Supplementary Material to Robust Structured Estimation with Single-Index Models

Sheng Chen [1] **Arindam Banerjee** [1]

Abstract

In this supplementary material, we present the deferred proofs of the results in the main paper.

1. Proof of Claim 1

Statement of Claim 1: *Suppose that each element x_i of $\mathbf{x}$ is sampled i.i.d. from Rademacher distribution, i.e., $\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = 0.5$. Under model (3) with noise $\epsilon = 0$, there exists a $\bar{\boldsymbol{\theta}} \in \mathbb{S}^{p-1}$ together with a monotone $\bar{f}$, such that $\text{supp}(\bar{\boldsymbol{\theta}}) = \text{supp}(\boldsymbol{\theta}^*)$ and $y_i = \bar{f}(\langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i \rangle)$ for data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with arbitrarily large sample size n, while $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 > \delta$ for some constant δ.*

Proof: In the noiseless setting with unknown f^*, provided that $\mathcal{S} \triangleq \text{supp}(\boldsymbol{\theta}^*)$ is given and $|\mathcal{S}| = s$, the estimation of $\boldsymbol{\theta}^*$ is simplified as

$$\begin{aligned} &\text{Find } \ \boldsymbol{\theta}_{\mathcal{S}} \in \mathbb{S}^{s-1} \\ &\text{s.t. } \ \text{sign}\left(\langle \boldsymbol{\theta}_{\mathcal{S}}, \mathbf{x}_{i\mathcal{S}} - \mathbf{x}_{j\mathcal{S}} \rangle\right) = \text{sign}(y_i - y_j), \\ &\forall\, 1 \leq i < j \leq n \ , \end{aligned} \tag{S.1}$$

any of whose solution $\boldsymbol{\theta}$ can be true $\boldsymbol{\theta}^*$ on the premise that no other information is available, since there always exists a monotone f satisfying $f(\langle \boldsymbol{\theta}, \mathbf{x}_i \rangle) = y_i$. Given the distribution of $\mathbf{x}$, $\mathbf{x}_{i\mathcal{S}} - \mathbf{x}_{j\mathcal{S}}$ only has 3^s possibilities even if $n \to +\infty$. We denote the feasible set of (S.1) by $\mathcal{C}$, which is basically an intersection of $\mathbb{S}^{s-1}$ and at most $\min\{n(n-1), 3^p\}$ halfspaces (or hyperplanes if $y_i = y_j$). Depending on the 3 different values of each $\text{sign}(y_i - y_j)$, this feasible set $\mathcal{C}$ has at most $3^{\min\{n(n-1), 3^p\}}$ possibilities, which is finite, and the union of them should be $\mathbb{S}^{s-1}$. When $s \geq 2$ and the constant δ is small enough, we can always find a $\mathcal{C}$, in which there exist two different points away by δ. Specify them as $\boldsymbol{\theta}_{*\mathcal{S}}$ and $\bar{\boldsymbol{\theta}}_{\mathcal{S}}$ respectively, and we are unable to distinguish between them, as both can be solution to (S.1) for any samples. ■

2. Proof of Lemma 1

Statement of Lemma 1: *Suppose the distribution of y in model (1) depends on $\mathbf{x}$ through $\langle \boldsymbol{\theta}^*, \mathbf{x} \rangle$ and we define accordingly*

$$\begin{aligned} &b_i\left(z_1, \ldots, z_m; \boldsymbol{\theta}^*\right) = \\ &\mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) | \langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle = z_1, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle = z_m\right] \ , \end{aligned} \tag{S.2}$$

With $\mathbf{x}$ being standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{u}$ defined in (4) satisfies

$$\mathbb{E}\left[\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right)\right] = \beta \boldsymbol{\theta}^* \ , \tag{S.3}$$

where $\beta = \sum_{i=1}^m \mathbb{E}[b_i\left(g_1, \ldots, g_m; \boldsymbol{\theta}^*\right) \cdot g_i]$, and $g_1, \ldots, g_m$ are i.i.d. standard Gaussian.

