Appendices

A Usefull properties of sub-Gaussian random variables

This section presents useful preliminary results satisfied by sub-Gaussian random variables. In particular, Lemma 5 provides a probabilistic upper-bound satisfied by the maximum of independent sub-Gaussian random variables.

A.1 Preliminary results

Under Assumption 3, the random variables $\sum_{i=1}^{n} \partial f \left(\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle, y_i \right) x_{ij}, \ \forall j$ are sub-Gaussian. They consequently satisfy the next Lemma 3:

Lemma 3 *Let $Z \sim$ subG(σ^2) for a fixed $\sigma > 0$. Then for any $t > 0$ it holds*

$$\mathbb{E} \left(\exp(tZ) \right) \leq e^{4\sigma^2 t^2}.$$

In addition, for any positive integer $\ell \geq 1$ we have:

$$\mathbb{E} \left(|Z|^\ell \right) \leq (2\sigma^2)^{\ell/2} \ell \Gamma(\ell/2)$$

where Γ is the Gamma function defined as $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx, \ \forall t > 0$.
Finally, let $Y = Z^2 - \mathbb{E}(Z^2)$ then we have

$$\mathbb{E} \left(\exp \left(\frac{1}{16\sigma^2} Y \right) \right) \leq \frac{3}{2}, \tag{19}$$

and as a consequence $\mathbb{E} \left(\exp \left(\frac{1}{16\sigma^2} Z^2 \right) \right) \leq 2$.

Proof: The two first results correspond to Lemmas 1.4 and 1.5 from Rigollet [2015].
In particular $\mathbb{E} \left(|Z|^2 \right) \leq 4\sigma^2$.
In addition, using the proof of Lemma 1.12 we have:

$$\mathbb{E} \left(\exp(tY) \right) \leq 1 + 128 t^2 \sigma^4, \ \forall |t| \leq \frac{1}{16\sigma^2}.$$

Equation (19) holds in the particular case where $t = 1/16\sigma^2$.
The last part of the lemma combines our precedent results with the observation that $\frac{3}{2} e^{1/4} \leq 2$. $\square$

A.2 Proof of Lemma 1

As a first consequence of Lemma 3, we derive the proof of Lemma 1 – stated in Section 2.3.

Proof: We note $S_i = \partial f\left(\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle, y_i\right), \ \forall i$.
Since $\boldsymbol{\beta}^*$ minimizes the theoretical loss, we have $\mathbb{E}(S_i x_{ij}) = 0, \ \forall i, j$.
By definition of a sub-Gaussian random variable, we fix $M > 0$ such that: $\forall t > 0$,

$$\mathbb{P}\left(|S_i x_{i,j}| > t\right) \leq 2 \exp\left(-\frac{t^2}{2L^2 M^2}\right), \ \forall i, j.$$

Then from Lemma 3 it holds:

$$\mathbb{E}\left(\exp(t S_i x_{ij})\right) \leq e^{4L^2 M^2 t^2}, \ \forall t > 0, \forall i, j.$$

As a consequence, using Lemma 3 for the independent random variables $(S_1 x_{1,j}, \ldots, S_n x_{n,j})$, it holds $\forall t > 0$,

$$\mathbb{E}\left(\exp\left(t \sum_{i=1}^{n} S_i x_{i,j}\right)\right) = \prod_{i=1}^{n} \mathbb{E}\left(\exp\left(t S_i x_{ij}\right)\right) \leq \prod_{i=1}^{n} e^{4L^2 M^2 t^2} = e^{4nL^2 M^2 t^2}.$$

Let $M_1 = 2\sqrt{2} M \sqrt{n}$, then with a Chernoff bound:

$$\mathbb{P}\left(\sum_{i=1}^{n} S_i x_{i,j} > t\right) \leq \min_{s>0} \ \exp\left(\frac{M_1^2 L^2 s^2}{2} - st\right) = \exp\left(-\frac{t^2}{2L^2 M_1^2}\right), \ \forall t > 0,$$

which concludes the proof. □

A.3 A bound for the maximum of independent sub-Gaussian variables

The next two technical lemmas derive a probabilistic upper-bound for the maximum of sub-Gaussian random variables. Lemma 4 extends Proposition E.1 [Bellec et al., 2016] to sub-Gaussian random variables.

Lemma 4 *Let $g_1, \ldots g_p$ be independent sub-Gaussian random variables with variance σ^2. Denote by $(g_{(1)}, \ldots, g_{(p)})$ a non-increasing rearrangement of $(|g_1|, \ldots, |g_p|)$. Then $\forall t > 0$ and $\forall j \in \{1, \ldots, p\}$:*

$$\mathbb{P}\left(\frac{1}{j\sigma^2} \sum_{k=1}^{j} g_{(k)}^2 > t \log\left(\frac{2p}{j}\right)\right) \leq \left(\frac{2p}{j}\right)^{1-\frac{t}{16}}.$$

Proof: Let $j \in \{1, \ldots, p\}$. We first apply a Chernoff bound:

$$\mathbb{P}\left(\frac{1}{j\sigma^2} \sum_{k=1}^{j} g_{(k)}^2 > t \log\left(\frac{2p}{j}\right)\right) \leq \mathbb{E}\left(\exp\left(\frac{1}{16 j\sigma^2} \sum_{k=1}^{j} g_{(k)}^2\right)\right) \left(\frac{2p}{j}\right)^{-\frac{t}{16}}.$$

Then we use Jensen inequality to obtain

$$\begin{aligned}
\mathbb{E}\left(\exp\left(\frac{1}{16 j\sigma^2} \sum_{k=1}^{j} g_{(k)}^2\right)\right) &\leq \frac{1}{j} \sum_{k=1}^{j} \mathbb{E}\left(\exp\left(\frac{1}{16\sigma^2} g_{(k)}^2\right)\right) \\
&\leq \frac{1}{j} \sum_{k=1}^{p} \mathbb{E}\left(\exp\left(\frac{1}{16\sigma^2} g_k^2\right)\right) \leq \frac{2p}{j} \text{ with Lemma 3.}
\end{aligned}$$

□

Using Lemma 4, we can derive the following bound which holds with high probability.

Lemma 5 *We consider the assumptions and notations of Lemma 4. In addition, we define the coefficients $\lambda_j = \sqrt{\log(2p/j)}, \ j = 1, \ldots p$. Then for $\delta \in \left(0, \frac{1}{2}\right)$, it holds with probability at least $1 - \delta$:*

$$\sup_{j=1,\ldots,p} \left\{\frac{g_{(j)}}{\sigma \lambda_j}\right\} \leq 12 \sqrt{\log(1/\delta)}.$$

Proof: We fix $\delta \in \left(0, \frac{1}{2}\right)$ and $j \in \{1, \ldots, p\}$. We upper-bound $g_{(j)}^2$ by the average of all larger variables:

$$g_{(j)}^2 \leq \frac{1}{j} \sum_{k=1}^{j} g_{(k)}^2.$$

Applying Lemma 4 gives, for $t > 0$:

$$\mathbb{P}\left(\frac{g_{(j)}^2}{\sigma^2 \lambda_j^2} > t\right) \leq \mathbb{P}\left(\frac{1}{j\sigma^2} \sum_{k=1}^{j} g_{(k)}^2 > t\lambda_j^2\right) \leq \left(\frac{j}{2p}\right)^{\frac{t}{16}-1}.$$

We fix $t = 144 \log(1/\delta)$ and use an union bound to get:

$$\mathbb{P}\left(\sup_{j=1,\ldots,p} \frac{g_{(j)}}{\sigma\lambda_j} > 12\sqrt{\log(1/\delta)}\right) \leq \left(\frac{1}{2p}\right)^{9\log(1/\delta)-1} \sum_{j=1}^{p} j^{9\log(1/\delta)-1}.$$

