X(\tilde\theta - \theta) \ge \tau^2(\tilde\theta,\theta) - \lambda^2 k_\theta \ge \tau^2(\tilde\theta,\theta)/2$. Now again apply the peeling device:

$$\begin{aligned} &P\left(\sup_{\tau(\vartheta;\theta)>R_t}\frac{\varepsilon^T X(\vartheta - \theta)/n}{\tau^2(\vartheta,\theta)} \ge \frac14\middle| X\right)\\ &\quad\le \sum_{s=1}^\infty P\left(\sup_{\vartheta\in\mathcal{G}_{2^s R_t}(\theta)}\frac{\varepsilon^T X(\vartheta - \theta)/n}{\tau^2(\vartheta,\theta)} \ge \frac{1}{16}2^{2s}R_t^2\middle| X\right). \end{aligned}$$

But if $\vartheta \in \mathcal{G}_R(\theta)$, we know that $\|X(\vartheta-\theta)\|_n \le R$ and that $k_\vartheta \le R^2/\lambda^2$. Hence, as in the proof of Lemma 2, we know that

$$P\Big(\sup_{\vartheta\in\mathcal{G}_R(\theta)} \varepsilon^T X(\vartheta-\theta)/n \ge 2CR\sqrt{\frac{2R^2\log p}{n\lambda^2}}\Big|X\Big) \le \exp\Big[-\frac{C^2R^2\log p}{\lambda^2}\Big].$$

As $\lambda = 32C\sqrt{2\log p/n}$, we get

$$P\Big(\sup_{\vartheta\in\mathcal{G}_R(\theta)} \varepsilon^T X(\vartheta-\theta)/n \ge \frac{R^2}{16}\Big|X\Big) \le \exp\Big[-\frac{nR^2}{2\times(32)^2}\Big].$$

We therefore have

$$P\Big(\sup_{\tau(\vartheta;\theta)>R_t} \frac{\varepsilon^T X(\vartheta-\theta)/n}{\tau^2(\vartheta,\theta)} \ge \frac{1}{4}\Big|X\Big) \le \sum_{s=1}^{\infty} \exp\Big[-\frac{n2^{2s}R_t^2}{2\times(32)^2}\Big] \le C_4\exp\Big[-\frac{nR_t^2}{C_5}\Big]$$

for some universal constants C_4 and C_5. □

COROLLARY 2. *Assume Condition 2 and let $\varepsilon \sim \mathcal{N}(0,I)$ be independent of X. Let $\tilde\theta$ be as in (30) with $\lambda^2 = (C_3\log p)/n$, where C_3 is as in Lemma 4, and let $k_0 = o(n/\log p)$. Then for some universal constants C_6, C_7, C_8, c, every $C \ge C_6$ and every n large enough,*

$$\sup_{\theta\in B_0(k_0)} P_\theta\Big(\|\tilde\theta-\theta\|^2 > C\frac{k_0\log p}{n}\Big) \le C_7\exp\Big[-\frac{k_0\log p}{C_8}\Big].$$

PROOF. By Lemma 4 with R_τ, τ equal to a suitable constant times $k_0\log p$, we see first $k_{\tilde\theta} \lesssim 3k_0$ on the event on which the exponential inequality holds. Then from Corollary 1 with $k = 3k_0$, on an event of sufficiently large probability, $\|\tilde\theta-\theta\|_2^2 \le C(\Lambda_{\min})\|X(\tilde\theta-\theta)\|_n^2$ for n large enough, so that the result follows from applying Lemma 4 again [this time to $\|X(\tilde\theta-\theta)\|_n^2$] and from combining the bounds. □

3.3.4. *Proof of Theorem 1 under Condition 2.* For p, n fixed, the random vectors $(Y_i, X_{i1}, \dots, X_{ip})_{i=1}^n$ are i.i.d., and if we split the n points into two subsamples, each of size of order n, then we have two independent replicates $Y^{(s)} = X^{(s)}\theta + \varepsilon^{(s)}$, $\hat\Sigma^{(s)} = (X^{(s)})^T X^{(s)}/n, s = 1, 2$, of the model. In abuse of notation, denote throughout this proof by $\tilde\theta \equiv \tilde\theta^{(1)}$ the estimator from (30) based on the subsample $s = 1$, with λ chosen as in Lemma 4, and by $(Y, X, \varepsilon) \equiv (Y^{(2)}, X^{(2)}, \varepsilon^{(2)})$ the variables from the second subsample. Define

$$\begin{aligned}\hat R_n &= \frac{1}{n}(Y-X\tilde\theta)^T(Y-X\tilde\theta) - 1\\ &= (\theta-\tilde\theta)^T\hat\Sigma^{(2)}(\theta-\tilde\theta) + \frac{2}{n}\varepsilon^T X(\theta-\tilde\theta) + \frac{1}{n}\varepsilon^T\varepsilon - 1.\end{aligned}$$

