Supplementary Material for "Subset Selection with Shrinkage: Sparse Linear Modeling when the SNR is Low" by Mazumder, Radchenko and Dedieu

A Computational details

A.1 Proof of Proposition 2

(a) It follows from (9) that for any $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}\|_0 \leq k$:

$$\begin{aligned}
F(\boldsymbol{\beta}) &= Q_L(\boldsymbol{\beta}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_q \\
&\geq \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; (Q_L(\boldsymbol{\eta}, \boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; \left(\frac{L}{2} \|\boldsymbol{\eta} - \boldsymbol{\beta}\|_2^2 + \langle \nabla f(\boldsymbol{\beta}), \boldsymbol{\eta} - \boldsymbol{\beta} \rangle + f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q \right) \\
&= \inf_{\|\boldsymbol{\eta}\|_0 \leq k} \; \left(\frac{L}{2} \left\| \boldsymbol{\eta} - \left(\boldsymbol{\beta} - \frac{1}{L} \nabla f(\boldsymbol{\beta}) \right) \right\|_2^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{\beta})\|_2^2 + f(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\eta}\|_q \right) \qquad \text{(A.20)} \\
&= \left(\frac{L}{2} \left\| \widehat{\boldsymbol{\eta}} - \left(\boldsymbol{\beta} - \frac{1}{L} \nabla f(\boldsymbol{\beta}) \right) \right\|_2^2 - \frac{1}{2L} \|\nabla f(\boldsymbol{\beta})\|_2^2 + f(\boldsymbol{\beta}) \right) + \lambda \|\widehat{\boldsymbol{\eta}}\|_q. \qquad \text{(A.21)}
\end{aligned}$$

Note that in (A.21) above we use the notation $\hat{\boldsymbol{\eta}}$ to denote a minimizer of (A.20). We now follow the proof in Proposition 6 in [9] to arrive at:

$$F(\boldsymbol{\beta}) \geq \frac{L - L_0}{2} \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\beta}\|_2^2 + F(\widehat{\boldsymbol{\eta}}). \qquad \text{(A.22)}$$

In particular, using $\hat{\boldsymbol{\eta}} = \boldsymbol{\beta}^{(m+1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ and $L \geq L_0$, we see that the sequence $F(\boldsymbol{\beta}^{(m)})$ is decreasing. Because $F(\boldsymbol{\beta}) \geq 0$, we observe that the sequence $F(\boldsymbol{\beta}^{(m)})$ converges to some $F^* \geq 0$.

(b) Summing inequalities (A.22) for $1 \leq m \leq M$, we obtain

$$\sum_{m=1}^{M} \left(F(\boldsymbol{\beta}^{(m)}) - F(\boldsymbol{\beta}^{(m+1)}) \right) \geq \frac{L - L_0}{2} \sum_{m=1}^{M} \|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2, \qquad \text{(A.23)}$$

leading to

$$F(\boldsymbol{\beta}^{(1)}) - F(\boldsymbol{\beta}^{(M+1)}) \geq \frac{M(L - L_0)}{2} \min_{m=1,\ldots,M} \|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|_2^2.$$

Because the decreasing sequence $F(\boldsymbol{\beta}^{(m)})$ converges to $F(\boldsymbol{\beta}^*) = F^*$, say, we arrive at the conclusion in Part (b).

A.2 MIO Formulations: adding implied inequalities

We use the following notation for the model matrix: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$. We consider a structured version of Problem (5) with additional implied inequalities (cuts) for improved lower bounds:

$$
\begin{aligned}
\text{minimize} \quad & \frac{u}{2} + \lambda v \\
\text{s.t.} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \leq u && \text{(A.24a)} \\
& \|\boldsymbol{\beta}\|_q \leq v && \text{(A.24b)} \\
& z_j \in \{0,1\}, j \in [p] \\
& \sum_j z_j = k \\
& -\mathcal{M}_i \leq \beta_i \leq \mathcal{M}_i, i \in [p] && \text{(A.24c)} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i^+, i \in [n] && \text{(A.24d)} \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1}, && \text{(A.24e)}
\end{aligned}
$$

where (a) $\mathcal{M}_i, i \in [p]$ denote bounds on β_i's via constraint (A.24c); (b) $-\bar{\mathcal{M}}_i^-, \bar{\mathcal{M}}_i^+$ denote bounds on the predicted values $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$ for $i \in [n]$ via constraint (A.24d); (c) $\mathcal{M}_{\ell_1}$, in constraint (A.24e), denotes an upper bound on the ℓ_1-norm of the regression coefficients $\|\boldsymbol{\beta}\|_1$.

The additional cuts in Problem (A.24) can potentially help the progress of the MIO solver. The caveat, however, is that the resulting formulation has additional variables – hence more work needs to be done within every node of the branch-and-bound tree. Section A.3 presents ways to compute these bounds – Section A.3.1 describes ways to compute them via convex optimization – these are bounds implied by an optimal solution to Problem (3). Section A.3.2 describes ways to compute these bounds based on good heuristic solutions.

A.3 Computing problem specific parameters

A.3.1 Computing parameters via convex optimization

Formulation (4) involves a BigM value $\mathcal{M}$ – tighter formulations can be obtained by using variable dependent BigM values for the β_i:

$$-\mathcal{M}_i z_i \leq \beta_i \leq \mathcal{M}_i z_i, \quad i \in [p].$$

In addition, implied constraints (or bounds) on $\langle \mathbf{x}_i, \boldsymbol{\beta} \rangle$'s can also be added:

$$-\bar{\mathcal{M}}_i \leq \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle \leq \bar{\mathcal{M}}_i, \quad i \in [n].$$

We discuss how to compute these from data using convex optimization. Note that, because $\boldsymbol{\beta}$ is k-sparse, we have $|\langle \mathbf{x}_i, \boldsymbol{\beta}\rangle| \leq \mathcal{M}\|\mathbf{x}_i\|_{k,1}$, where for a vector $\mathbf{a} \in \mathbb{R}^p$ the quantity $\|\mathbf{a}\|_{k,1}$ denotes the ℓ_1-norm of the k-largest (in absolute value) entries of $\mathbf{a}$. We can set $\bar{\mathcal{M}}_i \leq \mathcal{M}\|\mathbf{x}_i\|_{k,1}$. Note also that $\|\boldsymbol{\beta}\|_1 \leq \mathcal{M}k := \mathcal{M}_{\ell_1}$. We now upper bound each coefficient β_i by solving the optimization problems:

$$\begin{aligned}
\mathcal{M}_i^+ = \max \quad & \beta_i \\
\text{s.t.} \quad & \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \leq \text{UB} \\
& \|\boldsymbol{\beta}\|_\infty \leq \mathcal{M} \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \leq \bar{\mathcal{M}}_i^+, i \in [n]
\end{aligned}
\qquad
\begin{aligned}
\mathcal{M}_i^- = \max \quad & -\beta_i \\
\text{s.t.} \quad & \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \leq \text{UB} \\
& \|\boldsymbol{\beta}\|_\infty \leq \mathcal{M} \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \leq \bar{\mathcal{M}}_i^+, i \in [n]
\end{aligned}
\tag{A.25}$$

where UB is an upper bound to Problem (3) obtained via Algorithm 1, for example. Upon solving Problem (A.25), we set $\mathcal{M}_i = \max\{\mathcal{M}_i^+, \mathcal{M}_i^-\}$ for all $i \in [p]$. Consequently, we can update the bounds $\mathcal{M} = \|\mathcal{M}_i\|_\infty$, $\bar{\mathcal{M}}_i$ and $\mathcal{M}_{\ell_1}$ – such bound tightening methods have been proposed in [40] in the context of the Discrete Dantzig Selector problem.

Similarly, we can also refine bounds on $\langle \mathbf{x}_j, \boldsymbol{\beta}\rangle$ by solving the following pair of optimization problems for all $j \in [n]$.

$$\begin{aligned}
\bar{\mathcal{M}}_j^+ = \max \quad & \langle \mathbf{x}_j, \boldsymbol{\beta}\rangle \\
\text{s.t.} \quad & \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \leq \text{UB} \\
& -\mathcal{M}_i^- \leq \beta_i \leq \mathcal{M}_i^+, i \in [p] \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \leq \bar{\mathcal{M}}_i^+, i \in [n]
\end{aligned}
\qquad
\begin{aligned}
\bar{\mathcal{M}}_j^- = \max \quad & -\langle \mathbf{x}_j, \boldsymbol{\beta}\rangle \\
\text{s.t.} \quad & \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_q \leq \text{UB} \\
& -\mathcal{M}_i^- \leq \beta_i \leq \mathcal{M}_i^+, i \in [p] \\
& \|\boldsymbol{\beta}\|_1 \leq \mathcal{M}_{\ell_1} \\
& -\bar{\mathcal{M}}_i^- \leq \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \leq \bar{\mathcal{M}}_i^+, i \in [n].
\end{aligned}
\tag{A.26}$$

Upon solving Problem (A.26), we can set $\bar{\mathcal{M}}_i = \max\{|\bar{\mathcal{M}}_j^+|, |\bar{\mathcal{M}}_j^-|\}$. The bounds thus obtained can be used to tighten the bounds used in Problems (A.25) and (A.26). New bounds on $\{\mathcal{M}_i\}$ and $\{\bar{\mathcal{M}}_i\}$ can be obtained by solving the new problems with the updated bounds.