Since $\delta < \frac{1}{2}$ it holds that $9\log(1/\delta) - 1 \geq 9\log(2) - 1 > 0$, then the map $t > 0 \mapsto t^{9\log(1/\delta)-1}$ is increasing. An integral comparison gives:

$$\sum_{j=1}^{p} j^{9\log(1/\delta)-1} \leq \frac{1}{2}(p+1)^{9\log(1/\delta)} = \frac{1}{2}\delta^{-9\log(p+1)}.$$

In addition $9\log(1/\delta) - 1 \geq 7\log(1/\delta) = -7\log(\delta)$ and

$$\left(\frac{1}{2p}\right)^{9\log(1/\delta)-1} \leq \left(\frac{1}{2p}\right)^{-7\log(\delta)} = \delta^{7\log(2p)}.$$

Finally, by assuming $p \geq 2$, then we have $7\log(2p) - 9\log(p+1) > 1$, thus:

$$\mathbb{P}\left(\sup_{j=1,\ldots,p} \frac{g_{(j)}}{\sigma\lambda_j} > 12\sqrt{\log(1/\delta)}\right) \leq \delta,$$

which concludes the proof. □

B Proof of Theorem 2

We use the minimality of $\hat{\boldsymbol{\beta}}$ and Lemma 4 to derive the cone condition.

Proof: We assume without loss of generality that $|h_1| \geq \ldots \geq |h_p|$. We define $S_0 = \{1, \ldots, k^*\}$ as the set of the k^* highest coefficients of $\boldsymbol{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$.

$\hat{\boldsymbol{\beta}}$ is the solution of Problem (2) hence:

$$\frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \hat{\boldsymbol{\beta}}\rangle; y_i\right) + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^*\rangle; y_i\right) + \lambda\|\boldsymbol{\beta}^*\|_1. \tag{20}$$

Using the definition of $\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right)$ as introduced in Theorem 3, Equation (20) can be written in a compact form as:

$$\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right) \leq \lambda\|\boldsymbol{\beta}^*\|_1 - \lambda\|\hat{\boldsymbol{\beta}}\|_1.$$

Introducing the support S^* of $\boldsymbol{\beta}^*$ we have

$$\begin{aligned}\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right) &\leq \lambda \|\boldsymbol{\beta}^*_{S^*}\|_1 - \lambda \|\hat{\boldsymbol{\beta}}_{S^*}\|_1 - \lambda \|\hat{\boldsymbol{\beta}}_{(S^*)^c}\|_1 \\ &\leq \lambda \|\boldsymbol{h}_{S^*}\|_1 - \lambda \|\boldsymbol{h}_{(S^*)^c}\|_1 \\ &\leq \lambda \|\boldsymbol{h}_{S_0}\|_1 - \lambda \|\boldsymbol{h}_{(S_0)^c}\|_1,\end{aligned} \tag{21}$$

where this last relation holds by definition of S_0. We now want to lower bound $\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right)$. Exploiting the existence of a bounded sub-Gradient ∂f we obtain

$$\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right) \geq S\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right) := \frac{1}{n}\sum_{i=1}^{n} \partial f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* \rangle; y_i\right) \langle \boldsymbol{x_i}, \boldsymbol{h} \rangle.$$

In addition we have:

$$\begin{aligned}\left|S\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} \partial f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* \rangle; y_i\right) x_{ij} h_j\right| \\ &\leq \frac{1}{\sqrt{n}}\sum_{j=1}^{p}\left(\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n} \partial f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* \rangle; y_i\right) x_{ij}\right|\right)|h_j|.\end{aligned}$$

Let us define the independent random variables $g_j = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \partial f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* \rangle; y_i\right) x_{ij},\ j = 1, \ldots, p$. Assumption 3 guarantees that $g_1, \ldots, g_p$ are sub-Gaussian with variance L^2M^2. A first upper-bound of the quantity $|S(\boldsymbol{h})|$ could be obtained by considering the maximum of the sequence $\{g_j\}$. However Lemma 5 gives us a stronger result.

Indeed, since $\delta \leq 1$ we introduce a non-increasing rearrangement $(g_{(1)}, \ldots, g_{(p)})$ of $(|g_1|, \ldots, |g_p|)$. We recall that $S_0 = \{1, \ldots, k^*\}$ denotes the subset of indexes of the k^* highest elements of $\boldsymbol{h}$ and we use Lemma 5 to get, with probability at least $1 - \frac{\delta}{2}$:

$$\begin{aligned}\left|S\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right)\right| &\leq \frac{1}{\sqrt{n}}\sum_{j=1}^{p} g_j|h_j| = \frac{1}{\sqrt{n}}\sum_{j=1}^{p} g_{(j)}|h_{(j)}| = \frac{1}{\sqrt{n}}\sum_{j=1}^{p} \frac{g_{(j)}}{LM\lambda_j} LM\lambda_j |h_{(j)}| \\ &\leq \frac{1}{\sqrt{n}} \sup_{j=1,\ldots,p}\left\{\frac{g_{(j)}}{LM\lambda_j}\right\}\sum_{j=1}^{p} LM\lambda_j|h_{(j)}| \\ &\leq 12LM\sqrt{\frac{\log(2/\delta)}{n}}\sum_{j=1}^{p}\lambda_j|h_{(j)}| \text{ with Lemma 5} \\ &\leq 12LM\sqrt{\frac{\log(2/\delta)}{n}}\sum_{j=1}^{p}\lambda_j|h_j| \text{ since } \lambda_1 \geq \ldots \geq \lambda_p \text{ and } |h_1| \geq \ldots \geq |h_p| \\ &\leq 12LM\sqrt{\frac{\log(2/\delta)}{n}}\left(\sum_{j=1}^{k^*}\lambda_j|h_j| + \lambda_{k^*}\sum_{j=k^*}^{p}|h_j|\right) \\ &= 12LM\sqrt{\frac{\log(2/\delta)}{n}}\left(\sum_{j=1}^{k^*}\lambda_j|h_j| + \lambda_{k^*}\|\boldsymbol{h}_{(S_0)^c}\|_1\right).\end{aligned} \tag{22}$$

Cauchy-Schwartz inequality leads to:

$$\sum_{j=1}^{k^*} \lambda_j |h_j| \leq \sqrt{\sum_{j=1}^{k^*} \lambda_j^2} \|\boldsymbol{h}_{S_0}\|_2 \leq \sqrt{k^* \log(2pe/k^*)} \|\boldsymbol{h}_{S_0}\|_2,$$

where we have used the Stirling formula to get $\left(\frac{n}{e}\right)^n \leq n!$ and we have used:

$$\sum_{j=1}^{k^*} \lambda_j^2 = \sum_{j=1}^{k^*} \log(2p/j) = k^* \log(2p) - \log(k^*!)$$
$$\leq k^* \log(2p) - k^* \log(k^*/e) = k^* \log(2pe/k^*).$$

In the statement of Theorem 2 we have defined $\lambda = 12\alpha LM \sqrt{n^{-1} \log(2pe/k^*) \log(2/\delta)}$.
Because $\lambda_{k^*} \leq \sqrt{\log(2pe/k^*)}$, Equation (22) leads to:

$$|S(\boldsymbol{\beta}^*, \boldsymbol{h})| \leq \frac{1}{\alpha} \lambda \left(\sqrt{k^*} \|\boldsymbol{h}_{S_0}\|_2 + \|\boldsymbol{h}_{(S_0)^c}\|_1 \right)$$

Combined with Equation (21), it holds with probability at least $1 - \frac{\delta}{2}$:

$$-\frac{\lambda}{\alpha} \left(\sqrt{k^*} \|\boldsymbol{h}_{S_0}\|_2 + \|\boldsymbol{h}_{(S_0)^c}\|_1 \right) \leq \lambda \|\boldsymbol{h}_{S_0}\|_1 - \lambda \|\boldsymbol{h}_{(S_0)^c}\|_1,$$

which immediately leads to:

$$\|\boldsymbol{h}_{(S_0)^c}\|_1 \leq \frac{\alpha}{\alpha - 1} \|\boldsymbol{h}_{S_0}\|_1 + \frac{\sqrt{k^*}}{\alpha - 1} \|\boldsymbol{h}_{S_0}\|_2.$$

