Supplementary Material to
"Robustness and Tractability for Non-convex M-estimators"

Ruizhi Zhang[1], Yajun Mei[2], Jianjun Shi[2], Huan Xu[3]

[1]*University of Nebraska-Lincoln,* [2]*Georgia Institute of Technology*

[3]*Alibaba Inc.*

Lemma 1. *Under Assumption 1, for any $\pi > 0$, there exists a constant $C_\pi = C_0(C_h \vee \log(r\tau/\pi) \vee 1)$, where C_0 is a universal constant, C_h is a constant depending on $\gamma, r, \tau, \psi(z), h(z)$ but independent of π, p, n, δ and g, such that for any $\delta \geq 0$, the following hold:*

(a) *The sample gradient converges uniformly to the population gradient in Euclidean norm, i.e., if $n \geq C_\pi p \log n$, we have*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(0,r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau \sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi. \tag{1}$$

(b) *The sample Hessian converges uniformly to the population Hessian in operator norm, i.e., if $n \geq C_\pi p \log n$, we have*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(0,r)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \tau^2 \sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi. \tag{2}$$

Proof of Lemma 1: In order to prove the uniform convergency theorem, it is suffice to verify assumption 1, 2 and 3 in Mei et al. (2018). Specifically, first, we will verify that the directional gradient of the population risk is sub-Gaussian (Assumption 1 in Mei et al. (2018)). Note the directional gradient of the population risk is given by $\langle \nabla \rho(Y - \langle X, \theta \rangle), \nu \rangle = \psi(Y - \langle X, \theta \rangle)\langle X, \nu \rangle$.

Since $|\psi(Y - \langle X, \theta\rangle)| \leq L_\psi$, and $\langle X, \nu\rangle$ is mean zero and τ^2-sub-Gaussian by our assumption 1, due to Lemma 1 in Mei et al. (2018), there exists a universal constant C_1, such that $\langle \nabla\rho(Y - \langle X, \theta\rangle), \nu\rangle$ is $C_1 L_\psi \tau^2-$sub-Gaussian. Second, we will verify that the directional Hessian of the loss is sub-exponential (Assumption 2 in Mei et al. (2018)). The directional Hessian of the loss gives $\langle \nabla^2\rho(Y - \langle X, \theta\rangle)\nu, \nu\rangle = \psi'(Y - \langle X, \theta\rangle)\langle X, \nu\rangle^2$. Since $|\psi'(Y - \langle X, \theta\rangle)| \leq L_\psi$, by Lemma 1 in Mei et al. (2018), $\langle \nabla^2\rho(Y - \langle X, \theta\rangle)\nu, \nu\rangle$ is $C_2\tau^2$-sub-exponential. Third, let $H = ||\nabla^2 R(\theta_0)||_{op}$ and $J^* = \mathbf{E}\left[\sup_{\theta_1 \neq \theta_2} \frac{||(\psi'(Y - \langle X, \theta_1\rangle) - \psi'(Y - \langle X, \theta_2\rangle))xx^T||_{op}}{||\theta_1 - \theta_2||_2}\right]$. Then, we can show $H \leq L_\psi \tau^2$ and $J^* \leq L_\psi (p\tau^2)^{3/2}$. Therefore, there exists a constant C_h such that $H \leq \tau^2 p^{C_h}$ and $J^* \leq \tau^3 p^{C_h}$, which verifies the assumption 3 in Mei et al. (2018). Therefore, the uniform convergency of gradient and Hessian in theorem 1 in Mei et al. (2018) holds for our gross error model. $\square$

Proof of Theorem 1: Part (a): It is suffice to show that $\langle \theta - \theta_0, \nabla R(\theta)\rangle > 0$ for all $||\theta - \theta_0||_2 > \eta_0$. Note by Assumption 1(d), we have $h(z) = \int_{-\infty}^{+\infty} \psi(z + \epsilon) f_0(\epsilon) d\epsilon > 0$ as $z > 0$ and $h'(0) > 0$. Define $H(s) := \inf_{0 \leq z \leq s} \frac{h(z)}{z}$, it is easy to see that $H(s) > 0$ for all $s > 0$. Then, we have

$$\begin{aligned}
\langle \theta - \theta_0, \nabla R(\theta) \rangle &= \mathbf{E}\left[\mathbf{E}[\psi(z+\epsilon)z | z = \langle \theta_0 - \theta, X \rangle]\right] \\
&= (1-\delta)\mathbf{E}[h(\langle \theta - \theta_0, X \rangle)\langle \theta - \theta_0, X \rangle] + \delta \mathbf{E}\left[\mathbf{E}_g(\psi(z+\epsilon)z | z = \langle \theta_0 - \theta, X \rangle)\right] \\
&\geq (1-\delta)H(s)\mathbf{E}[\langle \theta - \theta_0, X \rangle^2 I_{(|\langle \theta - \theta_0, X \rangle| \leq s)}] - \delta L_\psi \mathbf{E}|\langle \theta_0 - \theta, X \rangle| \\
&= (1-\delta)H(s)\mathbf{E}[\langle \theta - \theta_0, X \rangle^2 - \langle \theta - \theta_0, X \rangle^2 I_{(|\langle \theta - \theta_0, X \rangle| > s)}] - \delta L_\psi \mathbf{E}|\langle \theta - \theta_0, X \rangle| \\
&\geq (1-\delta)H(s)\Big[\mathbf{E}[\langle \theta - \theta_0, X \rangle^2] - \big(\mathbf{E}[\langle \theta - \theta_0, X \rangle^4] \cdot \mathbf{P}(|\langle \theta - \theta_0, X \rangle| > s)\big)^{1/2}\Big] \\
&\quad -\delta L_\psi (\mathbf{E}|\langle \theta - \theta_0, X \rangle|^2)^{1/2} \\
&\overset{\text{(i)}}{\geq} (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{c_2 \mathbf{P}(|\langle \theta - \theta_0, X \rangle| > s)}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\overset{\text{(ii)}}{\geq} (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{\frac{c_2 \mathbf{E}(|\langle \theta - \theta_0, X \rangle|^4)}{s^4}}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \sqrt{\frac{c_2 \cdot c_2 \tau^4 ||\theta - \theta_0||_2^4}{s^4}}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \frac{c_2 \tau^2 ||\theta - \theta_0||_2^2}{s^2}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau \\
&\geq (1-\delta)H(s)||\theta - \theta_0||_2^2 \tau^2 \left(\gamma - \frac{16 c_2 \tau^2 r^2}{9 s^2}\right) - \delta L_\psi ||\theta - \theta_0||_2 \tau.
\end{aligned}$$