Remark 12. *Problems (A.25), (A.26) drop the cardinality constraint on $\boldsymbol{\beta}$ – hence the derived bounds need not be tight, i.e., $\mathcal{M}_i > |\hat{\beta}_i(\lambda; k)|$, where $\hat{\boldsymbol{\beta}}(\lambda; k)$ denotes an optimal solution to Prob-*

lem (3).

A.3.2 Computing parameters via Algorithm 1

We note that the BigM values $\mathcal{M}_i, i \in [p]$ can also be based on the solutions obtained from the heuristic algorithms. For example, we can set $\mathcal{M}_i = \tau \|\hat{\boldsymbol{\beta}}(\lambda; k)\|_\infty$ for all $i \in [p]$ for some multiplier $\tau \in \{1.5, 2\}$, for example. Similarly, the bounds $\bar{\mathcal{M}}_i$ can be set to $\tau |\langle \mathbf{x}_i, \hat{\boldsymbol{\beta}}(\lambda; k)\rangle|$ for all $i \in [n]$. Such bounds are usually tighter and are obtained as a simple by-product of Algorithm 1.

B Proofs of the results in Section 3

B.1 Proof of Theorem 1

We first note that the probability of event $\mathcal{F}$ is at least $1 - \delta_0/2$ by Theorem 4.1 in [5]. Next, we establish the probability bound for $\mathcal{E}_s$.

Because the columns of $\mathbf{X}$ have unit Euclidean norm, we can write $\|\mathbf{X}\mathbf{u}\| \leq \|\mathbf{u}\|_1 \leq \sqrt{s}\|\mathbf{u}\|$ for every $\mathbf{u} \in B_0(s)$. Hence, taking $\delta_0 = s/(2ep)$, we derive

$$\sqrt{\log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\| \leq \sqrt{s \log(2ep/s)}\|\mathbf{u}\|. \tag{B.27}$$

It follows from Stirling's formula that $\log(s!) \geq s\log(s/e)$, and hence

$$\sum_{j=1}^{s} \log(2p/j) = s\log(2p) - \log(s!) \leq s\log(2ep/s).$$

Thus, using the Cauchy-Schwarz inequality and taking into account $\|\mathbf{u}\|_0 \leq s$, we arrive at

$$\sum_{j=1}^{p} u_j^\sharp \sqrt{\log(2p/j)} \leq \|\mathbf{u}\| \sqrt{\sum_{j=1}^{s} \log(2p/j)} \leq \sqrt{s\log(2ep/s)}\|\mathbf{u}\|. \tag{B.28}$$

Inequalities (B.28) and (B.27) yield

$$[4+\sqrt{2}]\sigma \max\Big(\sum_{j=1}^{p} u_j^\sharp \sqrt{\log(2p/j)}, \sqrt{\log(1/\delta_0)}\|\mathbf{X}\mathbf{u}\|\Big) \leq [4+\sqrt{2}]\sigma\sqrt{s\log(2ep/s)}\|\mathbf{u}\|.$$

Consequently, when $\delta_0 = s/(2ep)$, we have $\mathcal{F} \subseteq \mathcal{E}_s$. Because the probability of event $\mathcal{F}$ is at least $1 - s/(4ep)$, we have established the stated probability bound for $\mathcal{E}_s$.

The result for $\mathcal{H}$ follows from the standard tail probability bounds for maxima of Gaussian random variables (for example, those in [14]). The result for $\mathcal{G}_s$ follows from the argument in the proof of Lemma 8 in [51], with appropriate modifications in order to incorporate the uncertainty parameter δ_0.

B.2 Proof of Theorem 2

We consider an arbitrary $\boldsymbol{\beta} \in B_0(k)$ and note that

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_2\| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_2\| \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|. \tag{B.29}$$

We will derive prediction error bounds for $\hat{\boldsymbol{\beta}}_2$ by controlling the term $\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta})$.

We first focus on establishing the slow rate. On the event $\mathcal{E}_{2k}$ we have

$$\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma\sqrt{2k\log(ep/k)}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2\|. \tag{B.30}$$

Combining this inequality with (B.29) and using the lower bound imposed on λ, we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|. \tag{B.31}$$

Thus, we have established the first slow rate prediction error bound.

Repeating the arguments in the proof of Theorem 1, we see that on the event $\mathcal{F}$ we have either (a) inequality (B.30), which implies (B.31), or (b) the following inequality:

$$\boldsymbol{\epsilon}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\|, \tag{B.32}$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\| + 2[4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\big). \tag{B.33}$$

We bound the last term in the above display by two applications of the inequality

$$2ab \leq \alpha a^2 + \alpha^{-1}b^2, \tag{B.34}$$

which holds for every $\alpha > 0$ and $a, b \in \mathbb{R}$. Setting $\alpha = 2$, we derive inequalities

$$2[4 + \sqrt{2}]\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| \leq 2[4 + \sqrt{2}]^2\sigma^2\log(1/\delta_0) + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2/2$$

and

$$\sigma\sqrt{\log(1/\delta_0)}\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\| \lesssim \sigma^2\log(1/\delta_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Taking into account inequality (B.33), we then arrive at the second slow rate prediction error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\| + \sigma^2 \log(1/\delta_0).$$

We now establish the fast rate. Starting with inequality (B.29) and restricting our attention to event $\mathcal{G}_{2k}$, we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\sigma\big[10k\log(ep/[2k]) + \log(1/\delta_0)\big]^{1/2}\big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\big) + \lambda\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2\|.$$

We bound the second term on the right-hand side by two applications of inequality (B.34), in which we set $\alpha = 4$ in order to have $\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2$ appear with the multiplier $1/4$. We bound the last term on the right-hand side using

$$\lambda\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2\| \leq \gamma_{2k}^{-1}\lambda\|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\| \leq \gamma_{2k}^{-1}\lambda\big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|\big),$$

and then apply (B.34) with $\alpha = 2$ again to derive

$$\gamma_{2k}^{-1}\lambda\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\| \leq \gamma_{2k}^{-2}\lambda^2 + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2/4$$

and

$$\gamma_{2k}^{-1}\lambda\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\| \lesssim \gamma_{2k}^{-2}\lambda^2 + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Rearranging the resulting terms we arrive at the fast rate prediction error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k\log(ep/[2k]) + \gamma_{2k}^{-2}\lambda^2 + \sigma^2\log(1/\delta_0).$$

B.3 Proof of Corollary 1

The first prediction error bound is a direct consequence of Theorems 1 and 2. The last two prediction error bounds are derived from Theorems 1 and the corresponding bounds in Theorem 2 by setting $\delta_0 = 1/p$ and $\delta_0 = (k/p)^k$, respectively. The estimation error bound follows from the inequality $\gamma_{2k}^2\|\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*\|^2 \leq \|\mathbf{X}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*)\|^2$.