We conclude that $\boldsymbol{h} \in \Lambda\left(S_0,\ \frac{\alpha}{\alpha-1},\ \frac{\sqrt{k^*}}{\alpha-1}\right)$ with probability at least $1 - \frac{\delta}{2}$. □

C Proof of Theorem 3:

Proof: Let $k \in \{1, \ldots, p\}$ and $S_1, \ldots S_q$ be a partition of $\{1, \ldots, p\}$ such that $q = \lceil p/k \rceil$ and $|S_\ell| \leq k, \forall \ell$. We divide the proof of the theorem in 3 steps. We first upper-bound the inner supremum for a sequence of k sparse vectors $\boldsymbol{z}_{S_1}, \ldots, \boldsymbol{z}_{S_q}$ satisfying $\|\boldsymbol{z}_{S_\ell}\|_1 \leq 3R, \forall \ell$. We then extend this bound for the supremum over the compact set of sequences considered through an ϵ-net argument.

Step 1: Let us fix a sequence $\boldsymbol{z}_{S_1}, \ldots, \boldsymbol{z}_{S_q} \in \mathbb{R}^p :\ \mathrm{Supp}(\boldsymbol{z}_{S_j}) \subset S_j, \forall j$ and $\|\boldsymbol{z}_{S_\ell}\|_1 \leq 3R, \forall \ell$.
In particular, $\|\boldsymbol{z}_{S_j}\|_0 \leq k, \forall j$. In the rest of the proof, we define $\boldsymbol{z}_{S_0} = \mathbf{0}$ and:

$$\boldsymbol{w}_\ell = \boldsymbol{\beta}^* + \sum_{j=1}^{\ell} \boldsymbol{z}_{S_j}, \quad \forall \ell \in \{1, \ldots, q\}. \tag{23}$$

In addition, we introduce $Z_{i\ell},\ \forall i \in \{1, \ldots, n\},\ \forall \ell \in \{1, \ldots, q\}$ as follows:

$$Z_{i\ell} = f(\langle \boldsymbol{x}_i, \boldsymbol{w}_\ell \rangle; y_i) - f(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1} \rangle; y_i) = f(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1} + \boldsymbol{z}_{S_\ell} \rangle; y_i) - f(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1} \rangle; y_i).$$

We fix $\ell \in \{1, \dots, q\}$. Let us note that:

$$
\begin{aligned}
\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) &= \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1} + \boldsymbol{z}_{S_\ell}\rangle; y_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1}\rangle; y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left\{ f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1} + \boldsymbol{z}_{S_\ell}\rangle; y_i\right) - f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell-1}\rangle; y_i\right) \right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_{i\ell}.
\end{aligned} \tag{24}
$$

Assumption 1 guarantees that $f(., y)$ is L-Lipschitz $\forall y$ then:

$$|Z_{i\ell}| \leq L \left|\langle \boldsymbol{x}_i, \boldsymbol{z}_{S_\ell}\rangle\right|.$$

Then using Assumption 4.1(k) on the k sparse vector $\boldsymbol{z}_{S_\ell}$ it holds:

$$\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n} |Z_{i\ell}| \leq \frac{1}{n}\sum_{i=1}^{n} L\left|\langle \boldsymbol{x}_i, \boldsymbol{z}_{S_\ell}\rangle\right| = \frac{L}{n}\|\mathbb{X}\boldsymbol{z}_{S_\ell}\|_1 \leq \frac{L\mu(k)}{\sqrt{nk}}\|\boldsymbol{z}_{S_\ell}\|_1.$$

Hence, with Hoeffding's lemma, the centered bounded random variable $\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)$ is sub-Gaussian with variance $\frac{L^2\mu(k)^2}{nk}\|\boldsymbol{z}_{S_\ell}\|_1^2$. It then hold, $\forall t > 0$,

$$\mathbb{P}\left(\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| \geq t\|\boldsymbol{z}_{S_\ell}\|_1\right) \leq 2\exp\left(-\frac{knt^2}{2L^2\mu(k)^2}\right). \tag{25}$$

Equation (25) holds for all values of ℓ. Thus, an union bound immediately gives:

$$\mathbb{P}\left(\sup_{\ell=1,\dots,q}\left\{\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| - t\|\boldsymbol{z}_{S_\ell}\|_1\right\} \geq 0\right) \leq 2\left\lceil\frac{p}{k}\right\rceil\exp\left(-\frac{knt^2}{2L^2\mu(k)^2}\right). \tag{26}$$

Step 2: We extend the result to any sequence of vectors $\boldsymbol{z}_{S_1}, \dots, \boldsymbol{z}_{S_q} \in \mathbb{R}^p :\ \text{Supp}(\boldsymbol{z}_{S_\ell}) \subset S_\ell$ and $\|\boldsymbol{z}_{S_\ell}\|_1 \leq 3R, \forall \ell$ throught an ϵ-net argument.

We recall that an ϵ-net of a set $\mathcal{I}$ is a subset $\mathcal{N}$ of $\mathcal{I}$ such that each element of I is at a distance at most ϵ of $\mathcal{N}$. We know from Lemma 1.18 from Rigollet [2015], that for any $\epsilon \in (0, 1)$, the ball $\left\{\boldsymbol{z} \in \mathbb{R}^d :\ \|\boldsymbol{z}\|_1 \leq R\right\}$ has an ϵ-net of cardinality $|\mathcal{N}| \leq \left(\frac{2R+1}{\epsilon}\right)^d$ – the ϵ-net is defined in term for the L1 norm. In addition, by following the proof of the lemma, we can create this set such that it contains $\mathbf{0}$.

Consequently, we use Equation (26) on a product of ϵ-nets $\mathcal{N}_{k,R} = \prod_{\ell=1}^{q} \mathcal{N}_{k,R}^{\ell}$. Each $\mathcal{N}_{k,R}^{\ell}$ is an ϵ-net of the bounded set of k sparse vectors $\mathcal{I}_{k,R}^{\ell} = \{\boldsymbol{z}_{S_\ell} \in \mathbb{R}^p :\ \text{Supp}(\boldsymbol{z}_{S_\ell}) \subset S_\ell\ ;\ \|\boldsymbol{z}_{S_\ell}\|_1 \leq 3R\}$ which contains $\mathbf{0}_{S_\ell}$. We note $\mathcal{I}_{k,R} = \prod_{\ell=1}^{q} \mathcal{I}_{k,R}^{\ell}$. It then holds:

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{(\boldsymbol{z}_{S_1},\dots,\boldsymbol{z}_{S_q}) \in \mathcal{N}_{k,R}}\left\{\sup_{\ell=1,\dots,q}\left\{\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| - t\|\boldsymbol{z}_{S_\ell}\|_1\right\} \geq 0\right\}\right) \\
&\leq 2\left\lceil\frac{p}{k}\right\rceil\left(\frac{6R+1}{\epsilon}\right)^k\left\lceil\frac{p}{k}\right\rceil\exp\left(-\frac{knt^2}{2L^2\mu(k)^2}\right) \leq 2\left(\frac{2p}{k}\right)^2\left(\frac{6R+1}{\epsilon}\right)^k\exp\left(-\frac{knt^2}{2L^2\mu(k)^2}\right).
\end{aligned} \tag{27}
$$

Step 3: We now extend Equation (27) to control any vector in $\mathcal{I}_{k,R}$. For $\boldsymbol{z}_{S_1}, \ldots, \boldsymbol{z}_{S_q} \in \mathcal{I}_{k,R}$, there exists $\tilde{\boldsymbol{z}}_{S_1}, \ldots, \tilde{\boldsymbol{z}}_{S_q} \in \mathcal{N}_{k,R}$ such that $\|\boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell}\|_1 \leq \epsilon, \forall \ell$. Similarly to Equation (23), we define:

$$\tilde{\boldsymbol{w}}_\ell = \boldsymbol{\beta}^* + \sum_{j=1}^{\ell} \tilde{\boldsymbol{z}}_{S_j}, \;\; \forall \ell \in \{1, \ldots, q\}.$$

For a given t, let us define

$$f_t\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) = \left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right| - t\|\boldsymbol{z}_{S_\ell}\|_1, \forall \ell.$$

We fix $\ell_0(t)$ such that $\ell_0(t) \in \underset{\ell=1,\ldots,q}{\operatorname{argmax}} \{f_{7t}\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\}$. The choice of $7t$ will be justified later. We fix t and will just note $\ell_0 = \ell_0(t)$ when no confusion can be made.