Proof: Let $\boldsymbol{\theta}_\perp$ be any vector orthogonal to $\boldsymbol{\theta}^*$. For convenience, we use the shorthand notation $\mathbf{u}$ for $\mathbf{u}\left((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\right)$. Then we have

$$\begin{aligned} \langle \mathbb{E}\mathbf{u}, \boldsymbol{\theta}_\perp \rangle &= \mathbb{E}\left[\sum_{i=1}^m q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle \cdot \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) | \mathbf{x}_1, \ldots, \mathbf{x}_m\right]\right] \ (*) \end{aligned}$$

As $\mathbf{x}_i$ follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle$ and $\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle$ are two zero-mean independent Gaussian random variables. Since the distribution of y_i depends on $\mathbf{x}$ only via $\langle \boldsymbol{\theta}^*, \mathbf{x}_i \rangle$, we can split the expectation and obtain

$$\begin{aligned} (*) &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle \cdot b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right)\right] \\ &= \sum_{i=1}^m \mathbb{E}\left[\langle \mathbf{x}_i, \boldsymbol{\theta}_\perp \rangle\right] \cdot \mathbb{E}\left[b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right)\right] \\ &= 0 \ . \end{aligned}$$

[1]Department of Computer Science & Engineering, University of Minnesota-Twin Cities, Minnesota, USA. Correspondence to: Sheng Chen <shengc@cs.umn.edu>, Arindam Banerjee <banerjee@cs.umn.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, 2017. JMLR: W&CP.

Hence $\mathbf{u}$ has to point towards either $\boldsymbol{\theta}^*$ or $-\boldsymbol{\theta}^*$, and note that

$$\begin{aligned}\langle \mathbb{E}\mathbf{u}, \boldsymbol{\theta}^* \rangle &= \sum_{i=1}^{m} \mathbb{E}\left[q_i\left(y_1, \ldots, y_m\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle\right] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[b_i\left(\langle \boldsymbol{\theta}^*, \mathbf{x}_1 \rangle, \ldots, \langle \boldsymbol{\theta}^*, \mathbf{x}_m \rangle; \boldsymbol{\theta}^*\right) \cdot \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle\right] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[b_i\left(g_1, \ldots, g_m; \boldsymbol{\theta}^*\right) \cdot g_i\right] = \beta\end{aligned}$$

We complete the proof by recalling that $\|\boldsymbol{\theta}^*\|_2 = 1$, thus $\mathbb{E}\mathbf{u} = \beta \boldsymbol{\theta}^*$. ■

3. Proof of Theorem 1

We first provide a lemma that is useful for bounding the Gaussian width of unions of sets, which originates in Maurer et al. (2014).

Lemma A (Lemma 2 in Maurer et al. (2014)) *Let $M > 4$, $\mathcal{A}_1, \cdots, \mathcal{A}_M \subset \mathbb{R}^p$, and $\mathcal{A} = \cup_m \mathcal{A}_m$. The Gaussian width of $\mathcal{A}$ satisfies*

$$w(\mathcal{A}) \leq \max_{1 \leq m \leq M} w(\mathcal{A}_m) + 2 \sup_{\mathbf{z} \in \mathcal{A}} \|\mathbf{z}\|_2 \sqrt{\log M} \qquad \text{(S.4)}$$

Statement of Theorem 1: *Suppose that the optimization (9) can be solved to global minimum. Then the following error bound holds for the minimizer $\hat{\boldsymbol{\theta}}$ with probability at least $1 - C'' \exp\left(-w^2\left(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)\right)\right)$,*

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq \frac{C\kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)) + C'}{\sqrt{n}}, \qquad \text{(S.5)}$$

where κ is the sub-Gaussian norm of a standard Gaussian random variable, and C, C', C'' are all absolute constant. *Proof:* We use the shorthand notation $\mathcal{A}_{\mathcal{K}}$ for the set $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$. As $\hat{\boldsymbol{\theta}}$ attains the global minimum of (9), we have

$$\begin{aligned}&\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \hat{\mathbf{u}} \rangle \geq 0 \iff \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* + \boldsymbol{\theta}^* \right\rangle \geq 0 \\ &\implies \langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^* \rangle \geq 1 - \left\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle \\ &\qquad \geq 1 - \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \cdot \sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} \left\langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle\end{aligned}$$

