A. Proof of the Main Results

In this section, we lay out the proofs of the main results presented in §3. We first establish the statistical rate of convergence for the proposed estimator, and then show the optimality of our procedure by deriving the minimax lower bound.

A.1. Proof of Theorem 1

Proof. For any stationary point $\widehat{\boldsymbol{\beta}}$ of the optimization problem in (1.2), by definition we have

$$\nabla L(\widehat{\boldsymbol{\beta}}) + \lambda \cdot \boldsymbol{\xi} = \mathbf{0}, \;\; \text{where } \boldsymbol{\xi} \in \partial \|\widehat{\boldsymbol{\beta}}\|_1.$$

For notational simplicity, we denote $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ as $\boldsymbol{\delta}$. By definition, we have

$$\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^* \rangle = \langle -\lambda \cdot \boldsymbol{\xi} - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle.$$

We denote the support of $\boldsymbol{\beta}^*$ as $\mathcal{S}$, that is, $\mathcal{S} = \{j \colon \beta_j^* \neq 0\}$. By writing $\boldsymbol{\xi} = \boldsymbol{\xi}_{\mathcal{S}} + \boldsymbol{\xi}_{\mathcal{S}^c}$ we have

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &= \langle -\lambda \cdot \boldsymbol{\xi}_{\mathcal{S}^c} - \lambda \cdot \boldsymbol{\xi}_{\mathcal{S}} - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle. \end{aligned} \tag{A.1}$$

Note that $\boldsymbol{\beta}^*_{\mathcal{S}} = \mathbf{0}$ and $\langle \boldsymbol{\xi}_{\mathcal{S}^c}, \widehat{\boldsymbol{\beta}}_{\mathcal{S}^c} \rangle = \|\widehat{\boldsymbol{\beta}}_{\mathcal{S}^c}\|_1 = \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1$. By Hölder's inequality, since $\|\boldsymbol{\xi}\|_\infty \leq 1$, the right-hand side of (A.1) can be bounded by

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \|\nabla L(\boldsymbol{\beta})\|_\infty \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{A.2}$$

Now we invoke Lemma 3 to bound the right hand side of (A.2). In what follows, we condition on the event that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \leq B\sigma \cdot \sqrt{\log d / n}$, which holds with probability at least $1 - \delta$, where $\delta \geq (2d)^{-1}$. By the definition of λ, we have $\lambda \geq L_1 \cdot \|\nabla L(\boldsymbol{\beta})\|_\infty$ with probability at least $1 - \delta$. By (A.2) we have

$$\begin{aligned} &\langle \nabla L(\widehat{\boldsymbol{\beta}}) - \nabla L(\boldsymbol{\beta}^*), \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \\ &\leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + L_1^{-1} \lambda \|\boldsymbol{\delta}\|_1. \end{aligned} \tag{A.3}$$

Now we invoke Lemma 4 to establish a lower bound of the left-hand side of (A.3). Combining (3.3), (A.3) and Lemma 4 we obtain that

$$\begin{aligned} 0 &\leq a^2 \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} \leq -\lambda \|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 + \mu \lambda \|\boldsymbol{\delta}\|_1 \\ &= -\lambda(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 + \lambda(1+\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1. \end{aligned} \tag{A.4}$$

where $\mu = L_1^{-1} + 3b\sqrt{D} L_2^{-1} \leq 0.1$. Hence it follows that $\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq (1+\mu)/(1-\mu)\|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \leq 1.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1$.

Now we invoke the Lemma 5 to bound $\boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta}$ from below. Under Assumption 1, we have

$$\rho_+(k^*)/\rho_-(s^* + 2k^*) \leq 1 + 0.5k^*/s^*.$$

Combining this inequality with Lemma 5 we obtain that

$$\begin{aligned} \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} &\geq \big(1 - 1.23\sqrt{0.5}\big) \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2 \\ &\geq 0.1 \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2, \end{aligned} \tag{A.5}$$

where $\mathcal{J}$ is the set of indices of the largest k^* entries of $\boldsymbol{\delta}_{\mathcal{S}^c}$ in absolute value and $\mathcal{I} = \mathcal{J} \cup \mathcal{S}$. Here the first inequality of (A.5) follows from Lemma 5 and that $\mathcal{S} \subset \mathcal{I}$. Combining (A.4) and (A.5) we obtain that