With Assumption 1 we obtain:

$$\begin{aligned}
&\left|\Delta\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - \Delta\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right)\right| \\
&= \frac{1}{n}\left|\sum_{i=1}^n f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell_0}\rangle; y_i\right) - \sum_{i=1}^n f\left(\langle \boldsymbol{x}_i, \tilde{\boldsymbol{w}}_{\ell_0}\rangle; y_i\right) + \sum_{i=1}^n f\left(\langle \boldsymbol{x}_i, \tilde{\boldsymbol{w}}_{\ell_0-1}\rangle; y_i\right) - \sum_{i=1}^n f\left(\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell_0-1}\rangle; y_i\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n L\left|\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell_0} - \tilde{\boldsymbol{w}}_{\ell_0}\rangle\right| + \frac{1}{n}\sum_{i=1}^n L\left|\langle \boldsymbol{x}_i, \boldsymbol{w}_{\ell_0-1} - \tilde{\boldsymbol{w}}_{\ell_0-1}\rangle\right| \\
&= \frac{1}{n}\sum_{i=1}^n L\left|\sum_{\ell=1}^{\ell_0}\langle \boldsymbol{x}_i, \boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell}\rangle\right| + \frac{1}{n}\sum_{i=1}^n L\left|\sum_{\ell=1}^{\ell_0-1}\langle \boldsymbol{x}_i, \boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell}\rangle\right| \\
&\leq \frac{2}{n}\sum_{i=1}^n\sum_{\ell=1}^q L\left|\langle \boldsymbol{x}_i, \boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell}\rangle\right| \\
&= \frac{2}{\sqrt{n}}\sum_{\ell=1}^q \frac{L}{\sqrt{n}}\|\boldsymbol{X}(\boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell})\|_1 \\
&\leq \frac{2}{\sqrt{n}}\sum_{\ell=1}^q \frac{L}{\sqrt{k}}\mu(k)\ \|\boldsymbol{z}_{S_\ell} - \tilde{\boldsymbol{z}}_{S_\ell}\|_1 \\
&\leq \frac{2p}{k\sqrt{kn}}L\mu(k)\epsilon \leq \eta\epsilon.
\end{aligned} \tag{28}$$

where $\eta = \frac{2L\mu(k)}{\sqrt{n}}$ and we have used Assumption 5.1(k). It then holds:

$$\begin{aligned}
f_t\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right) &\geq f_t\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - \left|\Delta\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - \Delta\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right)\right| \\
&\quad - \left|\mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - \Delta\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right)\right)\right| - t\|\boldsymbol{z}_{S_{\ell_0}} - \tilde{\boldsymbol{z}}_{S_{\ell_0}}\|_1 \\
&\geq f_t\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - 2\eta\epsilon - t\epsilon.
\end{aligned}$$

Case 1: Let us assume that $\|\boldsymbol{z}_{S_{\ell_0}}\|_1 \geq \epsilon/2$ and that $t \geq \eta$, then we have:

$$f_t\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right) \geq f_t\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - 2(2\eta + t)\|\boldsymbol{z}_{S_{\ell_0}}\|_1 \geq f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right). \tag{29}$$

Case 2: We now assume $\|\boldsymbol{z}_{S_{\ell_0}}\|_1 \leq \epsilon/2$. Since $\mathbf{0}_{S_{\ell_0}} \in \mathcal{N}_{k,R}$ we derive similarly to Equation (28):

$$\left|\Delta\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) - \Delta\left(\boldsymbol{w}_{\ell_0-1},\ \mathbf{0}_{S_{\ell_0}}\right)\right| \leq \frac{L\mu(k)}{\sqrt{nk}} \left\|\boldsymbol{z}_{S_{\ell_0}}\right\|_1,$$

which then implies that:

$$f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) \leq f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \mathbf{0}_{S_{\ell_0}}\right) + \frac{2L\mu(k)}{\sqrt{nk}} \left\|\boldsymbol{z}_{S_{\ell_0}}\right\|_1 - 7t\left\|\boldsymbol{z}_{S_{\ell_0}}\right\|_1,$$

and this quantity is smaller than $f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \mathbf{0}_{S_{\ell_0}}\right)$ as long as $7t \geq \frac{2L\mu(k)}{\sqrt{nk}}$. The latter condition is satisfied if $t \geq \eta$.

In this case, we can define a new $\tilde{\ell}_0$ for the sequence $\boldsymbol{z}_{S_1}, \ldots, \boldsymbol{z}_{S_{\ell_0-1}}, \mathbf{0}_{S_{\ell_0}}, \boldsymbol{z}_{S_{\ell_0+1}}, \ldots, \boldsymbol{z}_{S_q}$. After a finite number of iteration, by using the result in Equation (29) and the definition of ℓ_0, we finally get that $f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) \leq f_t\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right)$ for some $\tilde{\boldsymbol{z}}_{S_1}, \ldots, \tilde{\boldsymbol{z}}_{S_q} \in \mathcal{N}_{k,R}$.

As a consequence of cases 1 and 2, we obtain: $\forall t \geq \eta,\ \forall \boldsymbol{z}_{S_1}, \ldots, \boldsymbol{z}_{S_q} \in \mathcal{I}_{k,R},\ \exists \tilde{\boldsymbol{z}}_{S_1}, \ldots, \tilde{\boldsymbol{z}}_{S_q} \in \mathcal{N}_{k,R}$:

$$\sup_{\ell=1,\ldots,q} f_{7t}\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) = f_{7t}\left(\boldsymbol{w}_{\ell_0-1},\ \boldsymbol{z}_{S_{\ell_0}}\right) \leq f_t\left(\tilde{\boldsymbol{w}}_{\ell_0-1},\ \tilde{\boldsymbol{z}}_{S_{\ell_0}}\right) \leq \sup_{\ell=1,\ldots,q} f_t\left(\tilde{\boldsymbol{w}}_{\ell-1},\ \tilde{\boldsymbol{z}}_{S_\ell}\right).$$

This last relation is equivalent to saying that $\forall t \geq 7\eta$:

$$\sup_{\boldsymbol{z}_{S_1},\ldots,\boldsymbol{z}_{S_q} \in \mathcal{I}_{k,R}} \left\{\sup_{\ell=1,\ldots,q} f_t\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right\} \leq \sup_{\boldsymbol{z}_{S_1},\ldots,\boldsymbol{z}_{S_q} \in \mathcal{N}_{k,R}} \left\{\sup_{\ell=1,\ldots,q} f_{t/7}\left(\tilde{\boldsymbol{w}}_{\ell-1},\ \tilde{\boldsymbol{z}}_{S_\ell},\right)\right\}. \tag{30}$$

As a consequence, we have $\forall t \geq 7\eta$:

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\boldsymbol{z}_{S_1},\ldots,\boldsymbol{z}_{S_q} \in \mathcal{I}_{k,R}} \left\{\sup_{\ell=1,\ldots,q} \left\{\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| - t\|\boldsymbol{z}_{S_\ell}\|_1\right\}\right\} \geq 0\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{z}_{S_1},\ldots,\boldsymbol{z}_{S_q} \in \mathcal{N}_{k,R}} \left\{\sup_{\ell=1,\ldots,q} \left\{\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| - \frac{t}{7}\|\boldsymbol{z}_{S_\ell}\|_1\right\}\right\} \geq 0\right) \\
&\leq 2\left(\frac{2p}{k}\right)^2 \left(\frac{6R+1}{\epsilon}\right)^k \exp\left(-\frac{kn(t/7)^2}{2L^2\mu(k)^2}\right) \\
&\leq \left(\frac{4p}{k}\right)^2 3^k \exp\left(-\frac{knt^2}{98L^2\mu(k)^2}\right) \text{ by fixing } \epsilon = 2R \text{ and since } R \geq 1.
\end{aligned} \tag{31}$$

Thus we select t such that $t \geq 7\eta$ and that the condition $t^2 \geq \frac{98L^2\mu(k)^2}{kn}\left[k\log(3) + 2\log\left(\frac{4p}{k}\right) + \log\left(\frac{2}{\delta}\right)\right]$ holds [1]. To this end, we define:

$$\tau = 14L\mu(k)\sqrt{\frac{\log(3)}{n} + \frac{\log(4p/k)}{nk} + \frac{\log(2/\delta)}{nk}} \geq 7\eta.$$

We conclude that with probability at least $1 - \frac{\delta}{2}$:

$$\sup_{\boldsymbol{z}_{S_1},\ldots,\boldsymbol{z}_{S_q} \in \mathcal{I}_{k,R}} \left\{\sup_{\ell=1,\ldots,q} \left\{\left|\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right) - \mathbb{E}\left(\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{z}_{S_\ell}\right)\right)\right| - \tau\left(\|\boldsymbol{z}_{S_\ell}\|_1 \vee \eta\right)\right\}\right\} \leq 0.$$

□

[1] A somewhat faster proof would have consisted in fixing $\epsilon = 2R$ in the definition of the ϵ-net – of size now bounded by 3^k – and in noting that because of the L1-constraint, each element $\boldsymbol{z}_{S_\ell}$ is at a distance at most $R = \|\boldsymbol{z}_{S_\ell}\|_1/2$ of its closest neighborhood in the ϵ-net. However, we prefer the more general proof presented herein.

D Proof of Theorem 4:

Proof: The proof is divided in two steps. First, we lower-bound the quantity $\Delta\left(\boldsymbol{\beta}^{*}, \boldsymbol{h}\right)$ with Theorem 3. Second, we refine this lower-bound with the use of the cone condition derived in Theorem 2 and the restricted eigenvalue condition presented in Assumption 4.2.

Step 1: Let us fix the partition of $\{1, \ldots, p\}$: $S_1 = \{1, \ldots, k^*\}, S_2 = \{k^*+1, \ldots, 2k^*\}, \ldots, S_q$ – with $q = \lceil p/k^* \rceil$. Recall that $\boldsymbol{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$. Then it holds $|S_\ell| \leq k^*$ and $\|\boldsymbol{h}_{S_\ell}\|_1 \leq 3R$. We thus can use Theorem 3 for the corresponding sequence $\boldsymbol{h}_{S_1}, \ldots, \boldsymbol{h}_{S_q}$ of k^* sparse vectors.

$$\begin{aligned}
\Delta(\boldsymbol{\beta}^*, \boldsymbol{h}) &= \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* + \boldsymbol{h}\rangle; y_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^*\rangle; y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* + \sum_{j=1}^{q}\boldsymbol{h}_{S_j}\rangle; y_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^*\rangle; y_i\right) \\
&= \sum_{\ell=1}^{q}\left\{\frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* + \sum_{j=1}^{\ell}\boldsymbol{h}_{S_j}\rangle; y_i\right) - \frac{1}{n}\sum_{i=1}^{n} f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* + \sum_{j=0}^{\ell-1}\boldsymbol{h}_{S_j}\rangle; y_i\right)\right\} \qquad (32) \\
&= \sum_{\ell=1}^{q}\Delta\left(\boldsymbol{\beta}^* + \sum_{j=0}^{\ell-1}\boldsymbol{h}_{S_j},\ \boldsymbol{h}_{S_\ell}\right) \\
&= \sum_{\ell=1}^{q}\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{h}_{S_\ell}\right).
\end{aligned}$$

where we have defined $\boldsymbol{w}_\ell = \boldsymbol{\beta}^* + \sum_{j=1}^{\ell}\boldsymbol{h}_{S_j}, \forall \ell$ and $\boldsymbol{h}_{S_0} = \boldsymbol{0}$ as in the proof of Theorem 3. Consequently, with Theorem 3, it holds with probability at least $1 - \frac{\delta}{2}$:

$$\left|\Delta\left(\boldsymbol{w}_{\ell-1}, \boldsymbol{h}_{S_\ell}\right) - \mathbb{E}\left(\boldsymbol{w}_{\ell-1}, \boldsymbol{h}_{S_\ell}\right)\right| \geq \tau\|\boldsymbol{h}_{S_\ell}\|_1, \forall \ell,$$

where $\tau = \tau(k^*) = 14L\mu(k^*)\sqrt{\frac{\log(3)}{n} + \frac{\log(4p/k^*)}{nk^*} + \frac{\log(2/\delta)}{nk^*}}$ is fixed in the rest of the proof.
As a result, following Equation (32), we have with probability at least $1 - \frac{\delta}{2}$:

$$\begin{aligned}
\Delta(\boldsymbol{\beta}^*, \boldsymbol{h}) &\geq \sum_{\ell=1}^{q}\left\{\mathbb{E}\left(\boldsymbol{w}_{\ell-1}, \boldsymbol{h}_{S_\ell}\right) - \tau\|\boldsymbol{h}_{S_\ell}\|_1\right\} \\
&= \mathbb{E}\left(\sum_{\ell=1}^{q}\Delta\left(\boldsymbol{w}_{\ell-1},\ \boldsymbol{h}_{S_\ell}\right)\right) - \sum_{\ell=1}^{q}\tau\|\boldsymbol{h}_{S_\ell}\|_1 \qquad (33) \\
&= \mathbb{E}\left(\Delta(\boldsymbol{\beta}^*, \boldsymbol{h})\right) - \tau\|\boldsymbol{h}\|_1.
\end{aligned}$$

In addition, since the samples are identical drawn:

$$\mathbb{E}\left(\Delta(\boldsymbol{\beta}^*, \boldsymbol{h})\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^* + \boldsymbol{h}\rangle; y_i\right) - f\left(\langle \boldsymbol{x_i}, \boldsymbol{\beta}^*\rangle; y_i\right)\right\} = \mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{h}) - \mathcal{L}(\boldsymbol{\beta}^*).$$

Consequently, we conclude that with probability at least $1 - \frac{\delta}{2}$:

$$\Delta(\boldsymbol{\beta}^*, \boldsymbol{h}) \geq \mathcal{L}(\boldsymbol{\beta}^* + \boldsymbol{h}) - \mathcal{L}(\boldsymbol{\beta}^*) - \tau\|\boldsymbol{h}\|_1. \qquad (34)$$

Step 2: We now lower-bound the right-hand side of Equation (34). Since $\mathcal{L}$ is twice differentiable, a Taylor development around $\boldsymbol{\beta}^*$ gives:

$$\mathcal{L}(\boldsymbol{\beta}^*+\boldsymbol{h})-\mathcal{L}(\boldsymbol{\beta}^*)=\nabla\mathcal{L}(\boldsymbol{\beta}^*)^T\boldsymbol{h}+\frac{1}{2}\boldsymbol{h}^T\nabla^2\mathcal{L}(\boldsymbol{\beta}^*)^T\boldsymbol{h}+o\left(\|\boldsymbol{h}\|_2\right).$$

The optimality of $\boldsymbol{\beta}^*$ implies $\nabla L(\boldsymbol{\beta}^*)=0$. In addition, Theorem 2 states that $\boldsymbol{h}\in\Lambda\left(S_0,\gamma_1,\gamma_2\right)$ with probability at least $1-\frac{\delta}{2}$. Consequently, we can use the restricted eigenvalue condition defined in Assumption $4.2(k^*,\gamma)$. However we do not want to keep the term $o\left(\|\boldsymbol{h}\|_2\right)$ as it can hide non trivial dependencies. To overcome this difficulty, we use the convexity of $\mathcal{L}$ and the maximum radius $r(k^*,\gamma)$ introduced in the growth condition Assumption 5.2.