In order to bound the supremum above, we use the result from generic chaining. We define the stochastic process $\{Z_{\mathbf{v}} = \langle \mathbf{v}, \hat{\mathbf{u}}/\beta - \boldsymbol{\theta}^* \rangle\}_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}}$. First, we need to check the process has sub-Gaussian incremental. For simplicity, we denote $\mathbf{u}\left((\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_m}, y_{i_m})\right)$ by $\mathbf{u}_{i_1,\ldots,i_m}$. By the definitions and properties of sub-Gaussian norm (Vershynin, 2012), the sub-Gaussian norm of $\mathbf{u}_{i_1,\ldots,i_m}$ satisfies

$$\begin{aligned}\|\mathbf{u}_{i_1,\ldots,i_m}\|_{\psi_2} &= \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \left\| \sum_{j=1}^{m} q_j\left(y_{1_1}, \ldots, y_{i_m}\right) \cdot \langle \mathbf{x}_j, \mathbf{v} \rangle \right\|_{\psi_2} \\ &\leq \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \left\| \sum_{j=1}^{m} |\langle \mathbf{x}_j, \mathbf{v} \rangle| \right\|_{\psi_2} \\ &\leq m \cdot \sup_{\mathbf{v} \in \mathbb{S}^{p-1}} \left\||\langle \mathbf{x}_j, \mathbf{v} \rangle|\right\|_{\psi_2} \leq \kappa m,\end{aligned}$$

thus we know $\|\langle \mathbf{u}_{i_1,\ldots,i_m}, \mathbf{v} - \mathbf{w} \rangle\|_{\psi_2} \leq \kappa m \cdot \|\mathbf{v} - \mathbf{w}\|_2$. By Lemma 2, we have

$$\begin{aligned}&\mathbb{P}\left(|Z_{\mathbf{v}} - Z_{\mathbf{w}}| > \delta\right) = \mathbb{P}\left(\left|\left\langle \mathbf{v} - \mathbf{w}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle\right| > \delta\right) \\ &= \mathbb{P}\Bigg(\Bigg|\frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1,\ldots,i_m \leq n \\ i_1 \neq \ldots \neq i_m}} \frac{1}{\beta} \cdot \langle \mathbf{u}_{i_1,\ldots,i_m}, \mathbf{v} - \mathbf{w} \rangle \\ &\qquad - \langle \mathbf{v} - \mathbf{w}, \boldsymbol{\theta}^* \rangle \Bigg| > \delta\Bigg) \\ &\leq 2\exp\left(-C\left\lfloor \frac{n}{m} \right\rfloor \cdot \frac{\beta^2 \delta^2}{m^2 \kappa^2 \cdot \|\mathbf{v} - \mathbf{w}\|_2^2}\right) \\ &\leq 2\exp\left(-C' \cdot \frac{n\beta^2 \delta^2}{m^3 \kappa^2 \cdot \|\mathbf{v} - \mathbf{w}\|_2^2}\right),\end{aligned}$$

where we set $C' = C/2$. Therefore we can conclude that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental w.r.t. the metric $s(\mathbf{v}, \mathbf{w}) \triangleq \kappa m^{\frac{3}{2}} \cdot \|\mathbf{v} - \mathbf{w}\|_2 / \beta\sqrt{n}$. Now applying Lemma 3 to $\{Z_{\mathbf{v}}\}$, we obtain

$$\begin{aligned}&\mathbb{P}\Bigg(\sup_{\mathbf{v}, \mathbf{w} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} |Z_{\mathbf{v}} - Z_{\mathbf{w}}| \geq C_1\Big(\gamma_2\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, s\right) \\ &\qquad + \delta \cdot \operatorname{diam}\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, s\right)\Big)\Bigg) \leq C_2 \exp\left(-\delta^2\right) \\ &\implies \mathbb{P}\Bigg(\sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} |Z_{\mathbf{v}}| \geq \frac{C_1 \kappa m^{\frac{3}{2}}}{\beta\sqrt{n}} \cdot \Big(\gamma_2\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, \|\cdot\|_2\right) \\ &\qquad + 2\delta\Big)\Bigg) \leq C_2 \exp\left(-\delta^2\right)\end{aligned}$$

Using Lemma 4 $\gamma_2\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}, \|\cdot\|_2\right) \leq C_0 \cdot w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right)$ and taking $\delta = w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right)$, we get

$$\begin{aligned}&\sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} \left\langle \mathbf{v}, \frac{\hat{\mathbf{u}}}{\beta} - \boldsymbol{\theta}^* \right\rangle \leq \sup_{\mathbf{v} \in \mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}} |Z_{\mathbf{v}}| \\ &\leq \frac{C_3 \kappa m^{\frac{3}{2}}}{\beta\sqrt{n}} \cdot w\left(\mathcal{A}_{\mathcal{K}} \cup \{\mathbf{0}\}\right) \leq \frac{C_3 \kappa m^{\frac{3}{2}}}{\beta} \cdot \frac{w\left(\mathcal{A}_{\mathcal{K}}\right) + C_4}{\sqrt{n}}\end{aligned}$$