B.4 Proof of Corollary 2

We let $Q(\boldsymbol{\beta})$ denote the objective function in (16) when $q = 2$. Because $UB = Q(\widetilde{\boldsymbol{\beta}}_2)$, $LB \leq Q(\boldsymbol{\beta}^*)$, and $UB = LB/(1-\tau)$, we derive

$$Q(\widetilde{\boldsymbol{\beta}}_2) \leq Q(\boldsymbol{\beta}^*)/(1-\tau).$$

Because $Q(\widetilde{\boldsymbol{\beta}}_2) = \|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 + \lambda\|\widetilde{\boldsymbol{\beta}}\|$ and $Q(\boldsymbol{\beta}^*) = \|\boldsymbol{\epsilon}\|^2 + \lambda\|\boldsymbol{\beta}^*\|$, we then have

$$\|\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 + \lambda\|\widetilde{\boldsymbol{\beta}} \leq \|\boldsymbol{\epsilon}\|^2/(1-\tau) + \lambda\|\boldsymbol{\beta}^*\|/(1-\tau).$$

Repeating the arguments in the proof of the second slow rate in Theorem 2 while incorporating the optimality gap, we derive that

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \leq 2\lambda\|\boldsymbol{\beta}^*\|/(1-\tau) + 4[4+\sqrt{2}]^2\sigma^2\log(1/\delta_0) + 2\|\boldsymbol{\epsilon}\|^2\tau/(1-\tau) \tag{B.35}$$

on the event $\mathcal{F}$. Standard chi-square tail bounds imply that, with an appropriate multiplicative constant, inequality $\|\boldsymbol{\epsilon}\|^2 \lesssim \sigma^2[n \vee \log(p)]$ holds with probability at least $1 - 1/(2p)$. Letting $\delta_0 = 1/(2p)$, noting $\tau \leq 1$, and recalling that $1/(1-\tau)$ is upper-bounded by a universal constant, we then conclude that inequality

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \lesssim \lambda\|\boldsymbol{\beta}^*\| + \sigma^2[\log(p) + \tau n]$$

holds with probability at least $1 - 1/p$, establishing the first error bound in Corollary 2.

Revisiting inequality (B.35) with $\delta_0 = 1/p$, we note that (as $n \to \infty$) the right-hand side is of the order

$$2\lambda\|\boldsymbol{\beta}^*\|\left(1 + \frac{\tau}{1-\tau}\right) + 4[4+\sqrt{2}]^2\sigma^2\log(p)\left\{1 + \frac{\tau n}{2[4+\sqrt{2}]^2(1-\tau)\log(p)}\right\}.$$

Thus, the multiplicative increase in the error bound relative to the case $\tau = 0$ is at most

$$1 + \frac{\tau}{1-\tau}\left\{1 \vee \frac{n}{58\log(p)}\right\}.$$

We now focus on the second error bound in Corollary 2. Repeating the arguments in the proof of the fast rate in Theorem 2, incorporating the optimality gap, and keeping track of the constants, we arrive at the following error bound:

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 \leq 8\sigma^2\left[10k\log(ep/[2k]) + \log(1/\delta_0)\right] + 2\gamma_{2k}^{-2}\lambda^2\left(1 + \frac{\tau}{1-\tau}\right) + 2\|\boldsymbol{\epsilon}\|^2\tau/(1-\tau), \tag{B.36}$$

which holds on the event $\mathcal{G}_{2k}$. Letting $\delta_0 = (k/p)^k/2$, and again using the chi-square tail bounds to control $\|\boldsymbol{\epsilon}\|^2$, we then conclude that inequality

$$\|\mathbf{f}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_2\|^2 \lesssim \sigma^2[k\log(ep/k) + \tau n] + \gamma_{2k}^{-2}\lambda^2$$

holds with probability at least $1 - (k/p)^k$, establishing the second error bound in Corollary 2.

Revisiting inequality (B.36) with $\delta_0 = (k/p)^k$, we note that (as $n \to \infty$) the right-hand side is of the order

$$88\sigma^2 \log(ep/[2k]) \left\{ 1 + \frac{\tau n}{44(1-\tau)\log(ep/[2k])} \right\} + 2\gamma_{2k}^{-2}\lambda^2 \Big(1 + \frac{\tau}{1-\tau}\Big).$$

Thus, the multiplicative increase in the error bound relative to the case $\tau = 0$ is at most

$$1 + \frac{\tau}{1-\tau} \left\{ 1 \vee \frac{n}{43 \log(ep/[2k])} \right\}.$$

B.5 Proof of Corollary 3

Let c_0 be the universal constant from the second slow rate error bound in Theorem 2. Take an arbitrary $\boldsymbol{\beta} \in B_0(k)$ and define

$$W = \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_2\|^2 - c_0\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 - c_0\lambda\|\boldsymbol{\beta}\|.$$

By Theorems 1 and 2 we have $W \le c_0\sigma^2 \log(1/\delta_0)$ with probability at least $1 - \delta_0/2$. Thus,

$$2\mathbb{P}\big(W > w\big) \le e^{-w/[c_0\sigma^2]},$$

for every non-negative w. Consequently,

$$\mathbb{E}W \le \int_0^\infty \mathbb{P}\big(W > w\big) dw \le \tfrac{1}{2} \int_0^\infty e^{-w/[c_0\sigma^2]} dw \le \frac{c_0\sigma^2}{2},$$

and the first stated bound follows from the definition of W.

The second stated bound follows by an analogous argument, together with an additional observation that $k \log(ep/[2k])$ is bounded away from zero by a positive universal constant.

B.6 Proof of Theorem 3

We consider an arbitrary $\boldsymbol{\beta} \in B_0(k)$. In the ℓ_1 setting, inequality (B.29) becomes

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 + \lambda\|\hat{\boldsymbol{\beta}}_1\|_1 \le \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1. \tag{B.37}$$

On the event $\mathcal{H}$, we then have

$$2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \le 2\|\mathbf{X}^T\boldsymbol{\epsilon}\|_\infty \Big[\|\boldsymbol{\beta}\|_1 + \|\hat{\boldsymbol{\beta}}_1\|_1\Big] \le \lambda\Big[\|\hat{\boldsymbol{\beta}}_1\|_1 + \|\boldsymbol{\beta}\|_1\Big].$$

Consequently,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \le \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda\|\boldsymbol{\beta}\|_1,$$

which completes the proof of the first slow rate error bound.

We now restrict our attention to the event $\mathcal{F}$. Note that, because

$$\sum_{j=1}^{p} u_j^{\sharp} \sqrt{\log(2p/j)} \leq \sqrt{\log(2p)} \|\mathbf{u}\|_1,$$

we must have either (a) inequality

$$\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma \sqrt{\log(2p)} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|_1,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda \|\boldsymbol{\beta}\|_1;$$

or (b) the following inequality:

$$\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \leq [4 + \sqrt{2}]\sigma \sqrt{\log(1/\delta_0)} \|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_2)\|,$$

which implies

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 + [8 + 2\sqrt{2}]\sigma \sqrt{\log(1/\delta_0)} \Big(\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\| \Big).$$

Bounding the last term in the above display by two applications of (B.34) with $\alpha = 2$ yields

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 + \sigma^2 \log(1/\delta_0), \tag{B.38}$$

which establishes the second slow rate error bound.

We now move to the fast rate. Starting with (B.37), using inequalities

$$\lambda \|\boldsymbol{\beta}\|_1 - \lambda \|\hat{\boldsymbol{\beta}}_1\|_1 \leq \lambda \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|_1 \leq \lambda \sqrt{2k} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|,$$

and restricting our attention to the event $\mathcal{G}_{2k}$, we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma \Big[10k \log(ep/[2k]) + \log(1/\delta_0) \Big]^{1/2} \|\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1)\| + \lambda \sqrt{k} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_1\|.$$

Repeating the argument used to establish the fast rate part of Theorem 2, we arrive at

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_1\|^2 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 k \log(ep/[2k]) + \gamma_{2k}^{-2} \lambda^2 k + \sigma^2 \log(1/\delta_0).$$

B.7 Proof of Corollary 4

This result follows by an argument analogous to the one used in the proof of Corollary 3.

B.8 Proof of Corollary 5

This result follows directly from the slow rate parts of Theorems 2 and 3.

B.9 Proof of Theorem 4

The following result will allow us to lower-bound the magnitude of the cross-product term in the sum of squares function.

Lemma 1. *Let $S \subset \{1, ..., p\}$ have cardinality q, and let s be an integer in $[1, q]$. There exists a positive universal constant $\tilde{c}$, such that*

$$\max_{supp(\mathbf{v}) \subset S,\, \|\mathbf{v}\|_0 \leq s,\, \|\mathbf{X}\mathbf{v}\| = 1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| \gtrsim \sigma \gamma_{2s} \sqrt{s \log(eq/s)}$$

with probability at least $1 - 2(eq/s)^{-\tilde{c}\gamma_{2s}^2 s}$.

Lemma 1 is proved in the next subsection.