$$\begin{aligned} 0.1 \cdot \rho_-(s^* + k^*) \cdot \|\boldsymbol{\delta}_{\mathcal{I}}\|_2^2 &\leq \boldsymbol{\delta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\delta} \leq a^{-2} \lambda (1+\mu) \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \\ &\leq 1.1 \cdot a^{-2} \sqrt{s^*} \lambda \|\boldsymbol{\delta}_{\mathcal{I}}\|_2, \end{aligned}$$

which implies that $\|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \leq 11/\rho_-(s^* + k^*) \cdot a^{-2}\sqrt{s^*}\lambda$. Note that by Lemma 5 we also have $\|\boldsymbol{\delta}\|_2 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{I}}\|_2$. Combining this inequality with the fact that $\|\boldsymbol{\delta}_{\mathcal{S}^c}\|_1 \leq 1.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1$, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 &= \|\boldsymbol{\delta}\|_1 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{S}}\|_1 \leq 2.23 \sqrt{s^*} \|\boldsymbol{\delta}_{\mathcal{S}}\|_2 \\ &\leq 25/\rho_-(s^* + k^*) \cdot a^{-2} s^* \lambda; \\ \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 &= \|\boldsymbol{\delta}\|_2 \leq 2.23 \|\boldsymbol{\delta}_{\mathcal{I}}\|_2 \\ &\leq 25/\rho_-(s^* + k^*) \cdot a^{-2} \sqrt{s^*} \lambda. \end{aligned}$$

Finally, to show that Algorithm 1 indeed catches a stationary point, we note that the acceptance criterion of the Algorithm (Line 1) implies that $\phi(\boldsymbol{\beta}^{(1)}) \leq \phi(\boldsymbol{\beta}^{(0)})$ where $\phi(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$. Moreover, for $t = 2$, we also have $\phi(\boldsymbol{\beta}^{(2)}) \leq \max\{\phi(\boldsymbol{\beta}^{(0)}), \phi(\boldsymbol{\beta}^{(1)})\}$. By induction, we conclude that for all $t \geq 1$, $\phi(\boldsymbol{\beta}^{(t)}) \leq \phi(\boldsymbol{\beta}^{(0)})$. Therefore we have $\boldsymbol{\beta}^{(t)} \in \mathcal{C} := \big\{\boldsymbol{\beta} \in \mathbb{R}^d \colon \|\boldsymbol{\beta}\|_1 \leq \lambda^{-1} \cdot L(\boldsymbol{\beta}^{(0)}) + \|\boldsymbol{\beta}^{(0)}\|_1\big\}$. Since set $\mathcal{C}$ is compact and the loss function L is continuously differentiable, it is also Lipschitz on $\mathcal{C}$. Therefore, by the convergence result of in Theorem 1 of [51], we conclude that every accumulation point of Algorithm 1 is a stationary point of optimization problem (1.2). □

A.2. Proof of Theorem 2

In what follows, inspired by [39], we apply Fano's method to derive the minimax risk of estimation for the nonlinear regression model defined in (1.1).

Proof. Let $M = M(\delta_n)$ be the cardinality of a $2\delta_n$-packing set of $B_0(s)$ with respect to the ℓ_2-metric where δ_n will be specified later. We denote the elements of this packing set as $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M\}$. For any estimator $\widehat{\boldsymbol{\beta}}$, let $\psi = \operatorname{argmin}_{i \leq M} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2$, triangle inequality implies that

$$\begin{aligned} 2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2 &\geq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^i\|_2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\psi\|_2 \\ &\geq \|\boldsymbol{\beta}^i - \boldsymbol{\beta}^\psi\|_2 \geq 2\delta_n \;\; \text{for } \; i \neq \psi. \end{aligned}$$