with probability at least $1 - C_2 \exp\left(-w^2\left(\mathcal{A}_{\mathcal{K}}\right)\right)$. The last inequality follows from Lemma A. Now we turn to the quantity $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2$,

$$\begin{aligned}
&\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2^2 \leq 2-2\langle\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*\rangle \\
&\leq 2-2\left(1-\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\frac{C_3\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{A}_{\mathcal{K}}\right)+C_4}{\sqrt{n}}\right) \\
&\leq \|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\frac{2C_3\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{A}_{\mathcal{K}}\right)+C_4}{\sqrt{n}}\ .
\end{aligned}$$

We finish the proof by letting $C=2C_3$, $C'=C_4$ and $C''=C_2$. ■

4. Proof of Theorem 2

Statement of Theorem 2: *Define the following set for any $\rho>1$,*

$$\mathcal{A}_\rho\left(\boldsymbol{\theta}^*\right)=\operatorname{cone}\left\{\mathbf{v}\ \middle|\ \|\mathbf{v}+\boldsymbol{\theta}^*\|\leq\|\boldsymbol{\theta}^*\|+\frac{\|\mathbf{v}\|}{\rho}\right\}\bigcap\mathbb{S}^{p-1} \tag{S.6}$$

If we set $\lambda=\rho\left\|\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*\right\|_*=O(\rho m^{3/2}w(\mathcal{B}_{\|\cdot\|})/\sqrt{n})$ and it satisfies $\lambda<\|\hat{\mathbf{u}}\|_*$, then with probability at least $1-C'\exp\left(-w^2\left(\mathcal{B}_{\|\cdot\|}\right)\right)$, $\hat{\boldsymbol{\theta}}$ in (10) satisfies

$$\left\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2\leq\frac{C(1+\rho)\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)\cdot w\left(\mathcal{B}_{\|\cdot\|}\right)}{\sqrt{n}}\ , \tag{S.7}$$

where $\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)=\sup_{\mathbf{v}\in\mathcal{A}_\rho(\boldsymbol{\theta}^*)}\|\mathbf{v}\|$ and $\mathcal{B}_{\|\cdot\|}=\{\mathbf{v}\mid\|\mathbf{v}\|\leq1\}$ is the unit ball of norm $\|\cdot\|$.

Proof: Based on the optimality of $\hat{\boldsymbol{\theta}}$, we have

$$\begin{gathered}
-\langle\hat{\mathbf{u}},\hat{\boldsymbol{\theta}}\rangle+\lambda\|\hat{\boldsymbol{\theta}}\|\leq-\langle\hat{\mathbf{u}},\boldsymbol{\theta}^*\rangle+\lambda\|\boldsymbol{\theta}^*\| \quad\Longrightarrow \\
\langle\beta\boldsymbol{\theta}^*-\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle+\lambda\|\hat{\boldsymbol{\theta}}\| \\
\leq\langle\beta\boldsymbol{\theta}^*-\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\boldsymbol{\theta}^*\rangle+\lambda\|\boldsymbol{\theta}^*\| \quad\Longrightarrow \\
\beta(1-\langle\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle)\leq\langle\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle+\lambda(\|\boldsymbol{\theta}^*\|-\|\hat{\boldsymbol{\theta}}\|)
\end{gathered}$$

Since $\langle\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle\leq1$, we have

$$\begin{aligned}
&\langle\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle+\lambda\left(\|\boldsymbol{\theta}^*\|-\|\hat{\boldsymbol{\theta}}\|\right)\geq0 \quad\Longrightarrow \\
&\|\hat{\boldsymbol{\theta}}\|\leq\|\boldsymbol{\theta}^*\|+\frac{1}{\lambda}\cdot\langle\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle \\
&\quad\leq\|\boldsymbol{\theta}^*\|+\frac{1}{\lambda}\cdot\|\hat{\mathbf{u}}-\beta\boldsymbol{\theta}^*\|_*\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| \\
&\quad=\|\boldsymbol{\theta}^*\|+\frac{1}{\rho}\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\| \quad\Longrightarrow\quad \hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\in\mathcal{A}_\rho(\boldsymbol{\theta}^*)
\end{aligned}$$