Using Maurey's argument [49], we can bound the error in approximating $\mathbf{X}\boldsymbol{\beta}^*$ with $\mathbf{X}\boldsymbol{\beta}$, when $\boldsymbol{\beta}$ is restricted to an ℓ_0 ball. More specifically, by Lemma A.1 in [52], there exists a vector $\tilde{\boldsymbol{\beta}}^* \in B_0(k/2)$ such that $\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 \leq 2\|\boldsymbol{\beta}^*\|_1^2/k$. For convenience, we define $\boldsymbol{\Delta}^* = \mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\tilde{\boldsymbol{\beta}}^*$. Minimizing the sum of squares is equivalent to minimizing the function

$$G(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|^2 + 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}}^* - \mathbf{X}\boldsymbol{\beta}).$$

Given a vector $\mathbf{u} \in \mathbb{R}^p$, we define

$$H(\mathbf{u}) = \|\mathbf{X}\mathbf{u}\|^2 - 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{u}.$$

Given an index set $\mathcal{I}$ and a vector $\boldsymbol{\beta}$, we will write $\boldsymbol{\beta}_\mathcal{I}$ for the vector that (a) matches $\boldsymbol{\beta}$ element by element on the index set $\mathcal{I}$; and (b) has its support contained in $\mathcal{I}$. Let $\tilde{S}$ denote the support of $\tilde{\boldsymbol{\beta}}^*$. Note that if $\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*$ and $\|\boldsymbol{\beta}\|_0 \leq k$, then

$$G(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta}_{\tilde{S}^c}\|^2 - 2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\boldsymbol{\beta}_{\tilde{S}^c} = H(\boldsymbol{\beta}_{\tilde{S}^c}).$$

Note that $|\tilde{S}| \leq k/2$, and hence

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq \min_{\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*,\, \|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq \min_{\boldsymbol{\beta}_{\tilde{S}} = \tilde{\boldsymbol{\beta}}^*,\, \|\boldsymbol{\beta}\|_0 \leq k} H(\boldsymbol{\beta}_{\tilde{S}^c}) \leq \min_{\text{supp}(\mathbf{u}) \subseteq \tilde{S}^c,\, \|\mathbf{u}\|_0 \leq k/2} H(\mathbf{u}). \quad \text{(B.39)}$$

To simplify the notation, we define $\mathcal{V}_k = \{\mathbf{v} \in \mathbb{R}^p, \text{ s.t. } \text{supp}(\mathbf{v}) \subseteq \tilde{S}^c, \|\mathbf{v}\|_0 \leq k/2, \|\mathbf{X}\mathbf{v}\| = 1\}$ and $c_{\mathbf{v}} = (\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}$. In addition to the inequalities in (B.39) we also have

$$\min_{\text{supp}(\mathbf{u}) \subseteq \tilde{S}^c, \|\mathbf{u}\|_0 \leq k/2} H(\mathbf{u}) \leq \min_{\mathcal{V}_k} H(c_{\mathbf{v}}\mathbf{v}) = \min_{\mathcal{V}_k} \big[- c_{\mathbf{v}}^2 \big] = - \max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}|^2.$$

Consequently,

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \leq - \max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}|^2. \tag{B.40}$$

Note that if $\|\mathbf{X}\mathbf{v}\| = 1$, then $|(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}| \geq |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| - \|\boldsymbol{\Delta}^*\|$. Also note that

$$\|\boldsymbol{\Delta}^*\| \leq \|\boldsymbol{\beta}^*\|_1 / \sqrt{k/2} \lesssim \sigma\gamma_k \sqrt{k \log(ep/k)}, \tag{B.41}$$

with a sufficiently small multiplicative constant due to the assumption on $\|\boldsymbol{\beta}^*\|_1$. Note that the cardinality of $\tilde{S}^c$ is at least $p/2$. Thus, applying Lemma 1, with $s = k/2$ and $q = p/2$, to lower bound $\max_{\mathcal{V}_k} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}|$, we derive that

$$\max_{\mathcal{V}_k} |(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top \mathbf{X}\mathbf{v}| \gtrsim \sigma\gamma_k \sqrt{k \log(ep/k)},$$

with probability at least $1 - 2(ep/k)^{-\tilde{c}\gamma_k^2 k/2}$. Thus, inequality (B.40) and the definition of $\hat{\boldsymbol{\beta}}_{\ell_0}$ yield

$$G(\hat{\boldsymbol{\beta}}_{\ell_0}) \leq \min_{\boldsymbol{\beta}:\ \|\boldsymbol{\beta}\|_0 \leq k} G(\boldsymbol{\beta}) \lesssim -\sigma^2\gamma_k^2 k \log(ep/k).$$

Because $2(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\tilde{\boldsymbol{\beta}}^* - \mathbf{X}\boldsymbol{\beta}) \leq G(\boldsymbol{\beta})$ for each $\boldsymbol{\beta}$, we derive

$$|(\boldsymbol{\Delta}^* + \boldsymbol{\epsilon})^\top (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \gtrsim \sigma^2\gamma_k^2 k \log(ep/k). \tag{B.42}$$

Taking into account (B.41), which holds with a sufficiently small multiplicative constant, we derive

$$|\boldsymbol{\Delta}^{*\top} (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \leq \|\boldsymbol{\Delta}^*\| \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \lesssim \sigma\gamma_k \sqrt{k \log(ep/k)} \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|. \tag{B.43}$$

Furthermore, on the event $\mathcal{G}_{2k}$ with $\delta_0 = (ep/k)^{-k}$, which holds with probability at least $1 - \delta_0$ by Theorem 1, we have

$$|\boldsymbol{\epsilon}^\top (\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*)| \lesssim \sigma \sqrt{k \log(ep/k)} \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\|. \tag{B.44}$$

Combining inequalities (B.42), (B.43) and (B.44), we arrive at

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| + \|\boldsymbol{\Delta}^*\| \geq \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \gtrsim \sigma\gamma_k \sqrt{k \log(ep/k)}.$$

Note that $\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\tilde{\boldsymbol{\beta}}^*\| \le \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| + \|\boldsymbol{\Delta}^*\|$, by the triangle inequality. Let $\psi = \tilde{c}/2$. Applying (B.41), which holds with a sufficiently small multiplicative constant, we conclude that

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\| \gtrsim \sigma\gamma_k\sqrt{k\log(ep/k)},$$

with probability at least $1 - 2(ep/k)^{-c\gamma_k^2 k} - (ep/k)^{-k}$.

B.10 Proof of Lemma 1

Note that if $s > q/2$, then we can establish the bound for $s = \lfloor q/2 \rfloor$ and use

$$\max_{\|\mathbf{v}\|_0 \le s,\, \|\mathbf{X}\mathbf{v}\|=1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}| \ge \max_{\|\mathbf{v}\|_0 \le \lfloor q/2 \rfloor,\, \|\mathbf{X}\mathbf{v}\|=1} |\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}|.$$

Hence, we will focus on the case $s \le q/2$.

We write $|\cdot|$ for the cardinality of a set. Applying Lemma F.1 in [5], which is closely related to the results in [57], we deduce that there exists a subset $\mathcal{H}$ of the set $\{-1, 0, 1\}^p$, with

$$\log\left(|\mathcal{H}|\right) \gtrsim s\log(eq/s),$$

such that $\mathrm{supp}(\mathbf{v}) \subset S$, $\|\mathbf{v}\|_0 \le s$, $\|\mathbf{X}\mathbf{v}\|^2 \le s$ and $\|\mathbf{v}_1 - \mathbf{v}_2\|^2 \ge s/4$, for all $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}$. Note that the last inequality implies

$$\|\mathbf{X}\mathbf{v}_1 - \mathbf{X}\mathbf{v}_2\|^2 \ge \gamma_{2s}^2 s/4.$$

Consequently, by Sudakov's minoration [for example, Proposition 3.15 in [39]],

$$E\max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \gtrsim \sigma\gamma_{2s}\sqrt{s\log(|\mathcal{H}|)} \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)}.$$

Define $W = \max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}$ and $v = \max_{\mathbf{v}\in\mathcal{H}} SD(\boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v})$ by v. By the concentration inequality for the supremum of a Gaussian process [for example, Theorem 3.12 in [39]], we have, for all $t \ge 0$,

$$P\left(W \le EW - vt\right) \le 2\exp(-t^2/2).$$

Note that

$$v \le \sigma \max_{\mathbf{v}\in\mathcal{H}} \|\mathbf{X}\mathbf{v}\| \le \sigma\sqrt{s}.$$

Consequently, if $t \le \gamma_{2s}\sqrt{\tilde{c}s\log(eq/s)}$ with a sufficiently small positive universal constant $\tilde{c}$, then $EW - vt \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)}$, and hence

$$\max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \gtrsim \sigma\gamma_{2s}s\sqrt{\log(eq/s)},$$

with probability at least $1 - 2\exp(-\tilde{c}\gamma_{2s}^2 s\log(eq/s))$. We complete the proof by noting that

$$\max_{\|\mathbf{v}\|_0 \le s,\, \|\mathbf{X}\mathbf{v}\|=1} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v} \ge s^{-1/2} \max_{\mathbf{v}\in\mathcal{H}} \boldsymbol{\epsilon}^\top \mathbf{X}\mathbf{v}.$$

B.11 Proof of Proposition 3

Note that the assumptions imposed on b imply

$$b \gtrsim 1 \qquad \text{and} \qquad b \lesssim 1, \tag{B.45}$$

where the universal constant in the second bound can be chosen to be sufficiently small. Also note that

$$\|\boldsymbol{\beta}^*\| = b\sigma\sqrt{[\log(ep)]/k^*} \qquad \text{and} \qquad \|\boldsymbol{\beta}^*\|_1 = b\sigma\sqrt{\log(ep)}. \tag{B.46}$$

Thus, to establish the result of Proposition 3, we only need to demonstrate that

$$\min_{k\in[0,p]} \|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\ell_0}\|^2 \gtrsim \sigma^2 \log(ep), \tag{B.47}$$

with high probability.