Thus we conclude that

$$\begin{aligned}\mathcal{R}_f^*(s,n,d) &\geq \inf_{\psi} \sup_{1\leq i\leq M} \delta_n^2 \cdot \mathbb{P}_{\boldsymbol{\beta}^i}(\psi \neq i) \\ &\geq \inf_{\psi} \delta_n^2 \cdot \mathbb{P}_{\boldsymbol{\beta}^U}(\psi \neq U),\end{aligned}$$

where U is uniform distributed over $\{1,\ldots,N\}$. We consider the following data-generating process: For a continuously differentiable function f with $f'(u) \in [a,b], \forall u \in \mathbb{R}$, we first sample a random variable U uniformly over $1,\ldots M$, then generate data $y_i = f(\mathbf{x}_i^\top \boldsymbol{\beta}^U) + \epsilon_i$. Fano's inequality implies that

$$\mathbb{P}(\psi \neq U) \geq 1 - \big[I(U; y_1,\ldots,y_n) + \log 2\big] \big/ \log N.$$

In what follows, we establish an upper bound for the mutual information $I(U; y_1,\ldots,y_n)$. For $s \in \{1,\ldots,d\}$, we define the high-dimensional sparse hypercube as $\mathcal{C}_o(s) \coloneqq \big\{\mathbf{v} \in \{0,1\}^d, \|\mathbf{v}\|_0 = s\big\}$. We define the Hamming distance on $\mathcal{C}_0(s)$ as $\rho_H(\mathbf{v},\mathbf{v}') = \sum_{i=1}^d \mathbb{1}\{v_i \neq v_i'\}$. The following lemma, obtained from [40], is an extension of the Varhsamov-Gilbert lemma to $\mathcal{C}_0(s)$.

Lemma 6 (Sparse Varshamov-Gilbert lemma)**.** For any two integers s and d satisfying $1 \leq s \leq d/8$, there exist $\mathbf{v}_1,\ldots,\mathbf{v}_M \in \{0,1\}^d$ with $\|\mathbf{v}_i\|_0 = s$ for $1 \leq i \leq M$ such that

$$\begin{aligned}\rho_H(\mathbf{v}_i,\mathbf{v}_j) \geq s/2 \text{ for all } i \neq j, \text{ and} \\ \log(M) \geq s/8 \cdot \log[1 + d/(2s)].\end{aligned}$$

By Lemma 6 there exist $\mathcal{C}' \subset \mathcal{C}_0$ with $|\mathcal{C}'| \geq \exp\big\{s/8 \cdot \log[1+d/(2s)]\big\}$ such that $\rho_H(\mathbf{v},\mathbf{v}') \geq s/2$ for all $\mathbf{v},\mathbf{v}' \in \mathcal{C}'$. Then for $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathcal{C} \coloneqq \delta_n \cdot \sqrt{2/s} \cdot \mathcal{C}'$, we have

$$\begin{aligned}\delta_n^2 \cdot 2/s \cdot \rho_H(\boldsymbol{\beta},\boldsymbol{\beta}') &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 \\ &\leq 2(\|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}'\|_2^2) \leq 8\delta_n^2,\end{aligned}$$

which implies that $\delta_n^2 \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 \leq 8\delta_n^2$ for all $\boldsymbol{\beta},\boldsymbol{\beta}' \in \mathcal{C}$. By the convexity of mutual information, we have $I(U; y_1,\ldots,y_n) \leq M^{-2} \sum_{1\leq m,m'\leq M} D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'})$. Since given $\boldsymbol{\beta}$ and f, $y_i \sim N\big(f(\mathbf{x}_i^\top \boldsymbol{\beta}), \sigma^2\big)$, direct computation yields that

$$\begin{aligned}&D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'}) \\ &= 1/(2\sigma^2) \sum_{i=1}^n [f(\mathbf{x}_i^\top \boldsymbol{\beta}^m) - f(\mathbf{x}_i^\top \boldsymbol{\beta}^{m'})]. \qquad \text{(A.6)}\end{aligned}$$