Case 1: If $\|\boldsymbol{h}\|_2\leq r(k^*)$ – where $r(k^*)$ is a shorthand for $r(k^*,\gamma)$ – then with Theorem 2 and Assumption $4.2(k,\gamma)$, it holds with probability at least $1-\frac{\delta}{2}$:

$$\mathcal{L}(\boldsymbol{\beta}^*+\boldsymbol{h})-\mathcal{L}(\boldsymbol{\beta}^*)\geq\frac{1}{4}\kappa(k^*)\|\boldsymbol{h}\|_2^2. \tag{35}$$

Case 2: If now $\|\boldsymbol{h}\|_2\geq r(k^*)$, then using the convexity of $\mathcal{L}$ thus of $t\to\mathcal{L}\left(\boldsymbol{\beta}^*+t\boldsymbol{h}\right)$, we similarly obtain with same probability:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}^*+\boldsymbol{h})-\mathcal{L}(\boldsymbol{\beta}^*)&\geq\frac{\|\boldsymbol{h}\|_2}{r(k^*)}\left\{\mathcal{L}\left(\boldsymbol{\beta}^*+\frac{r(k^*)}{\|\boldsymbol{h}\|_2}\boldsymbol{h}\right)-\mathcal{L}(\boldsymbol{\beta}^*)\right\}\text{ by convexity}\\&\geq\frac{\|\boldsymbol{h}\|_2}{r(k^*)}\inf_{\substack{\boldsymbol{z}:\ \boldsymbol{z}\in\Lambda(S_0,\gamma_1,\gamma_2)\\\|\boldsymbol{z}\|_2=r(k^*)}}\left\{\mathcal{L}(\boldsymbol{\beta}^*+\boldsymbol{z})-\mathcal{L}(\boldsymbol{\beta}^*)\right\}\\&\geq\frac{\|\boldsymbol{h}\|_2}{r(k^*)}\,\frac{1}{4}\kappa(k^*)r(k^*)^2=\frac{1}{4}\kappa(k^*)r(k^*)\|\boldsymbol{h}\|_2.\end{aligned} \tag{36}$$

Combining Equations (34), (35) and (36), we conclude that with probability at least $1-\delta$ the following restricted strong convexity with L1 tolerance function holds:

$$\Delta\left(\boldsymbol{\beta}^*,\boldsymbol{h}\right)\geq\frac{1}{4}\kappa(k^*)\|\boldsymbol{h}\|_2^2\wedge\frac{1}{4}\kappa(k^*)r(k^*)\|\boldsymbol{h}\|_2-\tau\|\boldsymbol{h}\|_1. \tag{37}$$

To derive the condition for the L2 tolerance function, we use our cone condition derived in Theoreme 2. We recall that S_0 has been defined as the subset of the k^* highest elements of $\boldsymbol{h}$. It thus holds:

$$\begin{aligned}\|\boldsymbol{h}\|_1&=\|\boldsymbol{h}_{S_0}\|_1+\|\boldsymbol{h}_{(S_0)^c}\|_1\\&\leq|\boldsymbol{h}_{S_0}\|_1+\frac{\alpha}{\alpha-1}\|\boldsymbol{h}_{S_0}\|_1+\frac{\sqrt{k^*}}{\alpha-1}\|\boldsymbol{h}_{S_0}\|_2\text{ since }\boldsymbol{h}\in\Lambda\left(S_0,\gamma_1,\gamma_2\right)\\&=\frac{2\alpha-1}{\alpha-1}\|\boldsymbol{h}_{S_0}\|_1+\frac{\sqrt{k^*}}{\alpha-1}\|\boldsymbol{h}_{S_0}\|_2\\&\leq\frac{2\alpha-1}{\alpha-1}\sqrt{k^*}\|\boldsymbol{h}_{S_0}\|_2+\frac{\sqrt{k^*}}{\alpha-1}\|\boldsymbol{h}_{S_0}\|_2\text{ with Cauchy-Schwartz inequality on the }k^*\text{ sparse vector }\boldsymbol{h}_{S_0}\\&\leq\frac{2\alpha}{\alpha-1}\sqrt{k^*}\|\boldsymbol{h}\|_2.\end{aligned} \tag{38}$$

We thus conclude that it holds with probability at least $1-\delta$:

$$\Delta\left(\boldsymbol{\beta}^*,\boldsymbol{h}\right)\geq\frac{1}{4}\kappa(k^*)\|\boldsymbol{h}\|_2^2\wedge\frac{1}{4}\kappa(k^*)r(k^*)\|\boldsymbol{h}\|_2-\frac{2\alpha}{\alpha-1}\tau\sqrt{k^*}\|\boldsymbol{h}\|_2. \tag{39}$$

□

E Proof of Theorem 1

Proof: We now prove our main Theorem 1. Following Equation (21) we have:

$$\Delta\left(\boldsymbol{\beta}^*, \boldsymbol{h}\right) \leq \lambda \|\boldsymbol{h}_{S_0}\|_1 - \lambda \|\boldsymbol{h}_{(S_0)^c}\|_1.$$

Thus using the restricted strong convexity derived in Theorem 4, it holds with probability at least $1-\delta$:

$$\begin{aligned}\frac{1}{4}\kappa(k^*)\left\{\|\boldsymbol{h}\|_2^2 \wedge r(k^*)\|\boldsymbol{h}\|_2\right\} &\leq \frac{2\alpha}{\alpha-1}\tau\sqrt{k^*}\|\boldsymbol{h}\|_2 + \lambda\|\boldsymbol{h}_{S_0}\|_1 - \lambda\|\boldsymbol{h}_{(S_0)^c}\|_1 \\ &\leq \frac{2\alpha}{\alpha-1}\tau\sqrt{k^*}\|\boldsymbol{h}\|_2 + \lambda\sqrt{k^*}\|\boldsymbol{h}_{S_0}\|_2 \\ &\leq \left(\frac{2\alpha}{\alpha-1}\tau + \lambda\right)\sqrt{k^*}\|\boldsymbol{h}\|_2.\end{aligned} \tag{40}$$

With the definitions of τ and λ as in Theorem 2 and 3, Equation (40) leads to:

$$\begin{aligned}\frac{1}{4}\kappa(k^*)\left\{\|\boldsymbol{h}\|_2 \wedge r(k^*)\right\} \leq 12\alpha LM\sqrt{\frac{k^*\log(2pe/k^*)}{n}\log(2/\delta)} \\ + \frac{28\alpha}{\alpha-1}L\mu(k^*)\sqrt{\frac{\log(3)}{n} + \frac{\log\left(4p/k^*\right)}{nk^*} + \frac{\log\left(2/\delta\right)/k^*}{nk^*}}.\end{aligned}$$

Exploiting Assumption 5.2(k^*, γ, δ), and using that $\alpha \geq 2$, we obtain with probability at least $1-\delta$:

$$\|\boldsymbol{h}\|_2^2 \lesssim \left(\frac{\alpha LM}{\kappa(k^*)}\right)^2 \frac{k^*\log\left(p/k^*\right)\log\left(2/\delta\right)}{n} + \left(\frac{\alpha L\mu(k^*)}{\kappa(k^*)}\right)^2 \frac{\log(3) + \log\left(4p/k^*\right)/k^* + \log\left(2/\delta\right)/k^*}{n}.$$

which concludes the proof. □

F Proof of Corollary 1

Proof: In order to derive the bound in expectation, we define the bounded random variable:

$$Z = \frac{\kappa(k^*)^2}{\alpha^2 L^2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2.$$