Let $\mathbf{1}$ denote a p-dimensional vector of ones, and note that

$$\|\mathbf{X}\boldsymbol{\beta}^*\|^2 \geq \rho_l \boldsymbol{\beta}^{*\top}\mathbf{1}\mathbf{1}^\top\boldsymbol{\beta}^* = \rho_l\|\boldsymbol{\beta}^*\|_1^2 = \rho_l b^2\sigma^2\log(ep) \gtrsim \sigma^2\log(ep).$$

We conclude that for $k = 0$ bound (B.47) holds with probability one. For the remainder of the proof we focus on the case of $k \in [p]$.

Minimizing the sum of squares is equivalent to minimizing the function

$$L(\boldsymbol{\beta}) = \|\mathbf{X}\boldsymbol{\beta}\|^2 - 2\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}.$$

Define $c_j = \mathbf{y}^\top\mathbf{X}_j$ and let $\mathbf{e}_j$ denote the j-th coordinate vector in $\mathbb{R}^p$. Because $\mathbf{X}\mathbf{e}_j = \mathbf{X}_j$ and $\|\mathbf{X}_j\| = 1$, we have

$$\min_{\|\boldsymbol{\beta}\|_0=1} L(\boldsymbol{\beta}) \leq \min_j L(c_j\mathbf{e}_j) = \min_j -c_j^2 = -\max_j |\mathbf{y}^\top\mathbf{X}_j|^2.$$

We also have

$$\begin{aligned}
\max_j |\mathbf{y}^\top\mathbf{X}_j|^2 &= \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2 + 2(\boldsymbol{\epsilon}^\top\mathbf{X}_j)(\mathbf{X}_j^\top\mathbf{X}\boldsymbol{\beta}^*) + |\mathbf{X}_j^\top\mathbf{X}\boldsymbol{\beta}^*|^2\Big) \\
&\geq \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2 - 2|\boldsymbol{\epsilon}^\top\mathbf{X}_j|\|\mathbf{X}\boldsymbol{\beta}^*\|\Big) \\
&\geq \max_j \Big(|\boldsymbol{\epsilon}^\top\mathbf{X}_j|^2/2 - 2\|\mathbf{X}\boldsymbol{\beta}^*\|^2\Big),
\end{aligned}$$

where we used bound (B.34) with $a = |\boldsymbol{\epsilon}^\top\mathbf{X}_j|$, $b = \|\mathbf{X}\boldsymbol{\beta}^*\|$ and $\alpha = 1/2$ to get the last inequality. Applying Lemma 1, we derive that

$$\max_j |\boldsymbol{\epsilon}^\top\mathbf{X}_j| \gtrsim (1-\rho_u)\sigma^2\log(ep),$$

with probability at least $1 - 2(ep)^{-\tilde{c}(1-\rho_u)}$, for some positive universal constant $\tilde{c}$.

Inequalities (B.46) and (B.45), together with the fact that columns of $\mathbf{X}$ have unit norm, yield

$$\|\mathbf{X}\boldsymbol{\beta}^*\|^2 \leq \|\boldsymbol{\beta}^*\|_1^2 \leq b^2\sigma^2 \log(ep) \lesssim \sigma^2 \log(ep), \tag{B.48}$$

with a sufficiently small universal constant. Consequently,

$$\min_{\|\boldsymbol{\beta}\|_0 \leq k} L(\boldsymbol{\beta}) \leq \min_{\|\boldsymbol{\beta}\|_0 = 1} L(\boldsymbol{\beta}) \lesssim -\sigma^2 \log(ep), \tag{B.49}$$

uniformly over $k \in [p]$ and with probability at least $1 - 2(ep)^{-a}$, for some positive universal constant a.

We conduct the rest of the argument on the high-probability event where (B.49) holds. On this event we have the bound

$$L(\hat{\boldsymbol{\beta}}_{\ell_0}) = \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 - 2\mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} \lesssim -\sigma^2 \log(ep), \tag{B.50}$$

in which the universal constant does not depend on k. Given a set $S \subseteq \{1, ..., p\}$, we define $\hat{\boldsymbol{\beta}}_S = \arg\min_{\text{supp}(\boldsymbol{\beta}) \subseteq S} L(\boldsymbol{\beta})$ and note that $\|\mathbf{X}\hat{\boldsymbol{\beta}}_S\|^2 = \mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_S$. Consequently, $\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 = \mathbf{y}^\top \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}$, and hence bound (B.50) implies

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\|^2 \gtrsim \sigma^2 \log(ep).$$

Bound (B.47) then follows from the inequality $\|\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\| \geq \|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0}\| - \|\mathbf{X}\boldsymbol{\beta}^*\|$ and bound (B.48), applied with a sufficiently small universal constant.

B.12 Proof of Proposition 4

Let $\mathbf{1}$ denote a p-dimensional vector of ones, and note that

$$\mathbf{X}^\top \mathbf{X} = (1 - \rho)\mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top.$$

Hence, for every $\mathbf{u} \in \mathbb{R}^p$,

$$\|\mathbf{X}\mathbf{u}\|^2 = (1 - \rho)\|u\|^2 + \rho(\mathbf{1}^\top \mathbf{u})^2 \geq (1 - \rho)\|u\|^2,$$

which implies $\gamma_k^2 \geq 1 - \rho$. Also note that, by (B.46), $\|\boldsymbol{\beta}^*\| = b\sigma\sqrt{[\log(ep)]/k^*}$, which can be made smaller than any given multiple of $\sigma\sqrt{k \log(ep/k)}$ under the assumptions imposed on b, k and k^* in Proposition 4. Under this scenario, we can apply Theorem 4, which leads to

$$\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\ell_0} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \gtrsim \sigma^2 k \log(ep/k).$$

B.13 Proof of Theorem 5

We first establish the bound for $\hat{\boldsymbol{\beta}}_2^B$, and then establish the one for $\hat{\boldsymbol{\beta}}_1^B$. We note that throughout the proof the positive multiplicative factors in inequalities $\lesssim$ and $\gtrsim$ are universal constants, which are independent from all other parameters such as n, p, σ, $\boldsymbol{\beta}^*$, $\boldsymbol{\beta}$ and s.

Expected prediction error bound for $\hat{\boldsymbol{\beta}}_2^B$.

To simplify the expressions, we drop the subscript and the superscript in $\hat{\boldsymbol{\beta}}_2^B$ and simply write $\hat{\boldsymbol{\beta}}$.

Taking an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^p$, we note that

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\| + \mu_{\hat{\boldsymbol{\beta}}}|||\hat{\boldsymbol{\beta}}\|_0 \leq \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{B.51}$$

In the setting where $a \gtrsim \sigma$, with a sufficiently large multiplicative constant, we will bound the term $2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\|$. Similarly, in the case $b \gtrsim \sigma^2$ we will bound $2\boldsymbol{\epsilon}^\top \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \mu_{\hat{\boldsymbol{\beta}}}|||\hat{\boldsymbol{\beta}}\|_0$.