By mean-value theorem, (A.6) can be bounded by

$$\begin{aligned}&D_{KL}(\boldsymbol{\beta}^m, \boldsymbol{\beta}^{m'}) \\ &\leq n \cdot b^2/(2\sigma^2)(\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'})^\top \widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}) \\ &\leq n \cdot b^2 \cdot \rho_+(2s)/(2\sigma^2)\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}\|_2^2 \\ &\leq 4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2,\end{aligned}$$

where the second inequality follows from $\|\boldsymbol{\beta}^m - \boldsymbol{\beta}^{m'}\|_0 \leq 2s$. Therefore we conclude that $I(U; y_1,\ldots,y_n) \leq 4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2$, which yields that

$$\begin{aligned}\inf_{\psi} \mathbb{P}_{\boldsymbol{\beta}}(\psi \neq U) &\geq 1 - \frac{4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2 + \log 2}{\log M} \\ &\geq 1 - \frac{4nb^2 \cdot \rho_+(2s) \cdot \delta_n^2/\sigma^2 + \log 2}{s/8 \cdot \log[1 + d/(2s)]}.\end{aligned}$$

Setting $\delta_n^2 = \frac{\sigma^2 s \log[1+d/(2s)]}{96nb^2\rho_+(2s)}$, since $s \geq 4$ and $d \geq 8s$, we conclude that the right-hand side is no less than $1/2$. Now we obtain the following minimax lower bound

$$\mathcal{R}_f^*(s,n,d) \geq \frac{\sigma^2}{192b^2\rho_+(2s)} \frac{s\log[1+d/(2s)]}{n}.$$

This concludes the proof of Theorem 2. □

B. Proof of Auxiliary Results

In this appendix, we provide the proofs of the auxiliary lemmas appearing in the proof of the main results.

Proof of Lemma 3. By the definition of loss function L, for $j = 1,\ldots,d$, the j-th entry of $\nabla L(\boldsymbol{\beta}^*)$ can be written as $\nabla_j L(\boldsymbol{\beta}^*) = 1/n \cdot \sum_{i=1}^n \epsilon_i f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) x_{ij}$. Recall that ϵ_i's are i.i.d. centered sub-Gaussian random variables with variance proxy σ^2. Thus conditioning on $\mathbf{x}_1,\ldots,\mathbf{x}_n$, $\nabla_j L(\boldsymbol{\beta}^*)$ is a centered sub-Gaussian random variable with variance proxy bounded by

$$\sigma^2 \cdot \frac{1}{n^2} \sum_{i=1}^n f'(\mathbf{x}_i^\top \boldsymbol{\beta}^*)^2 x_{ij}^2 \leq \sigma^2 \cdot b^2 \cdot \widehat{\Sigma}_{j,j}/n,$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. Under Assumption *Bounded-Design*(D), the variance proxy of $\nabla_j L(\boldsymbol{\beta}^*)$ is bounded by $\sigma^2 \cdot b^2 \cdot D/n$. By the definition of variance proxy of sub-Gaussian random variables, we have

$$\begin{aligned}&\mathbb{P}\big(\big|\nabla_i \mathcal{L}(\boldsymbol{\beta}^*)\big| > \sigma \cdot b \cdot t \cdot \sqrt{D/n} \big| \mathbf{x}_1,\ldots,\mathbf{x}_n\big) \\ &\leq 2\exp(-t^2/2), \ \ \forall t > 0. \qquad \text{(B.1)}\end{aligned}$$

Taking a union bound over $j = 1,2,\ldots,d$ in for the left-hand side of (B.1) we obtain that

$$\begin{aligned}&\mathbb{P}\big(\big\|\nabla L(\boldsymbol{\beta}^*)\big\|_\infty > \sigma \cdot b \cdot t\sqrt{D/n} \big| \mathbf{x}_1,\ldots,\mathbf{x}_n\big) \\ &\leq 2\exp(-t^2/2 + \log d), \ \ \forall t > 0. \qquad \text{(B.2)}\end{aligned}$$

By choosing $t = C\sqrt{\log d}$ in (B.2) for a sufficiently large C, we conclude that there exist a constant $B = C \cdot b \cdot \sqrt{D} > 0$ such that $\|\nabla L(\boldsymbol{\beta}^*)\|_\infty \leq B\sigma\sqrt{\log d/n}$ with probability at least $1 - \delta$, where we have $\delta \leq (2d)^{-1}$. □