Since Assumption 5(k^*, γ, δ_0) is satisfied for a small enough δ_0, we can fix C such that $\forall \delta \in (0,1)$, it holds with probability at least $1-\delta$:

$$Z \leq CM^2 H\log(2/\delta) + C\frac{\mu(k^*)^2}{n}\log(2/\delta) \;\text{ where }\; H = \frac{k^*\log(p/k^*)}{n}.$$

Then it holds $\forall t \geq t_0 = \log(4)$:

$$\mathbb{P}\left(Z/C \geq M^2 Ht + \frac{\mu(k^*)^2}{n}t\right) \leq 2e^{-t}.$$

Let $q_0 = M^2 Ht_0 + \frac{\mu(k^*)^2}{n}t_0$, then $\forall q \geq q_0$

$$\mathbb{P}\left(Z/C \geq q\right) \leq 2\exp\left(-\frac{n}{nM^2H + \mu(k^*)^2}\;q\right) \leq 2\exp\left(-\frac{q}{M^2H}\right).$$

Consequently, by integration we have:

$$
\begin{aligned}
\mathbb{E}(Z) &= \int_0^{+\infty} C\mathbb{P}\left(|Z|/C \geq q\right) dq \\
&\leq \int_{q_0}^{+\infty} C\mathbb{P}\left(|Z|/C \geq q\right) dq + Cq_0 \\
&\leq \int_{q_0}^{+\infty} 2Ce^{-\frac{q}{M^2 H}} dq + Cq_0 \\
&\leq 2CM^2He^{-\frac{q_0}{M^2H}} + Cq_0 \\
&\leq 2CM^2H + CM^2H\log(4) + C\frac{\mu(k^*)}{n}\log(4) \\
&\leq C_1\left(M^2H + \frac{\mu(k^*)^2}{n}\right)
\end{aligned} \tag{41}
$$

for $C_1 = 2C + \log(4)$. Hence we conclude:

$$
\mathbb{E}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2\right) \lesssim \left(\frac{\alpha L}{\kappa(k^*)}\right)^2 \left\{M^2\frac{k^*\log(p/k^*)}{n} + \frac{\mu(k^*)}{\sqrt{n}}\right\}.
$$

□

G Proof of Theorem 5

Proof: We fix $\tau > 0$ and denote $\mathbb{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in \mathbb{R}^{n\times p}$ the design matrix.
For $\boldsymbol{\beta} \in \mathbb{R}^p$, we define $\boldsymbol{w}^\tau(\boldsymbol{\beta}) \in \mathbb{R}^n$ by:

$$
w_i^\tau(\boldsymbol{\beta}) = \min\left(1, \frac{1}{2\tau}|z_i|\right)\operatorname{sign}(z_i),\ \forall i
$$

where $z_i = 1 - y_i\boldsymbol{x}_i^T\boldsymbol{\beta},\ \forall i$. We easily check that

$$
\boldsymbol{w}^\tau(\boldsymbol{\beta}) = \underset{\|w\|_\infty \leq 1}{\operatorname{argmax}} \frac{1}{2n}\sum_{i=1}^n (z_i + w_i z_i) - \frac{\tau}{2n}\|w\|_2^2.
$$

Then the gradient of the smooth hinge loss is

$$
\nabla g^\tau(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^n (1 + w_i^\tau(\boldsymbol{\beta}))y_i\boldsymbol{x}_i \in \mathbb{R}^p.
$$

For every couple $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$ we have:

$$
\nabla g^\tau(\boldsymbol{\beta}) - \nabla g^\tau(\boldsymbol{\gamma}) = \frac{1}{2n}\sum_{i=1}^n (w_i^\tau(\boldsymbol{\gamma}) - w_i^\tau(\boldsymbol{\beta}))y_i\boldsymbol{x}_i. \tag{42}
$$

For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ we define the vector $\boldsymbol{a} * \boldsymbol{b} = (a_ib_i)_{i=1}^n$. Then we can rewrite Equation (42) as:

$$
\nabla g^\tau(\boldsymbol{\beta}) - \nabla g^\tau(\boldsymbol{\gamma}) = \frac{1}{2n}\mathbb{X}^T\left[\boldsymbol{y} * (\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta}))\right]. \tag{43}
$$

The operator norm associated to the Euclidean norm of the matrix $\mathbb{X}$ is $\|\mathbb{X}\| = \max_{\|\boldsymbol{z}\|_2=1} \|\mathbb{X}\boldsymbol{z}\|_2$.
Let us recall that $\|\mathbb{X}\|^2 = \|\mathbb{X}^T\|^2 = \|\mathbb{X}^T\mathbb{X}\| = \mu_{\max}(\mathbb{X}^T\mathbb{X})$ corresponds to the highest eigenvalue of the matrix $\mathbb{X}^T\mathbb{X}$.
Consequently, Equation (43) leads to:

$$\|\nabla L^\tau(\boldsymbol{\beta}) - \nabla L^\tau(\boldsymbol{\gamma})\|_2 \leq \frac{1}{2n}\|\mathbb{X}\| \, \|\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta})\|_2 \,. \tag{44}$$

In addition, the first order necessary conditions for optimality applied to $\boldsymbol{w}^\tau(\boldsymbol{\beta})$ and $\boldsymbol{w}^\tau(\boldsymbol{\gamma})$ give:

$$\sum_{i=1}^n \left\{ \frac{1}{2n}(1 - y_i \boldsymbol{x}_i^T \boldsymbol{\beta}) - \frac{\tau}{n} w_i^\tau(\boldsymbol{\beta}) \right\} \{ w_i^\tau(\boldsymbol{\gamma}) - w_i^\tau(\boldsymbol{\beta}) \} \leq 0, \tag{45}$$

and

$$\sum_{i=1}^n \left\{ \frac{1}{2n}(1 - y_i \boldsymbol{x}_i^T \boldsymbol{\gamma}) - \frac{\tau}{n} w_i^\tau(\boldsymbol{\gamma}) \right\} \{ w_i^\tau(\boldsymbol{\beta}) - w_i^\tau(\boldsymbol{\gamma}) \} \leq 0. \tag{46}$$

Then by adding Equations (45) and (46) and rearranging the terms we have:

$$\begin{aligned}
&\tau \|\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta})\|_2^2 \\
&\leq \frac{1}{2} \sum_{i=1}^n y_i \boldsymbol{x}_i^T (\boldsymbol{\beta} - \boldsymbol{\gamma}) \left(w_i^\tau(\boldsymbol{\gamma}) - w_i^\tau(\boldsymbol{\beta}) \right) \\
&\leq \frac{1}{2} \|\mathbb{X}\left(\boldsymbol{\beta} - \boldsymbol{\gamma}\right)\|_2 \|\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta})\|_2 \\
&\leq \frac{1}{2} \|\mathbb{X}\| \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2 \|\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta})\|_2,
\end{aligned}$$

where we have used Cauchy-Schwartz inequality. We then have:

$$\|\boldsymbol{w}^\tau(\boldsymbol{\gamma}) - \boldsymbol{w}^\tau(\boldsymbol{\beta})\|_2 \leq \frac{1}{2\tau} \|\mathbb{X}\| \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2. \tag{47}$$

We conclude the proof by combining Equations (44) and (47):

$$\begin{aligned}
\|\nabla L^\tau(\boldsymbol{\beta}) - \nabla L^\tau(\boldsymbol{\gamma})\|_2 &\leq \frac{1}{4n\tau} \|\mathbb{X}\|^2 \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2 \\
&= \frac{\mu_{\max}(n^{-1}\mathbb{X}^T\mathbb{X})}{4\tau} \|\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2.
\end{aligned}$$