Therefore it follows that

$$\begin{aligned}
&1-\langle\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}\rangle\leq\langle\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*,\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\rangle+\frac{\lambda}{\beta}\left(\|\boldsymbol{\theta}^*\|-\|\hat{\boldsymbol{\theta}}\|\right) \\
&\leq\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\left(\left\|\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\|_*\cdot\frac{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2}+\frac{\lambda}{\beta}\cdot\frac{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2}\right) \\
&\leq(1+\rho)\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\left\|\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\|_*\cdot\sup_{\mathbf{v}\in\mathcal{A}_\rho(\boldsymbol{\theta}^*)}\|\mathbf{v}\| \\
&=(1+\rho)\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2\cdot\left\|\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\|_*\cdot\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)
\end{aligned} \tag{S.8}$$

Now we try to bound $\left\|\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\|_*$. We first rewrite it as $\left\|\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*\right\|_*=\sup_{\mathbf{v}\in\mathcal{B}_{\|\cdot\|}}\left\langle\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*,\mathbf{v}\right\rangle$. Construct the stochastic process $\{Z_{\mathbf{v}}=\langle\mathbf{v},\hat{\mathbf{u}}/\beta-\boldsymbol{\theta}^*\rangle\}_{\mathbf{v}\in\mathcal{B}_{\|\cdot\|}}$, and it is not difficult to verify that $\{Z_{\mathbf{v}}\}$ has sub-Gaussian incremental using the proof in Theorem 1. Now applying Lemma 3 and 4, we have

$$\begin{aligned}
\sup_{\mathbf{v}\in\mathcal{B}_{\|\cdot\|}}\left\langle\frac{\hat{\mathbf{u}}}{\beta}-\boldsymbol{\theta}^*,\mathbf{v}\right\rangle&=\frac{1}{2}\cdot\sup_{\mathbf{v},\mathbf{w}\in\mathcal{B}_{\|\cdot\|}}|Z_{\mathbf{v}}-Z_{\mathbf{w}}| \\
&\leq\frac{C_1\kappa m^{\frac{3}{2}}}{\beta}\cdot\frac{w\left(\mathcal{B}_{\|\cdot\|}\right)}{\sqrt{n}}\ ,
\end{aligned} \tag{S.9}$$

with probability at least $1-C'\exp\left(-w^2\left(\mathcal{B}_{\|\cdot\|}\right)\right)$. Therefore we know that λ satisfies

$$\lambda=O\left(\frac{\rho m^{3/2}w(\mathcal{B}_{\|\cdot\|})}{\sqrt{n}}\right)$$

If $\hat{\boldsymbol{\theta}}=\mathbf{0}$ is the minimizer, the first-order optimality should hold, i.e.,

$$\hat{\mathbf{u}}\in\lambda\cdot\partial\|\mathbf{0}\| \quad\Longrightarrow\quad \|\hat{\mathbf{u}}\|_*\leq\lambda$$

Hence if $\lambda<\|\hat{\mathbf{u}}\|_*$, $\mathbf{0}$ cannot be the minimizer, which means that the minimum of (10) must be negative. So we can assert that $\|\hat{\boldsymbol{\theta}}\|_2=1$, otherwise we can normalize $\hat{\boldsymbol{\theta}}$ to get a smaller objective value. Combining (S.8) and (S.9), we finally get

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|&=\frac{2-2\langle\hat{\boldsymbol{\theta}},\boldsymbol{\theta}^*\rangle}{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|} \\
&\leq\frac{Cm\kappa(1+\rho)}{\beta}\cdot\frac{\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)\cdot w\left(\mathcal{B}_{\|\cdot\|}\right)}{\sqrt{n}}\ ,
\end{aligned}$$

where the equality uses the fact that $\|\hat{\boldsymbol{\theta}}\|_2=1$. ■

5. Proof of Corollary 1

Statement of Corollary 1: *Assume that $\{(\mathbf{x}_i,y_i)\}_{i=1}^n$ follow 1-bit CS model in (2) and $\hat{\mathbf{u}}$ is given as (14). For any*

s-sparse $\boldsymbol{\theta}^*$, with high probability, $\hat{\boldsymbol{\theta}}$ produced by both (15) and (17) (i.e., $\hat{\boldsymbol{\theta}}^{ks}$ and $\hat{\boldsymbol{\theta}}^{ps}$) satisfy

$$\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}}\right) \tag{S.10}$$

Proof: For the k-support norm estimator, the cone $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$ is given by

$$\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*) = \text{cone}\left\{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \;\middle|\; \|\hat{\boldsymbol{\theta}}\|_0 \leq s, \|\hat{\boldsymbol{\theta}}\|_2 \leq 1\right\} \bigcap \mathbb{S}^{p-1}$$
$$\implies \quad \mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*) \subseteq \mathcal{S} = \{\mathbf{v} \mid \|\mathbf{v}\|_0 \leq 2s\} \cap \mathbb{S}^{p-1}$$