We first consider the case $a \gtrsim \sigma$, which implies $\lambda_{\boldsymbol{\beta}} \gtrsim \sigma\sqrt{\|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0)}$. We let $\hat{\mathbf{u}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and restrict our attention to event $\mathcal{F}$, defined in Section 3.1, which holds with probability at least $1 - \delta_0/2$. On event $\mathcal{F}$, we have either $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)}$ or $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma\sqrt{\log(1/\delta_0)}\|\mathbf{X}\hat{\mathbf{u}}\|$. In the latter scenario, repeating the argument after inequality (B.32) in the proof of Theorem 2 yields

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \sigma^2 \log(1/\delta_0) + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{B.52}$$

We now focus on the event $\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \lesssim \sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)}$. In view of (B.28), we have

$$\begin{aligned}
\sigma \sum_{j=1}^p \hat{u}_j^\sharp \sqrt{\log(2p/j)} &\leq \sigma \sum_{j=1}^p \hat{\beta}_j^\sharp \sqrt{\log(2p/j)} + \sigma \sum_{j=1}^p \beta_j{}^\sharp \sqrt{\log(2p/j)} \\
&\leq \sigma\sqrt{\hat{k}\log(2ep/\hat{k})}\|\hat{\boldsymbol{\beta}}\| + \sigma\sqrt{\|\boldsymbol{\beta}\|_0 \log(2ep/\|\boldsymbol{\beta}\|_0)}\|\boldsymbol{\beta}\|.
\end{aligned}$$

Thus, if $a \gtrsim \sigma$ with a sufficiently large universal constant, then $2\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \leq \lambda_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\| + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|$. Combining this bound with inequality (B.51), we derive

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0. \tag{B.53}$$

Bounds (B.52) and (B.53) imply that, for each $\delta_0 \in (0,1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}_2}|||\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \sigma^2 \log(1/\delta_0) + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 \tag{B.54}$$

with probability at least $1 - \delta_0/2$.

We now focus on the case $b \gtrsim \sigma^2$, which implies $\mu_{\boldsymbol{\beta}} \gtrsim \sigma^2 \log(ep/\|\boldsymbol{\beta}\|_0)$. Given an $s \in [p]$, we consider the event $\mathcal{G}_s$, defined in Section 3.1, where we take $\delta_0 = (s/[ep])^s \epsilon_0$. Here, $\epsilon_0 \in (0,1)$ is an arbitrary value that does not depend on s. On the event $\mathcal{G} = \cap_{s=1}^p \mathcal{G}_s$ we have

$$\begin{aligned}\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} &\lesssim \sigma \sqrt{\|\hat{\mathbf{u}}\|_0 \log(ep/\|\hat{\mathbf{u}}\|_0) + \log(1/\epsilon_0)} \|\mathbf{X}\hat{\mathbf{u}}\| \\ &\lesssim \sigma \sqrt{\|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) + \|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0) + \log(1/\epsilon_0)} \|\mathbf{X}\hat{\mathbf{u}}\|.\end{aligned}$$

Noting that $\|\mathbf{X}\hat{\mathbf{u}}\| \le \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\| + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|$ and applying inequality (B.34) twice, we derive

$$2\boldsymbol{\epsilon}^\top \mathbf{X}\hat{\mathbf{u}} \le c\Big[\sigma^2 \|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) + \sigma^2 \|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0) + \sigma^2 \log(1/\epsilon_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2\Big] + \|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2,$$

for some universal constant c. Taking into account inequality (B.51), we deduce that

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + 2[b - c\sigma^2]\|\hat{\boldsymbol{\beta}}\|_0 \log(ep/\|\hat{\boldsymbol{\beta}}\|_0) \lesssim \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 + \sigma^2 \log(1/\epsilon_0) + \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|$$

on the event $\mathcal{G}$. Note that

$$\mathbb{P}(\mathcal{G}^c) \le \sum_{s=1}^p \mathcal{P}(\mathcal{G}_s^c) \le \sum_{s=1}^p (s/[ep])^s \epsilon_0 \le \sum_{s=1}^p e^{-s}\epsilon_0 \le \epsilon_0.$$

Consequently, if we let $b \ge 2c\sigma^2$, then

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 \log(1/\epsilon_0) + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0 \tag{B.55}$$

with probability at least $1 - \epsilon_0$.

Combining bounds (B.54) and (B.55), we conclude that, for each $\delta_0 \in (0,1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2 \log(1/\delta_0) \tag{B.56}$$

with probability at least $1 - \delta_0$. Repeating the argument in the proof of Corollary 3, we derive

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\| + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2, \tag{B.57}$$

which establishes the first bound in the statement of Theorem 5.

Expected prediction error bound for $\hat{\boldsymbol{\beta}}_1^B$.

To simplify the expressions, we drop the subscript and the superscript in $\hat{\boldsymbol{\beta}}_1^B$ and simply write $\hat{\boldsymbol{\beta}}$.

In the case $b \gtrsim \sigma^2$ we repeat the argument in the corresponding part of the proof for $\hat{\boldsymbol{\beta}}_2^B$ to derive a counterpart of inequality (B.56). We deduce that, for each $\delta_0 \in (0,1)$,

$$\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \mu_{\hat{\boldsymbol{\beta}}}\|\hat{\boldsymbol{\beta}}\|_0 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2 \log(1/\delta_0) \tag{B.58}$$

with probability at least $1-\delta_0$.

In the case $\lambda \geq c_0\sigma\sqrt{\log(ep)}$, we repeat the argument in the second slow rate part of the proof of Theorem 3 to derive a slight modification of inequality (B.38), containing the additional ℓ_0 penalty terms. Thus, we again deduce that inequality (B.58) holds for each $\delta_0 \in (0,1)$ with probability at least $1-\delta_0$.

As before, starting with probability bound (B.58) and repeating the argument in the proof of Corollary 3 we conclude that

$$\mathbb{E}\|\mathbf{f}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \lesssim \inf_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big[\|\mathbf{f}^* - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 + \mu_{\boldsymbol{\beta}}\|\boldsymbol{\beta}\|_0\Big] + \sigma^2, \tag{B.59}$$

which establishes the second bound in the statement of Theorem 5.

B.14 Proof of Corollary 6

The error rates in Corollary 6 follow directly from inequalities (B.57) and (B.59).

To control the sparsity of $\hat{\boldsymbol{\beta}}_2^B$, we first establish that inequality $\|\hat{\boldsymbol{\beta}}_2^B\|_0 \lesssim (k^* \vee 1)\{1+[\log(1/\delta_0)]^2\}$ holds with probability at least $1-\delta_0$, and then use this fact to bound $\mathbb{E}\|\hat{\boldsymbol{\beta}}_2^B\|_0$.

We define $g(x) = \log(ep/x)x$ for $x \in [0,p]$, with $g(0)=0$. We note that $g(x) \geq x$, and $g(x)$ is monotone increasing and continuous. We also note that if $C>0$, $x_1 \in [0,p]$ and $x_2 \in [0,p]$, then

$$g(x_1) \leq Cg(x_2) \quad \Rightarrow \quad x_1 \leq 2C[1 \vee \log(2C)]x_2. \tag{B.60}$$

To establish the above relationship we note that

$$g(x_1) \leq Cg(x_2) \leq g(2Cx_2)/2 + Cx_2\log(2C) \leq \max\{g(2Cx_2), 2Cx_2\log(2C)\}$$

and consider two cases. If $g(x_1) \leq g(2Cx_2)$, then $x_1 \leq 2Cx_2$. Alternatively, if $g(x_1) \leq 2Cx_2\log(2C)$, then $x_1 \leq 2C\log(2C)x_2$.

We write $\hat{k} = \|\hat{\boldsymbol{\beta}}_2^B\|_0$ and note that under each corresponding set of assumptions on the tuning parameters in Corollary 6, inequality (B.56) yields $(\mu_{\hat{\boldsymbol{\beta}}_2^B})\hat{k} \lesssim b\log(ep/[k^*\vee 1])[k^*\vee 1]\{1+\log(1/\delta_0)\}$. Taking into account $(\mu_{\hat{\boldsymbol{\beta}}_2^B})\hat{k} = bg(\hat{k})$, we rewrite the last inequality as $g(\hat{k}) \lesssim g(k^*\vee 1)\{1+\log(1/\delta_0)\}$. Hence, by property B.60, we have $\hat{k} \lesssim \{1+\log(1/\delta_0)\}[1\vee\log(1+2\log(1/\delta_0))](k^*\vee 1)$. Consequently, $\hat{k}/[k^*\vee 1] \lesssim 1+[\log(1/\delta_0)]^2$ with probability at least $1-\delta_0$. Finally, we bound $\mathbb{E}\hat{k}/[k^*\vee 1]$ using an argument analogous to the one in the proof of Corollary 3: $\mathbb{E}\hat{k}/[k^*\vee 1] \lesssim 1+\int_0^\infty e^{-w^{1/2}}dw \lesssim 1$.

To establish the sparsity bound for $\hat{\boldsymbol{\beta}}_1^B$, we note that for all of corresponding tuning parameter settings in Corollary 6, inequality (B.58) yields $(\mu_{\hat{\boldsymbol{\beta}}_1^B})\|\hat{\boldsymbol{\beta}}_1^B\|_0 \lesssim b\log(ep/[k^*\vee 1])[k^*\vee 1]\{1+\log(1/\delta_0)\}$, with probability at least $1-\delta_0$. The sparsity bound then follows by repeating the argument in the last paragraph of the proof for $\hat{\boldsymbol{\beta}}_2^B$.