The case of Quantile Regression: For the quantile regression loss, the same smoothing method applies. Let us simply note that:

$$\begin{aligned}
\rho_\theta(x) = \max\left((\theta - 1)x, \, \theta x\right) &= \frac{1}{2}((2\theta - 1)x + |x|) \\
&= \max_{|w| \leq 1} \frac{1}{2}((2\theta - 1)x + wx).
\end{aligned}$$

Hence we can immediately use the same steps than for the hinge loss – which is a particular case of the quantile regression loss – and define the smooth quantile regression loss g_θ^τ. Its gradient is:

$$\nabla g_\theta^\tau(\boldsymbol{\beta}) = -\frac{1}{2n} \sum_{i=1}^n (2\theta - 1 + w_i^\tau(\boldsymbol{\beta})) y_i \boldsymbol{x}_i \in \mathbb{R}^p, \tag{48}$$

where we still have $w_i^\tau = \min\left(1, \frac{1}{2\tau}|z_i|\right) \operatorname{sign}(z_i)$ but now $z_i = y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}, \; \forall i$. The Lipschitz constant of ∇g_θ^τ is still given by Theorem 5. □

H Proof of Theorem 6

Proof: We still assume $|h_1| \geq \ldots \geq |h_p|$. Following Equation (21) it holds:

$$S(\boldsymbol{h}) \leq \Delta(\boldsymbol{h}) \leq \eta|\boldsymbol{\beta}^*|_S - \eta|\hat{\boldsymbol{\beta}}|_S. \tag{49}$$

We want to upper-bound the right-hand side of Equation (49). We define the permutation $\phi \in \mathcal{S}_p$ such that $|\boldsymbol{\beta}^*|_S = \sum_{j=1}^{k^*} \lambda_j |\beta^*_{\phi(j)}|$ and $|\hat{\beta}_{\phi(k^*+1)}| \geq \ldots \geq |\hat{\beta}_{\phi(p)}|$ – ϕ is uniquely defined. Hence it holds:

$$\begin{aligned}
\frac{1}{\eta}\Delta(\boldsymbol{h}) &\leq \sum_{j=1}^{k^*} \lambda_j |\beta^*_{\phi(j)}| - \max_{\psi \in \mathcal{S}_p} \sum_{j=1}^{p} \lambda_j |\hat{\beta}_{\psi(j)}| \quad \text{by definition of Slope} \\
&\leq \sum_{j=1}^{k^*} \lambda_j \left(|\beta^*_{\phi(j)}| - |\hat{\beta}_{\phi(j)}|\right) - \sum_{j=k^*+1}^{p} \lambda_j |\hat{\beta}_{\phi(j)}| \quad \text{since } \phi \in \mathcal{S}_p \\
&= \sum_{j=1}^{k^*} \lambda_j |h_{\phi(j)}| - \sum_{j=k^*+1}^{p} \lambda_j |\hat{\beta}_{\phi(j)}| \\
&\leq \sum_{j=1}^{k^*} \lambda_j |h_{\phi(j)}| - \sum_{j=k^*+1}^{p} \lambda_j |h_{\phi(j)}|.
\end{aligned} \tag{50}$$

Since λ is monotonically non decreasing: $\sum_{j=1}^{k^*} \lambda_j |h_{\phi(j)}| \leq \sum_{j=1}^{k^*} \lambda_j |h_j|$.
Because $|h_{\phi(k^*+1)}| \geq \ldots \geq |h_{\phi(p)}|$: $\sum_{j=k^*+1}^{p} \lambda_j |h_j| \leq \sum_{j=k^*+1}^{p} \lambda_j |h_{\phi(j)}|$.
In addition, Equation (22) from Appendix B leads to, with probability at least $1 - \frac{\delta}{2}$:

$$|S(\boldsymbol{h})| \leq 14LM\sqrt{\frac{\log(2/\delta)}{n}} \sum_{j=1}^{p} \lambda_j |h_j| \leq \frac{\eta}{\alpha}|\boldsymbol{h}|_S,$$

where η is defined in the statement of the theorem. Thus, combining this last equation with Equation (50), it holds with probability at least $1 - \frac{\delta}{2}$:

$$-\frac{1}{\alpha}|\boldsymbol{h}|_S \leq \sum_{j=1}^{k^*} \lambda_j |h_j| - \sum_{j=k^*+1}^{p} \lambda_j |h_j|,$$

which is equivalent to saying that with probability at least $1 - \frac{\delta}{2}$:

$$\sum_{j=k^*+1}^{p} \lambda_j |h_j| \leq \frac{\alpha+1}{\alpha-1} \sum_{j=1}^{k^*} \lambda_j |h_j|, \tag{51}$$

that is $\boldsymbol{h} \in \Gamma\left(k^*, \frac{\alpha+1}{\alpha-1}\right)$. □

I Proof of Corollary 2

Proof: We follow the proof of Theorem 1. Theorem 3 still holds with L1 tolerance loss function – the results for L2 is however no longer true. In addition,the restricted strong convexity derived in Lemma 4 applies for Slope. We consequently obtain with probability at least $1-\delta$:

$$\begin{aligned}\frac{1}{4}\tilde{\kappa}(k^*,\omega)\left\{\|\boldsymbol{h}\|_2^2 \wedge r(k^*)\|\boldsymbol{h}\|_2\right\} &\leq \tau\|\boldsymbol{h}\|_1 + \eta\sum_{j=1}^{k^*}\lambda_j|h_j| - \eta\sum_{j=k^*+1}^{p}\lambda_j|h_j| \\ &\leq \tau\|\boldsymbol{h}_{S_0}\|_1 + \eta\sum_{j=1}^{k^*}\lambda_j|h_j| + \tau\|\boldsymbol{h}_{(S_0)^c}\|_1 - \eta\sum_{j=k^*+1}^{p}\lambda_j|h_j| \\ &\leq \tau\|\boldsymbol{h}_{S_0}\|_1 + \eta\sum_{j=1}^{k^*}\lambda_j|h_j| + (\tau-\eta\lambda_p)\|\boldsymbol{h}_{(S_0)^c}\|_1.\end{aligned} \tag{52}$$

We want $\tau \leq \eta\lambda_p$, that is $14L\mu(k^*)\sqrt{\frac{\log(3)}{n}+\frac{\log(4p/k)}{nk}+\frac{\log(2/\delta)}{nk}} \leq 14\alpha LM\sqrt{\frac{\log(2e)}{n}\log(2/\delta)}$, which is satisfied since $\mu(k^*) \leq \alpha M$. Hence we obtain, similarly to Section E:

$$\begin{aligned}\frac{1}{4}\tilde{\kappa}(k^*,\omega)\left\{\|\boldsymbol{h}\|_2^2 \wedge r(k^*)\|\boldsymbol{h}\|_2\right\} &\leq \tau\|\boldsymbol{h}_{S_0}\|_1 + \eta\sum_{j=1}^{k^*}\lambda_j|h_j| \\ &\leq \tau\sqrt{k^*}\|\boldsymbol{h}_{S_0}\|_2 + \eta\sqrt{k^*\log(2pe/k^*)}\|\boldsymbol{h}_{S_0}\|_2 \\ &\leq 2\eta\sqrt{k^*\log(2pe/k^*)}\|\boldsymbol{h}_{S_0}\|_2 \text{ since } \tau \leq \eta\lambda_p \leq \eta\lambda_{k^*} \\ &\leq 28\alpha LM\sqrt{\frac{k^*\log(2pe/k^*)}{n}\log(2/\delta)}\|\boldsymbol{h}\|_2.\end{aligned}$$

This last equation is very similar to Equation (40) in the proof of Theorem 1. We conclude the proof identically, and obtain a similar bound in expectation by following the proof of Corollary 1. □