By independence, and conditional on $(Y^{(1)}, X^{(1)})$, we have $E_\theta^{(2)}(\varepsilon^T X(\theta-\tilde\theta))^2 = n(\tilde\theta-\theta)^T\Sigma(\tilde\theta-\theta)$ and so, using Markov's inequality,

$$\frac{2}{n}\varepsilon^T X(\theta-\tilde\theta) = O_P\left(\sqrt{\frac{(\tilde\theta-\theta)^T\Sigma(\tilde\theta-\theta)}{n}}\right). \tag{31}$$

By Lemma 4, we have $\|X^{(1)}(\tilde\theta-\theta)\|_n^2 = O_P((k\log p)/n)$ and $k_{\tilde\theta} = O(k_1)$ and, hence, by Lemma 1, also $(\tilde\theta-\theta)^T\Sigma(\tilde\theta-\theta) = O_P((k\log p)/n) = o(1)$. Thus, the bound in (31) is $o_P(1/\sqrt{n})$ uniformly in $B_0(k_1)$, and this will be used in the following estimate. Let u_α be suitable quantile constants to be chosen below. Take as confidence set

$$C_n = \left\{\theta\in\mathbb{R}^p : \|\theta-\tilde\theta\|^2 \le 2\Lambda_{\min}^{-2}\left(\hat R_n + \frac{u_\alpha}{\sqrt{n}}\right)\right\}.$$

Uniformly in $\theta\in B_0(k_1)$ with $k_1 = o(n/\log p)$, we have again by Lemma 4 that $\tilde\theta\in B_0(2k_1)$ on events of probability approaching one, so that, using Corollary 1 on these events,

$$\begin{aligned}
P_\theta(\theta\notin C_n) &= P_\theta\left(\|\theta-\tilde\theta\|^2 > 2\Lambda_{\min}^{-2}\left(\hat R_n + \frac{u_\alpha}{\sqrt{n}}\right)\right)\\
&\le P_\theta\left((\theta-\tilde\theta)^T\hat\Sigma^{(2)}(\theta-\tilde\theta) > \hat R_n + \frac{u_\alpha}{\sqrt{n}}\right) + o(1)\\
&= P_\theta\left(-\frac{1}{n}\varepsilon^T\varepsilon + 1 > \frac{u_\alpha}{\sqrt{n}} + \frac{2}{n}\varepsilon^T X(\theta-\tilde\theta)\right) + o(1)\\
&= P_\theta\left(\frac{-1}{\sqrt{n}}\sum_{i=1}^n(\varepsilon_i^2-1) > (1+o(1))u_\alpha\right) + o(1) \le \alpha + o(1)
\end{aligned}$$

for a fixed constant u_α. Moreover, from the previous arguments and Corollary 2, we see that, for $\theta\in B_0(k)$, the diameter $\hat R_n = O_P(\|\tilde\theta-\theta\|^2 + n^{-1/2})$ is of order $O_P(\frac{k\log p}{n} + n^{-1/2})$.

Acknowledgements. We would like to thank the Editor, Associate Editor, two referees and Sasha Tsybakov for helpful remarks on this article. Richard Nickl is grateful to the Cafes Florianihof and Griensteidl in Vienna where parts of this research were carried out.

REFERENCES

ALEXANDER, K. S. (1985). Rates of growth for weighted empirical processes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). **2** 475–493. Wadsworth, Belmont, CA. MR0822047

ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. MR2060168

BERAN, R. and DÜMBGEN, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26** 1826–1856. MR1673280

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford Univ. Press, London.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** 889–919. MR3078289

CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240

CANDÈS, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

DUDLEY, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1** 290–330. MR0220340

GINÉ, E. and NICKL, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596. MR2475673

GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707

HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30** 325–396. MR1902892

HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

JAVANMARD, A. and MONTANARI, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. Available at arXiv:1306.3171.

JUDITSKY, A. and LAMBERT-LACROIX, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.* **12** 410–428. MR2054156

LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

PÖTSCHER, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā* **71** 1–18. MR2579644

PÖTSCHER, B. M. and SCHNEIDER, U. (2011). Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electron. J. Stat.* **5** 1876–1934. MR2970179

ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241

TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

VAN DE GEER, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Math. Methods Statist.* **10** 355–374. MR1867165

VAN DE GEER, S., BÜHLMANN, P. and RITOV, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Submitted. Available at arXiv:1303.0518.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer, New York. MR1385671

ZHANG, C. H. and ZHANG, S. S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. 2011. Available at arXiv:1110.2563v1.

STATISTICAL LABORATORY
DEPARTMENT OF PURE MATHEMATICS
AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
CB3 0WB CAMBRIDGE
UNITED KINGDOM
E-MAIL: r.nickl@statslab.cam.ac.uk

SEMINAR FOR STATISTICS
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: geer@stat.math.ethz.ch