Proof of Lemma 4. By the definition of $L(\boldsymbol{\beta})$, the gradient $\nabla L(\boldsymbol{\beta})$ is given by

$$\nabla L(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i^\top\boldsymbol{\beta})]f'(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i. \quad \text{(B.3)}$$

Hence for $\nabla L(\boldsymbol{\beta}^*)$, (B.3) can be reduced to

$$\nabla L(\boldsymbol{\beta}^*) = -\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*)\mathbf{x}_i, \quad \text{(B.4)}$$

where $\epsilon_1, \ldots, \epsilon_n$ are n i.i.d. realizations of the random noise ϵ in (1.1). For any $\boldsymbol{\beta} \in \mathbb{R}^d$, we denote $\boldsymbol{\eta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$. Recalling that $y_i = f(\mathbf{x}_i^\top\boldsymbol{\beta}^*) + \epsilon_i$, Taylor expansion of (B.3) implies that

$$\begin{aligned}\nabla L(\boldsymbol{\beta}) = &-\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f'(\mathbf{x}_i^\top\boldsymbol{\beta})\mathbf{x}_i \\ &+ \frac{1}{n}\sum_{i=1}^{n} f'(\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}})f'(\mathbf{x}_i^\top\boldsymbol{\beta})(\mathbf{x}_i^\top\boldsymbol{\eta})\mathbf{x}_i, \quad \text{(B.5)}\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ lies on the line segment between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$. Combining (B.4) and (B.5) we have

$$\langle \nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle = A_1 + A_2, \quad \text{(B.6)}$$

where A_1 and A_2 are defined respectively as

$$\begin{aligned}A_1 &= \frac{1}{n}\sum_{i=1}^{n} f'(\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}})f'(\mathbf{x}_i^\top\boldsymbol{\beta})(\mathbf{x}_i^\top\boldsymbol{\eta})^2 \ \text{ and} \\ A_2 &= \frac{1}{n}\sum_{i=1}^{n}\big\{f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*) - f'(\mathbf{x}_i^\top\boldsymbol{\beta})\big\}(\mathbf{x}_i^\top\boldsymbol{\eta})\epsilon_i.\end{aligned}$$

By the boundedness of f', we can lower bound A_1 by

$$A_1 \geq a^2\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\boldsymbol{\eta})^2 = a^2\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}. \quad \text{(B.7)}$$

For the second part A_2, by the sub-Gaussianity of the random noise ϵ_i's,

$$\big\{f'(\mathbf{x}_i^\top\boldsymbol{\beta}^*) - f'(\mathbf{x}_i^\top\boldsymbol{\beta})\big\}\cdot(\mathbf{x}_i^\top\boldsymbol{\eta})\cdot\epsilon_i$$

is a centered sub-Gaussian random variable with variance proxy

$$\sigma^2[f'\big(\mathbf{x}_i^\top\boldsymbol{\beta}^*\big) - f'\big(\mathbf{x}_i^\top\boldsymbol{\beta}\big)]^2\cdot(\mathbf{x}_i^\top\boldsymbol{\eta})^2 \leq 4\sigma^2b^2(\mathbf{x}_i^\top\boldsymbol{\eta})^2.$$

Therefore we conclude that A_2 is centered and sub-Gaussian with variance proxy bounded by

$$4b^2n^{-2}\sigma^2\sum_{i=1}^{n}(\mathbf{x}_i^\top\boldsymbol{\eta})^2 = 4b^2\sigma^2n^{-1}\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}.$$

By the tail bound for sub-Gaussian random variables, we obtain that for any $x > 0$,

$$\mathbb{P}\big(|A_2| \geq x\big) \leq 2\exp(-x^2/C),$$

where $C = 8b^2\sigma^2n^{-1}\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}$. With probability at least $1 - (2d)^{-1}$, it holds that

$$\begin{aligned}A_2 &\geq \sqrt{C\cdot\log(4d)} \geq -3b\sigma\sqrt{\log d/n}\sqrt{\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}} \\ &\geq -3b\sigma\sqrt{D\log d/n}\|\boldsymbol{\eta}\|_1, \quad \text{(B.8)}\end{aligned}$$

where the last inequality is derived from Hölder's inequality $\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} \leq \|\widehat{\boldsymbol{\Sigma}}\|_\infty\|\boldsymbol{\eta}\|_1^2 \leq D\|\boldsymbol{\eta}\|_1^2$. Therefore combining (B.6), (B.7) and (B.8) with probability at least $1 - (2d)^{-1}$, we have

$$\begin{aligned}&\langle \nabla L(\boldsymbol{\beta}) - \nabla L(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle \\ &\geq a^2\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta} - 3b\sigma\sqrt{D\log d/n}\|\boldsymbol{\eta}\|_1.\end{aligned}$$