Here (i) holds from the fact that if X has mean zero and is τ^2-sub-Gaussian, then for all $u \in \mathbb{R}^p$,

$$\begin{aligned}
\mathbf{E}|\langle u, X \rangle|^2 &\leq ||u||_2^2 \tau^2, \\
\mathbf{E}|\langle u, X \rangle|^4 &\leq c_2 ||u||_2^4 \tau^4,
\end{aligned}$$

where c_2 is a constant (Boucheron et al., 2013). (ii) holds from Chebyshev's inequality. Thus, a choice of $\tilde{s} = \frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}}$ will ensure that

$$\langle \theta - \theta_0, \nabla R(\theta) \rangle \geq (1-\delta)\frac{3}{4}H\left(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}}\right)||\theta - \theta_0||_2^2 \tau^2 \gamma - \delta L_\psi ||\theta - \theta_0||_2 \tau, \tag{3}$$

which is greater than 0 when

$$||\theta - \theta_0||_2 > \frac{\delta L_\psi}{(1-\delta)\frac{3}{4}H\left(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}}\right)\tau\gamma} := \eta_0. \tag{4}$$

Therefore, there are no stationary point outside of the ball $B_2^p(\theta_0, \eta_0)$.

Part(b): We first look at the minimum eigenvalue of the Hessian $\nabla^2 R(\theta)$ at $\theta = \theta_0$. For any $u \in \mathbb{R}^p, ||u||_2 = 1$,

$$\begin{aligned}
\langle u, \nabla^2 R(\theta_0) u \rangle &= (1-\delta)\mathbf{E}_{f_0}[\psi'(\epsilon)\langle X, u\rangle^2] + \delta \mathbf{E}_g[\psi'(\epsilon)\langle X, u\rangle^2] \\
&= (1-\delta)\mathbf{E}_{f_0}[\psi'(\epsilon)]\mathbf{E}[\langle X, u\rangle^2] + \delta \mathbf{E}_g[\psi'(\epsilon)\langle X, u\rangle^2] \\
&\geq (1-\delta)h'(0)\gamma\tau^2 - \delta L_\psi \tau^2.
\end{aligned}$$

Therefore, we have the minimum eigenvalue of $\nabla^2 R(\theta_0)$ is greater than 0 as long as $\delta < \frac{h'(0)\gamma}{h'(0)\gamma + L_\psi}$.

Similarly, we can get $\langle u, \nabla^2 R(\theta_0) u \rangle \leq (1-\delta)h'(0)\gamma\tau^2 + \delta L_\psi \tau^2$.

Then we look at the operator norm of $\nabla^2 R(\theta) - \nabla^2 R(\theta_0)$. For any $u \in \mathbb{R}^p, ||u||_2 = 1$,

$$\begin{aligned}
|\langle u, (\nabla^2 R(\theta) - \nabla^2 R(\theta_0))u\rangle| &= |\mathbf{E}[(\psi'(\langle X, \theta_0 - \theta\rangle + \epsilon) - \psi'(\epsilon))\langle X, u\rangle^2]| \\
&= |\mathbf{E}[\psi''(\xi)\langle X, \theta_0 - \theta\rangle\langle X, u\rangle^2]| \\
&\leq \mathbf{E}|\psi''(\xi)|\mathbf{E}|\langle X, \theta_0 - \theta\rangle\langle X, u\rangle^2| \\
&\leq L_\psi\{\mathbf{E}[\langle X, \theta_0 - \theta\rangle^2]\mathbf{E}[\langle X, u\rangle^4]\}^{1/2} \\
&\leq L_\psi(||\theta_0 - \theta||_2^2 \tau^2 c_2 \tau^4)^{1/2} \\
&= L_\psi \sqrt{c_2}||\theta_0 - \theta||_2 \tau^3.
\end{aligned}$$

Hence, taking

$$||\theta - \theta_0||_2 \leq \eta_1 := \frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2\sqrt{c_2}\tau L_\psi} \tag{5}$$

guarantees that $(\nabla^2 R(\theta) - \nabla^2 R(\theta_0))_{op} \leq \frac{(1-\delta)h'(0)\gamma\tau^2 - \delta L_\psi \tau^2}{2}$. Therefore, for all $\theta \in B_2^p(\theta_0, \eta_1)$, we have

$$\begin{aligned}
\lambda_{\min}(\nabla^2 R(\theta)) \geq \kappa &:= \frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2}\tau^2, &(6)\\
\lambda_{\max}(\nabla^2 R(\theta)) \leq \kappa' &:= [\frac{3}{2}(1-\delta)h'(0)\gamma + \frac{1}{2}\delta L_\psi]\tau^2, &(7)
\end{aligned}$$

which yields there is at most one minimizer of $R(\theta)$ in the ball $B_2^p(\theta_0, \eta_1)$, as long as $\delta < \frac{h'(0)\gamma}{h'(0)\gamma + L_\psi}$.

Part (c): Note $R(\theta)$ is a continuous function on $B_2^p(r)$. Thus there exists a global minimizer, denoted by θ^*. Since we have shown that there is no stationary points outside the ball $B_2^p(\theta_0, \eta_0)$, θ^* should be in the ball $B_2^p(\theta_0, \eta_0)$. Therefore, as long as $\eta_1 > \eta_0$, i.e.,

$$\frac{(1-\delta)h'(0)\gamma - \delta L_\psi}{2\sqrt{c_2}\tau L_\psi} > \frac{\delta L_\psi}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}, \tag{8}$$

there exists and only exists a unique stationary point of $R(\theta)$, which is also the global optimum θ^*. □

Proof of Theorem 2 Based on Lemma 1, there exists a constant C_π such that as n is large enough when $n \geq C_\pi p \log n$,

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau\sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi \tag{9}$$

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \tau^2\sqrt{\frac{C_\pi p \log n}{n}}\right) \geq 1 - \pi. \tag{10}$$

Let $\epsilon_0 = h'(0)H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\gamma^2\tau/(4\sqrt{c_2}L_\psi)$, which is a constant that does not depend on π, δ. Thus, if n is further large such that $\tau\sqrt{\frac{C_\pi p \log n}{n}} \leq \epsilon_0$ and $\tau^2\sqrt{\frac{C_\pi p \log n}{n}} \leq \kappa/2$, i.e., $n \geq Cp\log n$, where $C = \max\{C_\pi, \tau^2 C_\pi/\epsilon_0^2, 4\tau^4 C_\pi/\kappa^2\}$, we have