C Additional experiments

C.1 Ultra-high dimensional examples

Figures 6 and 7 summarize the results of additional experiments corresponding to the challenging ultra-high dimensional setting [57] with $k^*\log(p/k^*) > n/2$. These experiments complement the ones that are reported in Figure 2. Qualitatively, the results are overall similar to those in Figure 2, especially with respect to the effect of adding ℓ_1 or ℓ_2 regularization to best subset selection. However, the predictive performance of all the methods in the challenging ultra-high dimensional setting is significantly worse than before, while the corresponding relative standing of dense models such as Ridge, Elastic net, and Lasso is improved.

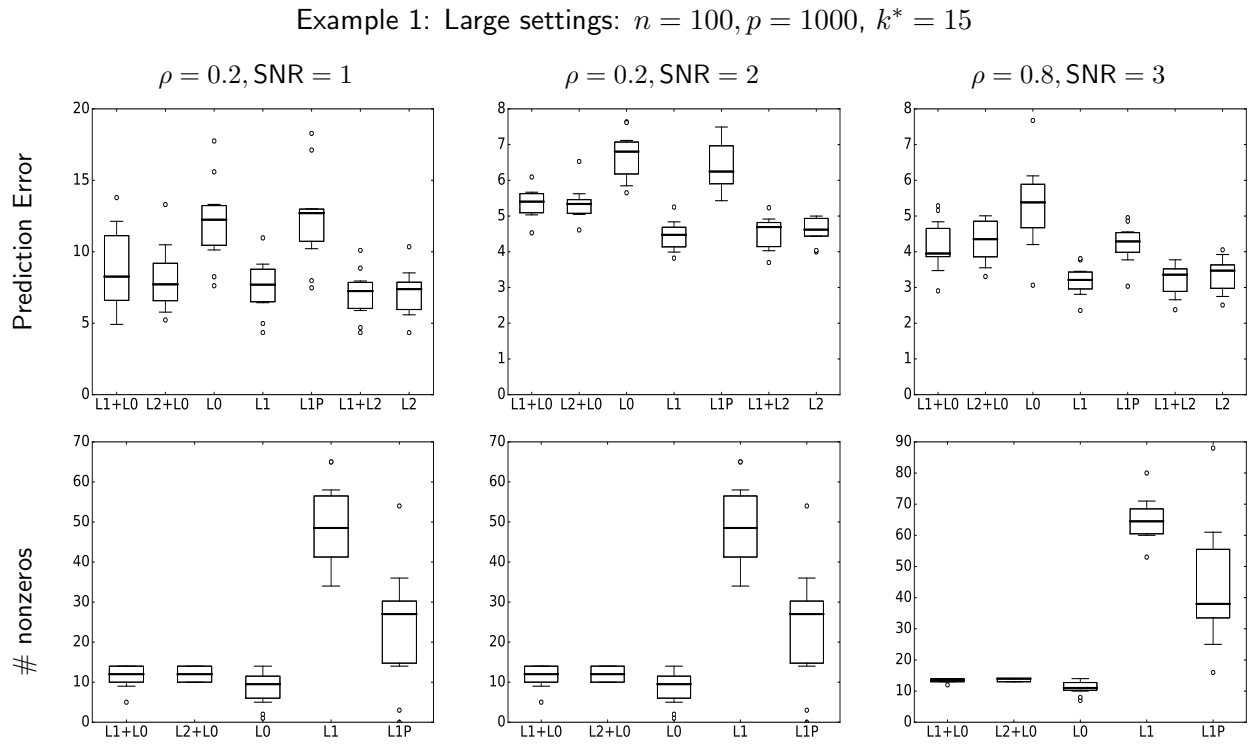


Figure 6: Example 1 simulations for different values of n, p, ρ, and SNR.

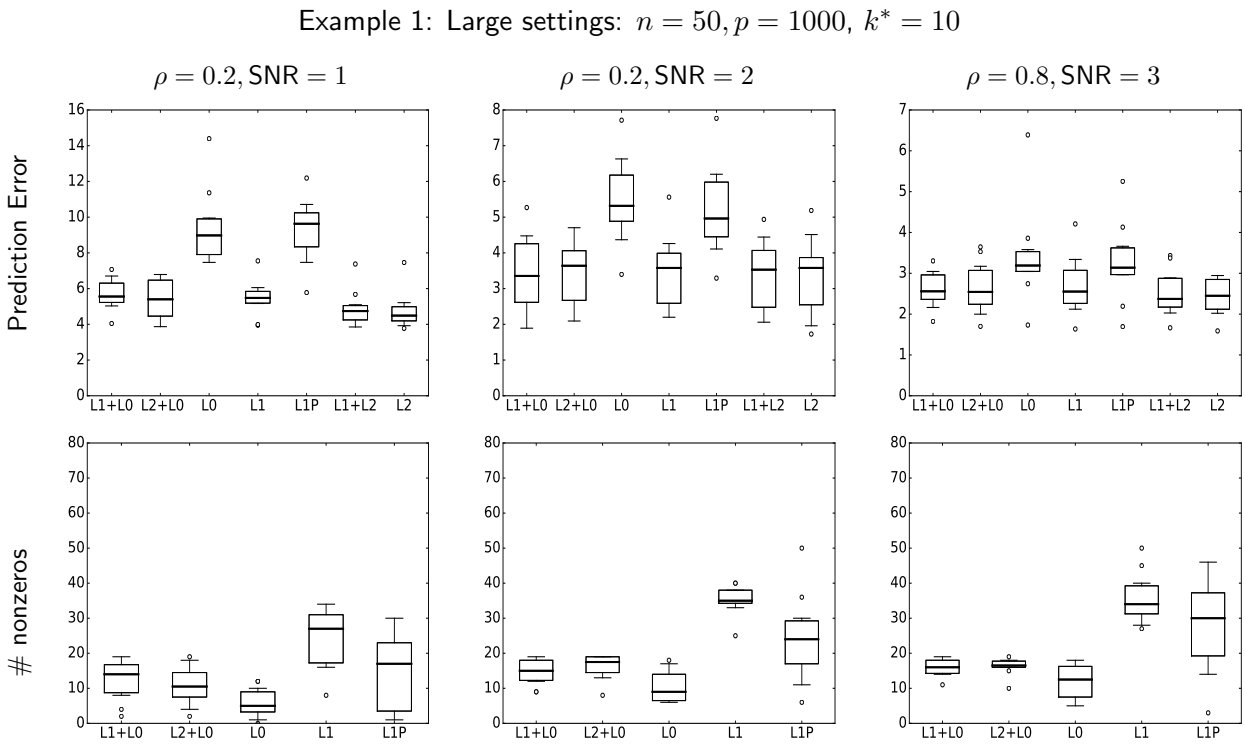


Figure 7: Example 1 simulations for different values of n, p, ρ, and SNR.

C.2 Comparisons with Bayesian methods

Figure 8 summarizes the results of additional experiments that include three state-of-the-art Bayesian approaches: the spike-and-slab Lasso method [53], implemented using R package `SSLASSO`; the empirical Bayes method of [38], implemented using R package `ebreg`; and the horseshoe regression [15], implemented using R package `horseshoe`. In the experiments that we consider, and with the default settings for the tuning parameters, the predictive performance of the last two methods is not quite as good as that of the competitors. The predictive performance of spike-and-slab Lasso is on par with the best performing methods (but somewhat worse overall than that of the proposed approach); however, their models are denser than those of the proposed approach. Overall, the proposed approach performed favorably in terms of both the model sparsity and the prediction accuracy. We note that the experiments in the top two panels of Figure 8 are the same as those in Figure 4; however, Figure 8 also includes the results for `horseshoe` and `ebreg`.

References

[1] E. H. Aarts and J. K. Lenstra. *Local search in combinatorial optimization*. Princeton University Press, 1997.

[2] A. Amini, U. S. Kamilov, and M. Unser. The analog formulation of sparsity implies infinite divisibility and rules out bernoulli-gaussian priors. In *2012 IEEE Information Theory Workshop*, pages 682–686. Ieee, 2012.

[3] A. Atamturk and A. Gomez. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

[4] P. L. Bartlett, S. Mendelson, and J. Neeman. L1-regularized linear regression: persistence and oracle inequalities. *Probability theory and related fields*, 154(1-2):193–224, 2012.

[5] P. C. Bellec, G. Lecué, and A. B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

[6] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[7] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

[8] D. Bertsimas and R. Weismantel. *Optimization over integers*. Dynamic Ideas Belmont, 2005.

[9] D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.