This concludes the proof of Lemma 4. □

Proof of Lemma 5. Recall that $\mathcal{J}$ is the set of indices of the largest k^* entries of $\boldsymbol{\eta}_{\mathcal{S}^c}$ in absolute value and let $\mathcal{I} = \mathcal{J}\cup\mathcal{S}$. The following Lemma establishes a lower-bound on $\boldsymbol{\eta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\eta}$.

Lemma 7. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be a positive semi-definite matrix and $\rho_-(k)$ and $\rho_+(k)$ be its k-sparse eigenvalues. Suppose that for some integer s and k, $\rho_-(s+2k) > 0$. For any $\mathbf{v} \in \mathbb{R}^d$, let $\mathcal{F}$ be any index set of size $d - s$, that is, $|\mathcal{F}^c| = s$. We let $\mathcal{J}$ be the set of indices of the k largest component of $\mathbf{v}_{\mathcal{F}^c}$ in absolute value and let $\mathcal{I} = \mathcal{F}^c \cup \mathcal{J}$. Then we have

$$\begin{aligned}\mathbf{v}^\top\boldsymbol{\Sigma}\mathbf{v} \geq &\rho_-(s+k)\cdot\Big[\|\mathbf{v}_{\mathcal{I}}\|_2 - \\ &\sqrt{\rho_+(k)/\rho_-(s+2k) - 1}\cdot\|\mathbf{v}_{\mathcal{F}}\|_1/\sqrt{k}\Big]\cdot\|\mathbf{v}_{\mathcal{I}}\|_2.\end{aligned}$$

By Assumption 1, $\rho_-(s^* + 2k^*) > 0$. Combining Lemma 7 with $\mathcal{F} = \mathcal{S}^c$ and that $\|\boldsymbol{\eta}_{\mathcal{S}^c}\|_1 \leq \gamma\|\boldsymbol{\eta}_{\mathcal{S}}\|_1 \leq \gamma\sqrt{s}\|\boldsymbol{\eta}_{\mathcal{S}}\|_2$ together yield inequality (6.5).

For the second part of the lemma, by the definition of $\mathcal{J}$ we obtain that

$$\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_\infty \leq \|\boldsymbol{\eta}_{\mathcal{J}}\|_1/k^* \leq \|\boldsymbol{\eta}_{\mathcal{S}^c}\|_1/k^* \leq \gamma/k^*\|\boldsymbol{\eta}_{\mathcal{S}}\|_1,$$

hence by Hölder's inequality we have

$$\begin{aligned}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_2 &\leq \|\boldsymbol{\eta}_{\mathcal{I}^c}\|_1^{1/2}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_\infty^{1/2} \\ &\leq (\gamma/k^*)^{1/2}\|\boldsymbol{\eta}_{\mathcal{S}}\|_1^{1/2}\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_1^{1/2} \\ &\leq \gamma k^{*-1/2}\cdot\|\boldsymbol{\eta}_{\mathcal{S}}\|_1,\end{aligned}$$

where we use the fact that $\mathcal{I}^c \subset \mathcal{S}^c$. Thus it holds that

$$\|\boldsymbol{\eta}_{\mathcal{I}^c}\|_2 \leq \gamma\sqrt{s^*/k^*}\cdot\|\boldsymbol{\eta}_{\mathcal{S}}\|_2 \leq \gamma\cdot\|\boldsymbol{\eta}_{\mathcal{I}}\|_2 \ \text{ and} \quad \text{(B.9)}$$

$$\big\|\boldsymbol{\eta}\big\|_2 \leq (1+\gamma)\cdot\|\boldsymbol{\eta}_{\mathcal{I}}\|_2. \quad \text{(B.10)}$$