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 \leq \tau\sqrt{\frac{C_\pi p \log n}{n}} \leq \epsilon_0\right) \geq 1 - \pi \tag{11}$$

$$\mathbf{P}\left(\sup_{\theta \in B^p(0,r)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \leq \tau^2\sqrt{\frac{C_\pi p \log n}{n}} \leq \kappa/2\right) \geq 1 - \pi. \tag{12}$$

Part (a): Note

$$\begin{aligned}
\langle \theta - \theta_0, \nabla \widehat{R}_n(\theta) \rangle &\geq \langle \theta - \theta_0, \nabla R(\theta) \rangle - ||\nabla \hat{R}_n(\theta) - \nabla R(\theta)||_2 ||\theta - \theta_0||_2 \qquad (13)\\
&\geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2 \tau^2\gamma - (\tau\delta L_\psi + \epsilon_0)||\theta - \theta_0||_2 \qquad (14)
\end{aligned}$$

which is greater than 0 when

$$||\theta - \theta_0||_2 > \frac{\tau\delta L_\psi + \epsilon_0}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau^2\gamma} = \eta_0 + \frac{1}{1-\delta}\zeta, \tag{15}$$

where $\zeta := \frac{h'(0)\gamma}{3\sqrt{c_2}\tau L_\psi}$ is a constant does not depend on δ. Therefore, there are no stationary points outside of the ball $B_2^p(\theta_0, \eta_0 + \frac{1}{1-\delta}\zeta)$.

Part (b): For the least eigenvalue of the empirical Hessian in $B_2^p(\theta_0, \eta_1)$, we have

$$\begin{aligned} \inf_{||\theta-\theta_0||_2 \le \eta_1} \lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) &\ge \inf_{||\theta-\theta_0||_2 \le \eta_1} \lambda_{\min}(\nabla^2 R(\theta)) - \sup_{\theta \in B^p(0,\eta_1)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \\ &\ge \kappa - \kappa/2 = \kappa/2 > 0. \end{aligned} \tag{16}$$

This lead to the conclusion that, $\widehat{R}_n(\theta)$ is strong convex inside the ball $B_2^p(\theta_0, \eta_1)$.

For the largest eigenvalue of the empirical Hessian in $B_2^p(\theta_0, \eta_1)$, we have

$$\begin{aligned} \sup_{||\theta-\theta_0||_2 \le \eta_1} \lambda_{\max}(\nabla^2 \widehat{R}_n(\theta)) &\le \sup_{||\theta-\theta_0||_2 \le \eta_1} \lambda_{\max}(\nabla^2 R(\theta)) + \sup_{\theta \in B^p(0,\eta_1)} ||\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta)||_{op} \\ &\le \kappa' + \kappa/2 < 2\kappa', \end{aligned} \tag{17}$$

where κ' is defined in (6) .

Part(c): When $\eta_0 + \frac{1}{1-\delta}\zeta < \eta_1$, by strong convexity of $\widehat{R}_n(\theta)$ in $B_2^p(\theta_0, \eta_1)$, there exists a unique local minimizer, which is in $B_2^p(\theta_0, \eta_0 + \frac{1}{1-\delta}\zeta)$. We denote the unique local minimizer as $\widehat{\theta}_n$.

By Theorem 1, there is a unique stationary point of the population risk function $R(\theta)$ in the ball $B_2^p(\theta_0, \eta_0)$. Suppose θ^* is the unique stationary point of $R(\theta)$. By Taylor expansion of $\widehat{R}_n(\theta)$ at the point θ^*, there exists a $\tilde{\theta}$ in $B_2^p(\theta_0, \eta_0 + \frac{1}{1-\delta}\zeta)$, such that

$$\widehat{R}_n(\widehat{\theta}_n) = \widehat{R}_n(\theta^*) + \langle \widehat{\theta}_n - \theta^*, \nabla \widehat{R}_n(\theta^*) \rangle + \frac{1}{2}(\widehat{\theta}_n - \theta^*)' \nabla^2 \widehat{R}_n(\tilde{\theta})(\widehat{\theta}_n - \theta^*) \le \widehat{R}_n(\theta^*). \tag{18}$$

Since by equation (16), the least eigenvalue of $\nabla^2 \widehat{R}_n(\tilde{\theta})$ is greater than $\kappa/2$, which lead to

$$\frac{\kappa}{4}||\widehat{\theta}_n - \theta^*||_2^2 \le \langle \theta^* - \widehat{\theta}_n, \nabla \widehat{R}_n(\theta^*) \rangle \le ||\theta^* - \widehat{\theta}_n||_2 ||\nabla \widehat{R}_n(\theta^*)||_2, \tag{19}$$

which yield

$$||\widehat{\theta}_n - \theta^*||_2 \leq \frac{4}{\kappa}||\nabla \widehat{R}_n(\theta^*)||_2. \tag{20}$$

By Theorem 1, $||\theta_0 - \theta^*||_2 < \eta_0$, combined with equation (20) and the uniform convergency theorem in Lemma 1 yield

$$||\widehat{\theta}_n - \theta_0||_2 \leq \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{Cp\log n}{n}}. \tag{21}$$

Part(d): Let $\theta_n(k)$ be the $k-$th iterate of gradient descent defined by

$$\theta_n(k+1) = \theta_n(k) - h\nabla \widehat{R}_n(\theta_n(k))$$

First, we assume that we initialize at $\theta_n(0) \notin B_2^p(\theta_0, \eta_1)$ and all the iterates up to $\theta_n(k)$ are outside the ball $B_2^p(\theta_0, \eta_1)$. We will show that the gradient descent will converge exponentially to the ball $B_2^p(\theta_0, \eta_1)$. Note

$$\|\theta_n(k+1) - \theta_0\|_2^2 - \|\theta_n(k) - \theta_0\|_2^2 = -2h\langle \nabla \widehat{R}_n(\theta_n(k)), \theta_n(k) - \theta_0\rangle + h^2\|\nabla \widehat{R}_n(\theta_n(k))\|_2^2 \tag{22}$$

The lower bound of the inner product term can be derived by (13).