[10] A. Bhadra, J. Datta, Y. Li, N. G. Polson, and B. Willard. Prediction risk for the horseshoe regression. *The Journal of Machine Learning Research*, 20(1):2882–2920, 2019.

[11] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[13] L. Breiman. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24 (6):2350–2383, 1996.

[14] P. Bühlmann and S. van-de-Geer. *Statistics for high-dimensional data*. Springer, 2011.

[15] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[16] L. Comminges, A. S. Dalalyan, et al. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012.

[17] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23 (1):552–581, 2017.

[18] Y. Fan and J. Lv. Asymptotic properties for combined L1 and concave regularization. *Biometrika*, 101 (1):57–70, 2013.

[19] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–148, 1993.

[20] D. Gamarnik and I. Zadik. High dimensional regression with binary coefficients. estimating squared error and a phase transtition. In *Conference on Learning Theory*, pages 948–953, 2017.

[21] E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

[22] E. Greenshtein. Best subset selection, persistence in high-dimensional statistical learning and optimiza-

tion under ℓ_1 constraint. *The Annals of Statistics*, 34(5):2367–2386, 2006.

[23] T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16:3367–3402, 2015.

[24] T. Hastie, R. Tibshirani, and R. Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

[25] H. Hazimeh and R. Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *arXiv preprint arXiv:1803.01454*, 2018.

[26] H. Hazimeh, R. Mazumder, and A. Saab. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *arXiv preprint arXiv:2004.06152*, 2020.

[27] H. Hazimeh, R. Mazumder, and P. Radchenko. Grouped variable selection with discrete optimization: Computational and statistical perspectives. *arXiv preprint arXiv:2104.07084*, 2021.

[28] A. E. Hoerl and R. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[29] J. Huang, P. Breheny, S. Lee, S. Ma, and C. Zhang. The mnet method for variable selection. *Statistica Sinica*, 26:903?923, 2016.

[30] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

[31] V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[32] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8), 2009.

[33] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904, 2017.

[34] J. T. Linderoth and A. Lodi. MILP software. *Wiley encyclopedia of operations research and management science*, 2010.

[35] Y. Liu and Y. Wu. Variable selection via a combination of the l0 and l1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798, 2007.

[36] K. Lounici, M. Pontil, A. Tsybakov, and S. Geer. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164–2204, 2011.

[37] R. Martin and Y. Tang. Empirical priors for prediction in sparse high-dimensional linear regression. *Journal of Machine Learning Research*, 21(144):1–30, 2020.

[38] R. Martin, R. Mess, and S. G. Walker. Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017.

[39] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[40] R. Mazumder and P. Radchenko. The Discrete Dantzig Selector: Estimating sparse linear models via mixed integer linear optimization. *IEEE Transactions on Information Theory*, 63 (5):3053 – 3075, 2017.

[41] N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[42] A. Miller. *Subset selection in regression*. CRC Press Washington, 2002.

[43] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[44] N. Mladenović and P. Hansen. Variable neighborhood search. *Computers & operations research*, 24(11): 1097–1100, 1997.

[45] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):

227–234, 1995.

[46] G. L. Nemhauser and L. A. Wolsey. Integer programming and combinatorial optimization. *Wiley, Chichester. GL Nemhauser, MWP Savelsbergh, GS Sigismondi (1992). Constraint Classification for Mixed Integer Programming Formulations. COAL Bulletin*, 20:8–12, 1988.

[47] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140 (1):125–161, 2013.

[48] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer, Norwell, 2004.

[49] G. Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire Analyse fonctionnelle (dit" Maurey-Schwartz")*, pages 1–12, 1980.

[50] N. G. Polson and L. Sun. Bayesian ℓ_0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

[51] G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[52] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

[53] V. Rocková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[54] C. Soussen, J. Idier, D. Brie, and J. Duan. From bernoulli–gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584, 2011.

[55] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[57] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[58] J. P. Vielma, I. Dunning, J. Huchette, and M. Lubin. Extended formulations in mixed integer conic quadratic programming. *Mathematical Programming Computation*, pages 1–50, 2016.

[59] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using l1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 2009.

[60] H. Weng, Y. Feng, and X. Qiao. Regularization after retention in ultrahigh dimensional linear regression models. *arXiv preprint arXiv:1311.5625*, 2013.

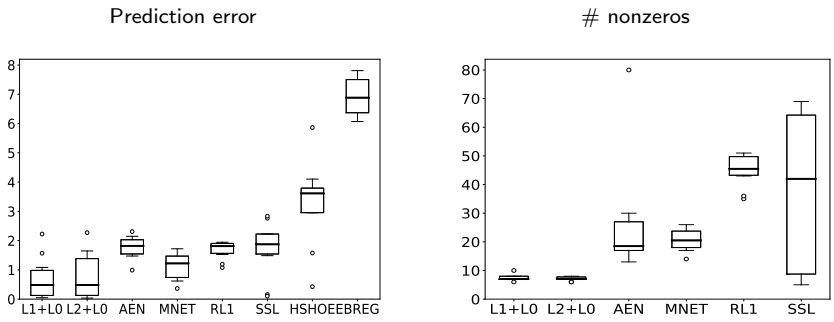
[61] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

[62] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

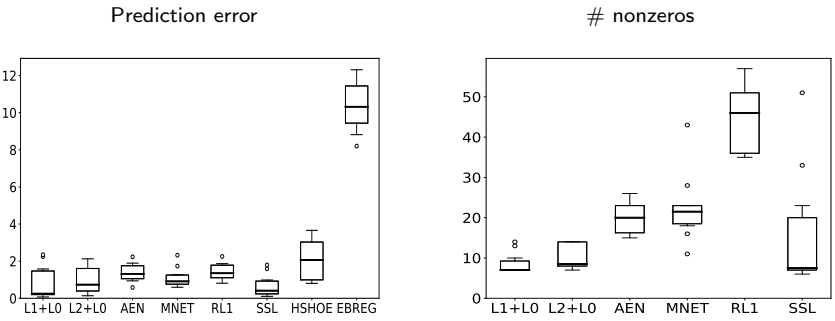
[63] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B.*, 67(2):301–320, 2005.

[64] H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

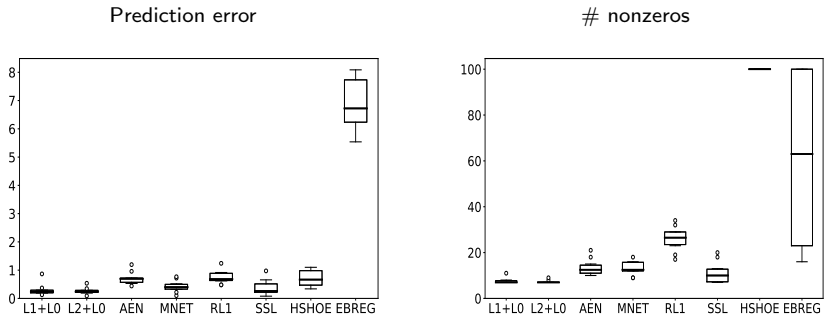
Example 1: $n = 100, p = 1000,\ \rho = 0.2$, SNR=2.



Example 2: $n = 100, p = 1000,\ \rho = 0.1$, SNR=3.



Example 1: $n = 100, p = 100,\ \rho = 0.2$, SNR=2.



Example 2: $n = 100, p = 100,\ \rho = 0.1$, SNR=3.

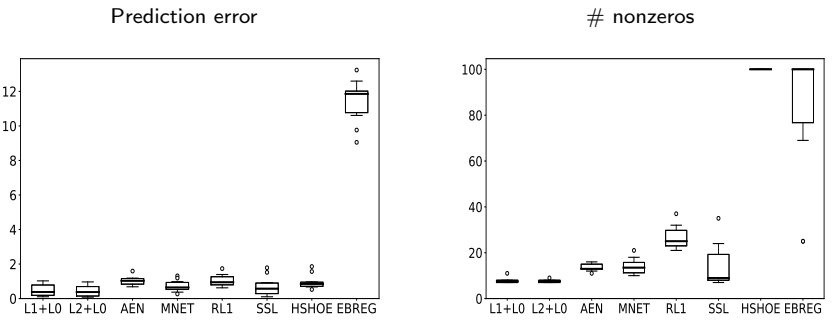


Figure 8: Experimental results for the proposed methods, L0+L1 and L0+L2, as well as adaptive elastic net (AEN), Mnet, relaxed Lasso (RL1), `SSLASSO` (SSL), `horseshoe` (HSHOE), and `ebreg` (EBREG) methods. Due to the density of the corresponding solutions, we do not report the sparsity for HSHOE and EBREG in the top two panels.