Thus we conclude the proof of Lemma 5. □

Proof of Lemma 7. Without loss of generality, we assume that $\mathcal{F}^c = \{1, \ldots, s\}$. We also assume that for $\mathbf{v} \in \mathbb{R}^d$,when $j > s$, v_j is arranged in descending order of $|v_j|$. That is, we rearrange the components of $\mathbf{v}$ such that $|v_j| \geq |v_{j+1}|$ for all j greater than s. Let $\mathcal{J}_0 = \{1, \ldots, s\}$ and $\mathcal{J}_i = \{s + (i-1)k + 1, \ldots, \min(s + ik, d)\}$ for $i \geq 1$. By definition, we have $\mathcal{J} = \mathcal{J}_1$ and $\mathcal{I} = \mathcal{J}_0 \cup J_1$. Moreover, we have $\|\mathbf{v}_{\mathcal{J}_i}\|_\infty \leq \|\mathbf{v}_{\mathcal{J}_{i-1}}\|_1/k$ when $i \geq 2$ because of the descending order of $|v_j|$ for $j > s$. Then we further have $\sum_{i \geq 2} \|\mathbf{v}_{\mathcal{J}_i}\|_\infty \leq \|\mathbf{v}_{\mathcal{F}}\|_1/k$.

We define the restricted correlation coefficients of $\boldsymbol{\Sigma}$ as

$$\pi(s,k) \coloneqq \sup\Bigg\{ \frac{\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{J}} \|\mathbf{v}_{\mathcal{I}}\|_2}{\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}} \|\mathbf{v}_{\mathcal{J}}\|_\infty} : I \cap \mathcal{J} = \emptyset, \\ |\mathcal{I}| \leq s, |\mathcal{J}| \leq k, \mathbf{v} \in \mathbb{R}^d \Bigg\}.$$

As shown in [56], if $\rho_-(s+k) > 0$ we have

$$\pi(s,k) \leq \frac{\sqrt{k}}{2} \cdot \sqrt{\rho_+(k)/\rho_-(s+k) - 1}. \tag{B.11}$$

Then by the definition of $\pi(s+k,k)$ we obtain

$$\big|\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{J}_i}\big| \leq \pi(s+k,k) \cdot \big(\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \cdot \|\mathbf{v}_{\mathcal{J}_i}\|_\infty / \|\mathbf{v}_{\mathcal{I}}\|_2.$$

Thus we have the following upper bound for $\big|\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}^c}\big|$:

$$\begin{aligned} \big|\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}^c}\big| &\leq \sum_{i \geq 2} \big|\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{J}_i}\big| \\ &\leq \pi(s+k,k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \big(\mathbf{v}_I^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \sum_{i \geq 2} \|\mathbf{v}_{\mathcal{J}_i}\|_\infty \\ &\leq \pi(s+k,k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \big(\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \|\mathbf{v}_{\mathcal{F}}\|_1/k. \end{aligned} \tag{B.12}$$

Because $\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} \geq \mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}} + 2\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}^c}$, by (B.12) we have

$$\begin{aligned} &\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} \\ &\quad \geq \mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}} - 2\pi(s+k,k) \|\mathbf{v}_I\|_2^{-1} \big(\mathbf{v}_I^\top \boldsymbol{\Sigma} \mathbf{v}_I\big) \|\mathbf{v}_{\mathcal{F}}\|_1/k \\ &\quad = \big(\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}}\big) \big[1 - 2\pi(s+k,k) \|\mathbf{v}_{\mathcal{I}}\|_2^{-1} \|\mathbf{v}_{\mathcal{F}}\|_1/k\big]. \end{aligned} \tag{B.13}$$

Combining (B.13), the fact that $\mathbf{v}_{\mathcal{I}}^\top \boldsymbol{\Sigma} \mathbf{v}_{\mathcal{I}} \geq \rho_-(s+k) \cdot \|\mathbf{v}_{\mathcal{I}}\|_2^2$ and (B.11) for $\pi(s+k,k)$, we conclude the proof of Lemma 7. □