$$\begin{aligned}\langle \nabla \widehat{R}_n(\theta_n(k)), \theta_n(k) - \theta_0\rangle &\geq \delta L_\psi \tau\left[\frac{1}{\eta_0}\|\theta_n(k) - \theta_0\|_2^2 - 2\|\theta_n(k) - \theta_0\|_2\right] \\ &\geq \frac{(\eta_1 - 2\eta_0)}{\eta_0\eta_1}\|\theta_n(k) - \theta_0\|_2^2 \delta L_\psi \tau, \end{aligned} \tag{23}$$

where the last inequality holds by the fact that $\theta_n(k) \notin B_2^p(\theta_0, \eta_1)$. Moreover, since $\|\nabla R(\theta)\|_2 \leq 2L_\psi\tau$, under the event (11), with probability $1-\pi$, $\|\nabla \widehat{R}_n(\theta)\|_2 \leq (2+\delta)L_\psi\tau$. Thus, by (22) and (23),

$$\|\theta_n(k+1) - \theta_0\|_2^2 \leq \|\theta_n(k) - \theta_0\|_2^2\left[1 - 2h\frac{(\eta_1 - 2\eta_0)}{\eta_0\eta_1}\delta L_\psi\tau\right] + h^2(2+\delta)^2L_\psi^2\tau^2. \tag{24}$$

Thus, by choosing $h \leq h_{\max,1} := \frac{\eta_1(\eta_1 - 2\eta_0)\delta}{\eta_0(2+\delta)^2L_\psi\tau}$, for all $\theta_n(k) \notin B_2^p(\theta_0, \eta_1)$, we have

$$\begin{aligned}\|\theta_n(k+1) - \theta_0\|_2^2 &\leq \|\theta_n(k) - \theta_0\|_2^2\left[1 - 2h\frac{(\eta_1 - 2\eta_0)}{\eta_0\eta_1}\delta L_\psi\tau\right] + h^2(2+\delta)^2L_\psi^2\tau^2 \\ &\leq \|\theta_n(k) - \theta_0\|_2^2\left[1 - h\frac{(\eta_1 - 2\eta_0)}{\eta_0\eta_1}\delta L_\psi\tau\right].\end{aligned}$$

Define $r_1 = 1 - h\frac{(\eta_1-2\eta_0)}{\eta_0\eta_1}\delta L_\psi \tau < 1$. We have the following chain of inequalities

$$\begin{aligned}
\|\theta_n(k) - \widehat{\theta}_n\|_2 &\leq \|\theta_n(k) - \theta_0\|_2 + \|\widehat{\theta}_n - \theta_0\|_2 \leq \|\theta_n(k) - \theta_0\|_2 + 2\eta_0 \\
&\leq 2\|\theta_n(k) - \theta_0\|_2 \leq 2r_1^{k/2}\|\theta_n(0) - \theta_0\|_2 \leq 2r_1^{k/2}(\|\theta_n(0) - \widehat{\theta}_n\|_2 + \|\widehat{\theta}_n - \theta_0\|_2) \\
&\leq 4r_1^{k/2}(\|\theta_n(0) - \widehat{\theta}_n\|_2, \qquad (25)
\end{aligned}$$

which implies the exponential convergence of the gradient descent outside $B_2^p(\theta_0, \eta_1)$.

Next, we will establish an exponential convergence inside $B_2^p(\theta_0, \eta_1)$. By (16), we have

$$\inf_{||\theta-\theta_0||_2 \leq \eta_1} \lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) \geq \kappa/2, \qquad \sup_{||\theta-\theta_0||_2 \leq \eta_1} \lambda_{\max}(\nabla^2 \widehat{R}_n(\theta)) \leq 2\kappa'.$$

Thus, $\widehat{R}_n(\theta)$ is $\kappa/2$-strongly convex in $B_2^p(\theta_0, \eta_1)$. By standard convex optimization results, if we start from a point inside $B_2^p(\theta_0, \eta_1)$, and take $h \leq h_{\max,2} := 1/(2\kappa')$, we have

$$\|\theta_n(k) - \widehat{\theta}_n\|_2 \leq 2\sqrt{\frac{\kappa'}{\kappa}}(1 - \frac{1}{2}\kappa h)^{k/2}\|\theta_n(0) - \widehat{\theta}_n\|_2.$$

Combined with the result (25) in the first step yields for any initialization $\theta_n(0) \in B_2^p(0, r)$, running gradient descent gives

$$\|\theta_n(k) - \widehat{\theta}_n\|_2 \leq 4\sqrt{\frac{\kappa'}{\kappa}} s^k \|\theta_n(0) - \widehat{\theta}_n\|_2, \qquad (26)$$

where $s = \max\{\sqrt{1 - h\frac{(\eta_1-2\eta_0)}{\eta_0\eta_1}\delta L_\psi\tau}, \sqrt{1 - \frac{1}{2}\kappa h}\}$, and the step size h satisfies $h \leq h_{\max} = \min\{h_{\max,1}, h_{\max,2}\} = \min\{\frac{\eta_1(\eta_1-2\eta_0)\delta}{\eta_0(2+\delta)^2 L_\psi \tau}, 1/(2\kappa')\}$.

□

Lemma 2. *Under assumption 1 and 2, there exist constants C_1, C_2, T_0, L_0 that depend on $r, \tau, \pi, \delta, L_\psi$, but independent of n, p, and g, such that the following hold:*

a *The sample directional gradient converges uniformly to the population directional gradient, along the direction $(\theta - \theta_0)$.*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(r)\setminus\{0\}} \frac{|\langle \nabla R_n(\theta) - \nabla R(\theta), \theta - \theta_0\rangle|}{||\theta - \theta_0||_1} \leq (T_0 + L_0\tau)\sqrt{\frac{C_1 \log(np)}{n}}\right)$$
$$\geq 1 - \pi.$$

b *As* $n \geq C_2 s_0 \log(np)$, *we have*

$$\mathbf{P}\left(\sup_{\theta \in B_2^p(r)\cap B_2^p(s_0), \nu \in B_2^p(1)\cap B_0^p(s_0)} |\langle \nu, \left(\nabla^2 R_n(\theta) - \nabla^2 R(\theta)\right)\nu\rangle| \leq \tau^2 \sqrt{\frac{C_2 s_0 \log(np)}{n}}\right)$$
$$\geq 1-\pi.$$

Proof of Lemma 2: From the Theorem 3 in Mei et al. (2018), the uniform convergency theorem of our Lemma 2 holds if Assumption 4, 5 in Mei et al. (2018) hold under the contaminated model with outliers. Here we will show under our assumption 1 and 2, there exist constants T_0 and L_0 such that

a For all $\theta \in B_2^p(r)$, $Y \in \mathbb{R}, X \in \mathbb{R}^p$, $||\nabla_\theta \rho(Y - \langle X, \theta\rangle)||_\infty \leq T_0 M$

b There exist functions $h_1 : \mathbb{R} \times \mathbb{R}^{p+1} \to \mathbb{R}$, and $h_2 : \mathbb{R}^{p+1} \to \mathbb{R}^p$, such that

$$\langle \nabla_\theta \rho(Y - \langle X, \theta\rangle), \theta - \theta_0\rangle = h_1(\langle \theta - \theta_0, h_2(Y, X)\rangle), Y, X). \tag{27}$$

In addition, $h_1(t, Y, X)$ is $L_0 M$- Lipschitz to its first argument t, $h_1(0, Y, X) = 0$, and $h_2(Y, X)$ is mean-zero and τ^2-sub-Gaussian.