Using (19) from (Chen & Banerjee, 2015), we have

$$w(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)) \leq w(\mathcal{S}) \leq O\left(\sqrt{s \log p}\right) .$$

By Theorem 1, the error of k-support norm estimator satisfies

$$\left\|\hat{\boldsymbol{\theta}}^{\text{ks}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}}\right)$$

For the passive algorithm, if we choose $\rho = 2$, the restricted norm compatibility $\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right)$ for L_1 norm satisfies

$$\Psi\left(\mathcal{A}_\rho(\boldsymbol{\theta}^*)\right) \leq 4\sqrt{s} \tag{S.11}$$

according to the results in (Negahban et al., 2012; Banerjee et al., 2014). Chen & Banerjee (2015) also show that the Gaussian width of the L_1-norm ball is bounded by

$$w(\mathcal{B}_{L_1}) \leq O\left(\sqrt{\log p}\right) . \tag{S.12}$$

Now combining (S.11), (S.12) and Theorem 2, we can conclude that

$$\left\|\hat{\boldsymbol{\theta}}^{\text{ps}} - \boldsymbol{\theta}^*\right\|_2 \leq O\left(\sqrt{\frac{s \log p}{n}}\right) ,$$

which completes the proof. ■

6. Proof of Proposition 1

Statement of Proposition 1: *Given $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, let $\pi^\downarrow$ be the permutation of $\{1, \ldots, n\}$ such that $y_{\pi_1^\downarrow} > y_{\pi_2^\downarrow} > \ldots > y_{\pi_n^\downarrow}$. Then we have*

$$\hat{\mathbf{h}} = \frac{2}{n(n-1)} \sum_{i=1}^{n} (n + 1 - 2i) \cdot \mathbf{x}_{\pi_i^\downarrow} \tag{S.13}$$

Proof: We rearrange the terms inside the summation of (21) based on $\pi^\downarrow$,

$$\begin{aligned}
\hat{\mathbf{h}} &= \frac{1}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \text{sign}(y_i - y_j) \cdot (\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{2}{n(n-1)} \sum_{\substack{1 \leq i,j \leq n \\ i \neq j}} \text{sign}(y_i - y_j) \cdot \mathbf{x}_i \\
&= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq \pi_i^\downarrow} \text{sign}\left(y_{\pi_i^\downarrow} - y_j\right) \cdot \mathbf{x}_{\pi_i^\downarrow} \\
&= \frac{2}{n(n-1)} \sum_{i=1}^{n} (n + 1 - 2i) \cdot \mathbf{x}_{\pi_i^\downarrow} ,
\end{aligned}$$

where the last inequality uses the fact that there are $(i-1)$ y_j larger than and $(n-i)$ smaller than $y_{\pi_i^\downarrow}$, thus $\sum_{j \neq \pi_i^\downarrow} \text{sign}\left(y_{\pi_i^\downarrow} - y_j\right) = (n-i) - (i-1) = n+1-2i$. ■

7. Proof of Proposition 2

Statement of Proposition 2: *For s-fused-sparse $\boldsymbol{\theta}^*$, the Gaussian width of set $\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)$ with $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s,\ \|\boldsymbol{\theta}\|_2 = 1\}$ satisfies*

$$w(\mathcal{A}_{\mathcal{K}}(\boldsymbol{\theta}^*)) \leq O(\sqrt{s \log p}) \tag{S.14}$$

Proof: Define the following sets

$$\begin{aligned}
\mathcal{T}_{i,j} = \Big\{ \alpha \mathbf{u} \in \mathbb{R}^p \;\Big|\; & u_1 = \ldots = u_{i-1} = u_{j+1} = \ldots = u_p = 0, \\
& u_i = \ldots = u_j = \frac{1}{\sqrt{j - i + 1}},\ |\alpha| \leq \sqrt{2s + 1} \Big\}
\end{aligned} \tag{S.15}$$