Part (a). The gradient of the loss is

$$\nabla_\theta \rho(Y - \langle X, \theta\rangle) = -\psi(Y - \langle X, \theta\rangle)X. \tag{28}$$

By assumption 1, we have $| - \psi(Y - \langle X, \theta\rangle)| \leq L_\psi$. By assumption 2, we have $||X||_\infty \leq M\tau$. Therefore, (a) is satisfied with parameter $T_0 = L_\psi \tau$.

Part (b). Note

$$\langle \nabla_\theta \rho(Y - \langle X, \theta\rangle), \theta - \theta_0\rangle = -\psi(Y - \langle X, \theta\rangle)\langle X, \theta - \theta_0\rangle. \tag{29}$$

We take $h_2(Y, X) = X$, $t = \langle X, \theta - \theta_0\rangle$ and $h_1(t, Y, X) = -\psi(Y - t - \langle X, \theta_0\rangle)t$. Clearly, we have $h_1(0, Y, X) = 0$ and $h_2(Y, X)$ is mean 0 and τ^2-sub-Gaussian. Furthermore, note $|t| \leq 2rM\tau$, we have

$$
\begin{aligned}
|\frac{\partial}{\partial t} h_1(t, Y, X)| &= |\psi'(Y - t - \langle X, \theta_0\rangle)t - \psi(Y - t - \langle X, \theta_0\rangle)| && (30)\\
&\leq 2ML_\psi r\tau + L_\psi && (31)\\
&\leq (2L_\psi r\tau + L_\psi)M. && (32)
\end{aligned}
$$

Therefore, $h_1(t, X, Y)$ is at most $(2L_\psi r\tau + L_\psi)M$-Lipschitz in its first argument t. By part (a) and part (b), we can see assumption 4, 5 are satisfied under the gross error model, which prove the uniform convergency theorem in our Lemma 2. □

Proof of Theorem 3: We decompose the proof into four technical lemmas. First, in Lemma 3, we prove there cannot be any stationary points of the regularized empirical risk $\hat{L}_n$ in (10) outside the region $\mathbb{A}$, which is a cone with $\mathbb{A} = \{\theta_0 + \Delta : ||\Delta_{S_0^c}||_1 \leq 3||\Delta_{S_0}||_1\}$. Then in Lemma 4, we show there cannot be any stationary points outside the region $B_2^p(\theta_0, r_s)$ where r_s is the statistical radius which is not less than η_0 in Theorem 1. In Lemma 5, we argue that all stationary points should have support size less or equal to $cs_0 \log p$. Finally, in Lemma 6, we show there cannot be two stationary points in $B_2^p(\theta_0, \eta_1) \cap \mathbb{A}$. Note $\hat{L}_n(\theta)$ is a continuous function, which indicates the existence of the global minimizer. Therefore, we can conclude there is and only is one unique stationary point of the regularized empirical risk $\hat{L}_n$ as long as $r_s < \eta_1$.

To start with those lemmas, we define the subgradient of $\hat{L}_n$ at θ as:

$$\partial\hat{L}_n(\theta) = \{\nabla R_n(\theta) + \lambda_n \nu : \nu \in \partial||\theta||_1\}. \tag{33}$$

Therefore, the optimality condition implies that θ is a stationary point of $\hat{L}_n$ if and only if $\mathbf{0} \in \partial\hat{L}_n(\theta)$. To simplify notations, all constants in the following lemmas are dependent on $(\rho, L_\psi, \tau^2, r, \gamma, \pi)$ but independent on δ, s_0, n, p, M. □

Lemma 3. Let $S_0 = supp(\theta_0)$ and $s_0 = |S_0|$. Define a cone $\mathbb{A} = \{\theta_0 + \Delta : ||\Delta_{S_0^c}||_1 \leq 3||\Delta_{S_0}||_1\} \subseteq \mathbb{R}^p$. For any $\pi > 0$, there exist constants C_π, such that letting $\lambda_n \geq 2C_\pi M\sqrt{\frac{\log p}{n}} + 2\delta L_\psi \tau$, with probability at least $1-\pi$, $\hat{L}_n(\theta)$ has no stationary points in $B_2^p(0,r) \cap \mathbb{A}^c$:

$$\langle z(\theta), \theta - \theta_0\rangle > 0, \quad \forall \theta \in B_2^p(0,r) \cap \mathbb{A}^c, z(\theta) \in \partial \hat{L}_n(\theta) \tag{34}$$

Proof of Lemma 3: For any $z(\theta) \in \partial \hat{L}_n(\theta)$, it can be written as $z(\theta) = \nabla \hat{R}_n(\theta) + \lambda_n \nu(\theta)$, where $\nu(\theta) \in \partial||\theta||_1$. Therefore, we have

$$\langle z(\theta), \theta - \theta_0\rangle = \langle \nabla R(\theta), \theta - \theta_0\rangle + \langle \nabla \hat{R}_n(\theta) - \nabla R(\theta), \theta - \theta_0\rangle + \lambda_n \langle \nu(\theta), \theta - \theta_0\rangle \tag{35}$$

Note by (3) we have

$$\langle \theta - \theta_0, \nabla R(\theta)\rangle \geq (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta-\theta_0||_2\tau. \tag{36}$$

By Lemma 2, for any $\pi > 0$, there exists a constant C_π such that

$$\mathbf{P}(\sup_{0<||\theta||_2<r} \frac{|\langle \nabla \hat{R}_n(\theta) - \nabla R(\theta), \theta - \theta_0\rangle|}{||\theta - \theta_0||_1} \leq C_\pi M\sqrt{\frac{\log p}{n}}) > 1 - \pi. \tag{37}$$

Letting $\Delta = \theta - \theta_0$, we have

$$\langle \nu(\theta), \theta - \theta_0\rangle = \langle \nu(\theta)_{S_0^c}, \Delta_{S_0^c}\rangle + \langle \nu(\theta)_{S_0}, \Delta_{S_0}\rangle \geq ||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1 \tag{38}$$

Plugging (36),(37),(38) into (35) yields

$$\begin{aligned}\langle z(\theta), \theta - \theta_0\rangle \quad &\geq \quad (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta-\theta_0||_2\tau &(39)\\ &- \quad C_\pi M\sqrt{\frac{\log p}{n}}(||\Delta_{S_0^c}||_1 + ||\Delta_{S_0}||_1) + \lambda_n(||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1). &(40)\end{aligned}$$

Let $\lambda_n \geq 2C_\pi M\sqrt{\frac{\log p}{n}} + C_2$, we have

$$\begin{aligned}\langle z(\theta), \theta - \theta_0\rangle \quad &\geq \quad (1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2\tau^2\gamma - \delta L_\psi||\theta-\theta_0||_2\tau \\ &+ \quad C_\pi M\sqrt{\frac{\log p}{n}}(||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1) + C_2(||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1). &(41)\end{aligned}$$

Next, we will find the lower bound of $||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1$ under the constraint of $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 \geq 0$. Note

$$\begin{aligned} ||\Delta_{S_0^c}||_1 - ||\Delta_{S_0}||_1 &= \frac{1}{2}(||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 + ||\Delta_{S_0^c}||_1 + ||\Delta_{S_0}||_1) \\ &= \frac{1}{2}(||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 + \|\Delta\|_1) \\ &\geq \frac{1}{2}\|\Delta\|_1 \geq \frac{1}{2}\|\Delta\|_2. \end{aligned} \tag{42}$$

Combined with (41), setting $C_2 \geq 2\delta L_\psi \tau$ yield $C_2/2 \geq \delta L_\psi \tau$, which implies $\langle z(\theta), \theta - \theta_0 \rangle > 0$, as long as $\theta \in \mathbb{A}^c$, i.e., $||\Delta_{S_0^c}||_1 - 3||\Delta_{S_0}||_1 > 0$. □

Lemma 4. For any $\pi > 0$, $\theta \in \mathbb{A}, z(\theta) \in \partial \hat{L}_n(\theta)$, there exist constants C_0, C_1 such that with probability at least $1 - \pi$,

$$\langle z(\theta), \theta - \theta_0 \rangle > 0 \tag{43}$$

as long as $||\theta - \theta_0||_2 > r_s$, where

$$r_s = \frac{\delta}{1-\delta} C_0 + \frac{4\sqrt{s_0}}{1-\delta}(M\sqrt{\frac{\log p}{n}} + \lambda_n)C_1. \tag{44}$$

Proof of Lemma 4: Since for any $\theta \in \mathbb{A}$, we have $||\theta - \theta_0||_1 \leq 4\sqrt{s_0}||\theta - \theta_0||_2$. Combining with (35) yields

$$\begin{aligned} \langle z(\theta), \theta - \theta_0 \rangle &\geq \langle \nabla R(\theta), \theta - \theta_0 \rangle - C_\pi M \sqrt{\frac{\log p}{n}} ||\theta - \theta_0||_1 - \lambda_n ||\theta - \theta_1||_1 & (45) \\ &\geq (1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})||\theta - \theta_0||_2^2 \tau^2 \gamma - \delta L_\psi ||\theta - \theta_0||_2 \tau & (46) \\ &\quad -(C_\pi M \sqrt{\frac{\log p}{n}} + \lambda_n) 4\sqrt{s_0}||\theta - \theta_0||_2, & (47) \end{aligned}$$

which is greater than 0 as long as

$$||\theta - \theta_0||_2 \geq \frac{\delta L_\psi + (C_\pi M \sqrt{\frac{\log p}{n}} + \lambda_n) 4\sqrt{s_0}}{(1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} := r_s. \tag{48}$$

Taking $C_0 = \frac{L_\psi}{\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}$ and $C_1 = \frac{\max(1, C_\pi)}{\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}$ give the result of r_s in equation (44). □

Lemma 5. If $\delta \leq 1/2$, for any π, there exist constants C_0, C_1 such that letting $\lambda_n \geq 2L_\psi \tau(C_0\sqrt{\frac{\log p}{n}} + \delta)$, with probability at least $(1-\pi)$, any stationary points of $\hat{L}_n(\theta)$ in $B_2^p(\theta_0, r_s) \cap \mathbb{A}$ has support size $|S(\hat{\theta})| \leq C_1 s_0 \log p$.

Proof of Lemma 5: Let $\hat{\theta} \in B_2^p(\theta_0, r_s) \cap \mathbb{A}$ be a stationary point of $\hat{L}_n(\theta)$ in (10). Then we have

$$\nabla R_n(\hat{\theta}) + \lambda_n \nu(\hat{\theta}) = 0, \tag{49}$$

where $\nu(\hat{\theta}) \in ||\hat{\theta}||_1$. Thus, we have

$$\left(\nabla R_n(\hat{\theta})\right)_j = \pm\lambda_n, \quad \forall j \in S(\hat{\theta}) \tag{50}$$

Note $|\psi(y_i - \langle x_i, \theta_0\rangle)| \leq L_\psi$ and $\langle x_i, e_j\rangle$ is τ^2-subgaussian with mean 0. Then there exists an absolute constant c_0 such that $\psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle$ is $c_0 L_\psi^2 \tau^2$-subgaussian, see Lemma 1(d) in Mei et al. (2018). Thus we have $\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle$ is $c_0 L_\psi^2 \tau^2/n$-subgaussian with mean $\langle \nabla R(\theta_0), e_j\rangle$. Moreover, note $|\langle \nabla R(\theta_0), e_j\rangle| = |\delta \mathbf{E}_g \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| \leq \delta L_\psi \mathbf{E}|\langle x_i, e_j\rangle| \leq \delta L_\psi \tau$, we have for any $t > 0$,

$$\begin{aligned} & \mathbf{P}(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| \geq t + \delta L_\psi \tau) \\ \leq \quad & \mathbf{P}(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle - \langle \nabla R(\theta_0), e_j\rangle| \geq t) \leq 2\exp(-\frac{t^2 n}{2c_0 L_\psi^2 \tau^2}). \end{aligned} \tag{51}$$