$$\mathcal{T} = \bigcup_{i \leq j} \mathcal{T}_{i,j} \tag{S.16}$$

For each $\mathcal{T}_{i,j}$, its Gaussian width can be calculated as

$$\begin{aligned}
w(\mathcal{T}_{i,j}) &= \mathbb{E}\left[\sup_{\mathbf{v} \in \mathcal{T}_{i,j}} \langle \mathbf{v}, \mathbf{g} \rangle\right] = \sqrt{2s+1} \cdot \mathbb{E}\left[|\langle \mathbf{u}, \mathbf{g} \rangle|\right] \\
&= \sqrt{2s+1} \cdot \mathbb{E}\,|g| = O(\sqrt{2s+1}) ,
\end{aligned}$$

where $\mathbf{u}$ is defined in (S.15) and g is a standard Gaussian random variable. We apply Lemma A to $\mathcal{T}$, and obtain

$$\begin{aligned}
w(\mathcal{T}) &\leq \max_{i \leq j} w(\mathcal{T}_{i,j}) + 2 \sup_{\mathbf{z} \in \mathcal{T}} \|\mathbf{z}\|_2 \sqrt{\log\left(\binom{p}{2} + p\right)} \\
&\leq O(\sqrt{2s+1}) + O(\sqrt{2s+1} \cdot \sqrt{\log p}) \\
&= O(\sqrt{s \log p})
\end{aligned}$$

Next we show that $\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*) \subseteq \mathrm{conv}(\mathcal{T})$. Since $\mathcal{K} = \{\boldsymbol{\theta} \mid |\mathcal{F}(\boldsymbol{\theta})| \leq s,\ \|\boldsymbol{\theta}\|_2 = 1\}$ and $\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*) = \mathrm{cone}\left\{\mathbf{v} \,\middle|\, \mathbf{v} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*,\ \hat{\boldsymbol{\theta}} \in \mathcal{K}\right\} \bigcap \mathbb{S}^{p-1}$ by definition, we have $|\mathcal{F}(\mathbf{v})| \leq 2s$ for any $\mathbf{v} \in \mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*)$. Suppose $|\mathcal{F}(\mathbf{v})| = t \leq 2s$ and $\mathcal{F}(\mathbf{v}) = \{i_1, i_2, \ldots, i_t\}$. For simplicity, we also let $i_0 = 0$ and $i_{t+1} = p$. Then any $\mathbf{v} \in \mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*)$ can be written as a convex combination of $t+2$ points in $\mathcal{T}$. To see this, we rewrite $\mathbf{v}$ as

$$\mathbf{v} = \sum_{r=0}^{t} \mathbf{v}_{i_r+1:i_{r+1}} = \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \cdot \frac{\sqrt{t+1}\mathbf{v}_{i_r+1:i_{r+1}}}{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2} + \left(1 - \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}}\right) \cdot \mathbf{0}\,, \tag{S.17}$$

where $\mathbf{v}_{i_r+1:i_{r+1}}$ is obtained from $\mathbf{v}$ by keeping the entries from index i_r+1 to i_{r+1} while zeroing out the rest. Let $\mathbf{u}_{i_r+1:i_{r+1}} = \frac{\sqrt{t+1}\mathbf{v}_{i_r+1:i_{r+1}}}{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}$, and we have

$$\|\mathbf{u}_{i_r+1:i_{r+1}}\|_2 = \sqrt{t+1} \leq \sqrt{2s+1} \implies \mathbf{u}_{i_r+1:i_{r+1}} \in \mathcal{T}_{i_r+1:i_{r+1}} \subseteq \mathcal{T}\,.$$

It follows from $\|\mathbf{v}\|_2 = 1$ that

$$\sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \leq \frac{\sqrt{(t+1)\sum_{r=0}^{t}\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2^2}}{\sqrt{t+1}} = 1 \implies 1 - \sum_{r=0}^{t} \frac{\|\mathbf{v}_{i_r+1:i_{r+1}}\|_2}{\sqrt{t+1}} \geq 0$$

Hence (S.17) is indeed a convex combination of $t+2$ points in $\mathcal{T}$, which implies $\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*) \subseteq \mathrm{conv}(\mathcal{T})$. Finally, by the properties of Gaussian width, we conclude that

$$w(\mathcal{A}_\mathcal{K}(\boldsymbol{\theta}^*)) \leq w(\mathrm{conv}(\mathcal{T})) = w(\mathcal{T}) \leq O(\sqrt{s\log p})$$