Thus, we can get

$$\begin{aligned} \mathbf{P}\left(||\nabla R_n(\theta_0)||_\infty > t + \delta L_\psi \tau\right) \quad & \leq \quad p \max_{1\leq j\leq p} \mathbf{P}\left(|\frac{1}{n}\sum_{i=1}^n \psi(y_i - \langle x_i, \theta_0\rangle)\langle x_i, e_j\rangle| > t + \delta L_\psi \tau\right) \\ & \leq \quad 2p\exp(-\frac{t^2 n}{2c_0 L_\psi^2 \tau^2}). \end{aligned} \tag{52}$$

Thus, a choice of $t = L_\psi \tau\sqrt{\frac{2c_0(\log p + \log 6/\pi)}{n}}$ and $C = \sqrt{c_0 \log 6/\pi}$ will guarantee that

$$\mathbf{P}\left(||\nabla \hat{R}_n(\theta_0)||_\infty > L_\psi \tau(C\sqrt{\frac{\log p}{n}} + \delta)\right) \quad \leq \pi/3 \tag{53}$$

Let $\lambda_n \geq 2L_\psi\tau(C\sqrt{\frac{\log p}{n}} + \delta)$, we have the event $(||\nabla R_n(\theta_0)||_\infty < \lambda_n/2)$ happens with the probability at least $1-\pi/3$. Under this event, combing with (50) yields

$$\lambda_n/2 \leq \left|\left(\nabla R_n(\theta_0) - \nabla R_n(\hat{\theta})\right)_j\right|, \quad \forall j \in S(\hat{\theta}). \tag{54}$$

Squaring and summing over $j \in S(\hat{\theta})$, we have

$$\begin{aligned}
\lambda_n^2|S(\hat{\theta})| &\leq 4\left\|\left(\nabla \hat{R}_n(\theta_0) - \nabla \hat{R}_n(\hat{\theta})\right)_{S(\hat{\theta})}\right\|_2^2 && (55)\\
&= 4\left\|\left(\frac{1}{n}\sum_{i=1}^n \Big(\psi(y_i - \langle\theta_0, x_i\rangle) - \psi(y_i - \langle\hat{\theta}, x_i\rangle)\Big)\, x_i\right)_{S(\hat{\theta})}\right\|_2^2 && (56)\\
&= 4\left\|\left(\frac{1}{n}\sum_{i=1}^n \left(\psi'(y_i - \langle\beta_i, x_i\rangle)\right)\langle\theta_0 - \hat{\theta}, x_i\rangle x_i\right)_{S(\hat{\theta})}\right\|_2^2 && (57)\\
&\leq 4L_\psi^2\left\|\left(\frac{1}{n}\sum_{i=1}^n \langle\theta_0 - \hat{\theta}, x_i\rangle x_i\right)_{S(\hat{\theta})}\right\|_2^2 && (58)
\end{aligned}$$

where β_i are located on the line between θ_0 and $\hat{\theta}$ obtained by intermediate value theorem. Moreover, by Minkowski inequality and Cauchy-Schwarz inequality yield

$$\begin{aligned}
\left\|\left(\frac{1}{n}\sum_{i=1}^n \langle\theta_0 - \hat{\theta}, x_i\rangle x_i\right)_{S(\hat{\theta})}\right\|_2 &\leq \frac{1}{n}\sum_{i=1}^n |\langle\theta_0 - \hat{\theta}, x_i\rangle| \left\|(x_i)_{S(\hat{\theta})}\right\|_2\\
&\leq \frac{1}{n}\left((\sum_{i=1}^n |\langle\theta_0 - \hat{\theta}, x_i\rangle|^2)(\sum_{i=1}^n \|\,(x_i)_{S(\hat{\theta})}\,\|_2^2)\right)^{1/2} && (59)
\end{aligned}$$

Due to the restricted smoothness property of the sub-Gaussian random variables Mei et al. (2018), there exists a constant c_1 depending on π such that with probability at least $1-\pi/3$, as $n \geq c_1 s_0 \log p$, we have

$$\sup_{\theta\in\mathbb{A}} \frac{\frac{1}{n}(\sum_{i=1}^n |\langle\theta_0 - \theta, x_i\rangle|^2)}{||\theta - \theta_0||_2^2} \leq 3\tau^2. \tag{60}$$

Therefore, with probability at least $1-\pi/3$, we have

$$\sup_{\theta\in\mathbb{A}\cap B^p(\theta_0, r_s)} \frac{1}{n}(\sum_{i=1}^n |\langle\theta_0 - \hat{\theta}, x_i\rangle|^2) \leq 3\tau^2 \sup_{\theta\in\mathbb{A}\cap B^p(\theta_0, r_s)} ||\theta - \theta_0||_2^2 \leq 3\tau^2 r_s^2. \tag{61}$$

Moreover, by Lemma 13 in Mei et al. (2018), for any π, there exists constant c_2 depending on π such that

$$\mathbf{P}(\frac{1}{n}\sum_{i=1}^{n} \| (x_i)_{S(\hat{\theta})} \|_2^2 > c_2\tau^2 \log p) \leq \pi/3. \tag{62}$$

By (53),(61),(62), as well as (59), at least $1-\pi$,

$$\begin{aligned} \lambda_n^2 |S(\hat{\theta})| &\leq 4L_\psi^2 3\tau^2 r_s^2 c_2 \tau^2 \log p \\ &= C r_s^2 \log p \end{aligned}$$

By equation (44) we have

$$r_s^2 \leq C_0(\frac{\delta}{1-\delta})^2 + \frac{s_0}{(1-\delta)^2}(M^2 \frac{\log p}{n} + \lambda_n^2)C_1 \tag{63}$$

Taking $\lambda_n \geq 2L_\psi \tau(C\sqrt{\frac{\log p}{n}} + \delta)$ gives us

$$\begin{aligned} |S(\hat{\theta})| &\leq (C_4 \frac{s_0}{(1-\delta)^2} + s_0 C_5) \log p \\ &= C s_0 \log p \end{aligned}$$

□

Lemma 6. For any positive constants C_0 and π, letting $r_0 = C_0 s_0 \log p$, there exist constant C_1 such that when $n \geq C_1 s_0 \log^2 p$,

$$\mathbf{P}(\sup_{\theta \in B_2^p(\theta_0, r) \cap B_0^p(0, r_0)} \sup_{\nu \in B_2^p(0,1) \cap B_0^p(0, r_0)} \langle \nu, (\nabla^2 \hat{R}_n(\theta) - \nabla^2 R(\theta))\nu \rangle \leq \kappa/2) \geq 1 - \pi. \tag{64}$$

Moreover, the regularized empirical risk $\hat{L}_n(\theta)$ in (10) cannot have two stationary points in the region $B_2^p(\theta_0, \eta_1) \cap B_0^p(0, r_0/2)$.