■

8. Proof of Lemma 2

Statement of Lemma 2: *Define the U-statistic*

$$U_{n,m}(h) = \frac{(n-m)!}{n!} \sum_{\substack{1 \leq i_1,\ldots,i_m \leq n \\ i_1 \neq i_2 \neq \ldots \neq i_m}} h\left(\mathbf{z}_{i_1}, \ldots, \mathbf{z}_{i_m}\right) \tag{S.18}$$

with order m and kernel $h : \mathbb{R}^{d\times m} \mapsto \mathbb{R}$ based on n independent copies of random vector $\mathbf{z} \in \mathbb{R}^d$, denoted by $\mathbf{z}_1, \cdots, \mathbf{z}_n$. If $h(\cdot, \ldots, \cdot)$ is sub-Gaussian with $\|h\|_{\psi_2} \leq \kappa$, then the following inequality holds for $U_{n,m}(h)$ with any $\delta > 0$,

$$\mathbb{P}\left(|U_{n,m}(h) - \mathbb{E}U_{n,m}(h)| > \delta\right) \leq 2\exp\left(-C\left\lfloor\frac{n}{m}\right\rfloor \cdot \frac{\delta^2}{\kappa^2}\right)\,, \tag{S.19}$$

in which C is an absolute constant.

Proof: Our proof is based on Hoeffding's decomposition for U-statistics. For simplicity, we use U as shorthand for $U_{n,m}(h)$. Given a permutation π of $\{1, \ldots, n\}$, define

$$W_\pi = \frac{1}{\left\lfloor\frac{n}{m}\right\rfloor} \sum_{k=0}^{\lfloor\frac{n}{m}\rfloor - 1} h\left(\mathbf{z}_{\pi_{mk+1}}, \ldots, \mathbf{z}_{\pi_{m(k+1)}}\right)\,,$$

The U-statistic can be rewritten as $U = \frac{1}{n!}\sum_\pi W_\pi$, and the summation is over all possible permutations of $\{1, \ldots, n\}$. As no copy of $\mathbf{z}$ appears more than twice in a single W_π, W_π is an average of $\lfloor\frac{n}{m}\rfloor$ independent sub-Gaussian random variables. Hence the ψ_2-norm of its centered version satisfies $\|W_\pi - \mathbb{E}W_\pi\|_{\psi_2} \leq c\kappa/\sqrt{\lfloor\frac{n}{m}\rfloor}$. Using Chernoff technique, we have for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(U - \mathbb{E}U > \delta\right) &\leq e^{-t\delta} \cdot \mathbb{E}\left[\exp(t(U - \mathbb{E}U))\right] \\
&= e^{-t\delta} \cdot \mathbb{E}\left[\exp\left(\frac{t}{n!}\sum_\pi (W_\pi - \mathbb{E}U)\right)\right] \\
&\leq e^{-t\delta} \cdot \mathbb{E}\left[\frac{1}{n!}\sum_\pi \exp\left(t(W_\pi - \mathbb{E}U)\right)\right] \\
&= e^{-t\delta} \cdot \mathbb{E}\left[\exp\left(t(W_\pi - \mathbb{E}W_\pi)\right)\right] \\
&\leq \exp\left(-t\delta + ct^2 \cdot \frac{\kappa^2}{\lfloor\frac{n}{m}\rfloor}\right)\,,
\end{aligned} \tag{S.20}$$

where the second inequality is obtained via Jensen's inequality and the last one follows the moment generating function bound for centered sub-Gaussian random variable. Choosing $t = \lfloor\frac{n}{m}\rfloor \delta/2c\kappa^2$ to minimize right-hand side of (S.20), we obtain

$$\mathbb{P}\left(U - \mathbb{E}U > \delta\right) \leq \exp\left(-C\left\lfloor\frac{n}{m}\right\rfloor \cdot \frac{\delta^2}{\kappa^2}\right)\,,$$

where $C = 1/2c$. To complete the proof, we just need to repeat the argument above for $\mathbb{P}\left(U - \mathbb{E}U < -\delta\right)$. ■

References

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Chen, S. and Banerjee, A. Structured estimation with atomic norms: General bounds and applications. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

Maurer, A., Pontil, M., and Romera-Paredes, B. An Inequality with Applications to Structured Sparsity and Multitask Dictionary Learning. In *Conference on Learning Theory (COLT)*, 2014.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for the analysis of regularized M-estimators. *Statistical Science*, 27(4):538–557, 2012.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G. (eds.), *Compressed Sensing*, chapter 5, pp. 210–268. Cambridge University Press, 2012.