Proof of Lemma 6: According to (6), we have

$$\inf_{\theta \in B_2^p(\theta_0, \eta_1)} \lambda_{\min}(\nabla^2 R(\theta)) \geq \kappa. \tag{65}$$

By Lemma 2, there exists constant C such that when $n \geq Cs_0 \log^2 p$,

$$\mathbf{P}\left(\inf_{\theta \in B_2^p(\theta_0,\eta_1)\cap B_0^p(0,r_0)} \inf_{\nu \in B_2^p(0,1)\cap B_0^p(0,r_0)} \langle \nu, (\nabla^2 \hat{R}_n(\theta))\nu\rangle \geq \kappa/2\right) \leq \pi. \tag{66}$$

Suppose θ_1, θ_2 are two distinct stationary points of $\hat{L}_n(\theta)$ in $B_2^p(\theta_0,\eta_1) \cap B_0^p(0, r_0/2)$. Define $u = \frac{\theta_2-\theta_1}{||\theta_1-\theta_2||_2}$. Since θ_1 and θ_2 are $r_0/2$-sparse, u is r_0 sparse, as well as $\theta_1 + tu$ for any $t \in \mathbb{R}$. Therefore,

$$\begin{aligned}
\langle \nabla \hat{R}_n(\theta_2), u\rangle &= \langle \nabla \hat{R}_n(\theta_1), u\rangle + \int_0^{||\theta_1-\theta_2||_2} \langle u, \nabla^2 \hat{R}_n(\theta_1+tu)u\rangle dt \\
&\geq \langle \nabla \hat{R}_n(\theta_1), u\rangle + \frac{\kappa}{2}||\theta_2-\theta_1||_2.
\end{aligned} \tag{67}$$

Note the regularization term $\lambda_n||\theta||_1$ is convex, we have for any subgradients $\nu(\theta_1) \in \partial||\theta_1||_1$, $\nu(\theta_2) \in \partial||\theta_2||_1$,

$$\lambda_n \langle \nu(\theta_2), u\rangle \geq \lambda_n \langle \nu(\theta_1), u\rangle. \tag{68}$$

Adding (67) with (68) gives

$$\langle \nabla \hat{R}_n(\theta_2) + \lambda_n \nu(\theta_2), u\rangle \geq \langle \nabla \hat{R}_n(\theta_1) + \lambda_n \nu(\theta_1), u\rangle + \frac{\kappa}{2}||\theta_2-\theta_1||_2, \tag{69}$$

which is contradict with the assumption that θ_1 and θ_2 are two distinct stationary points of $\hat{L}_n(\theta)$. □

Proof of Theorem 3. Now we are ready to prove Theorem 3. By Lemma 3 and Lemma 4, as $n \geq Cs_0 \log p$, letting $\lambda_n \geq 2CM\sqrt{\frac{\log p}{n}} + 2\delta L_\psi \tau$, all stationary points of $L_n(\theta)$ are in $B_2^p(\theta_0, r_s) \cap \mathbb{A} \cap B_0^p(C_1 s_0 \log p)$, where r_s is defined in (44), $\mathbb{A}$ is the cone defined in Lemma 3. This proves Theorem 3(a). Moreover, by Lemma 5, Lemma 6, as $n \geq C_2 s_0 \log^2 p$, $\hat{L}_n(\theta)$ cannot have two distinct stationary points in $B_2^p(\theta_0,\eta_1) \cap \mathbb{A} \cap B_0^p(C_1 s_0 \log p)$. Thus, as long as $\eta_1 \geq r_s$, there is only one unique stationary point of the regularized empirical risk function $\hat{L}_n(\theta)$, which is the corresponding regularized M-estimator of (10). This proves Theorem 3 (b). □

Proof of Corollary 1: Note the Welsch's loss function is defined by $\rho_\alpha(t) = \frac{1-e^{-\alpha t^2/2}}{\alpha}$. The corresponding score function is $\psi_\alpha(t) = \rho'_\alpha(t) = te^{-\alpha t^2/2}$. Moreover, we can get $\psi'_\alpha(t) = e^{-\alpha t^2/2}(1-\alpha t^2)$ and $\psi''_\alpha(t) = e^{-\alpha t^2/2}\alpha(\alpha t^2 - 3)$. Note for any $\alpha > 0$, all of $\psi_\alpha(t)$, $\psi'_\alpha(t)$ and $\psi''_\alpha(t)$ are bounded.

$$
\begin{aligned}
|\psi_\alpha(t)| &\leq \sqrt{\frac{e}{\alpha}} \\
|\psi'_\alpha(t)| &\leq \max\{1, 2e^{-1.5}\} = 1 \\
|\psi''_\alpha(t)| &\leq \max\{e^{-(3+\sqrt{6})/2}\sqrt{(18+6\sqrt{6})\alpha}, e^{-(3-\sqrt{6})/2}\sqrt{(18-6\sqrt{6})\alpha}\} \leq 1.5\sqrt{\alpha}.
\end{aligned}
$$

Therefore, the Assumption 1 is satisfied. It is suffice to find the explicit expression of η_0 and η_1 in equation (4) and (5). In order to have an accurate expression, we will use the individual bound of $\psi_\alpha(t), \psi'_\alpha(t), \psi''_\alpha(t)$ instead of the universal bound L_ψ. Specifically, according to Assumption 4, x_i is τ^2-sub-Gaussian, $c_2 = 3, \gamma = 1/3$. Thus, we can calculate $h(z) = \int_{-\infty}^{+\infty} \psi_\alpha(z+\epsilon) f_0(\epsilon) d\epsilon = \frac{z}{(1+\alpha\sigma^2)^{3/2}} e^{-\frac{\alpha z^2}{2(1+\alpha\sigma^2)}}$ and $H(s) = \frac{1}{(1+\alpha\sigma^2)^{3/2}} e^{-\frac{\alpha s^2}{2(1+\alpha\sigma^2)}}$. Similarly, we can calculate $h'(0) = E_{f_0}\psi'_\alpha(\epsilon) = \frac{1}{(1+\alpha\sigma^2)^{3/2}}$. By (15), we have $\zeta = \frac{h'(0)\gamma}{3\sqrt{c_2}\tau L_\psi} = \frac{1}{13.5\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}$.

By equation (4) in the proof of Theorem 1 yields

$$
\begin{aligned}
\eta_0(\delta, \alpha) &= \frac{\delta L_\psi}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} \\
&= \frac{\delta}{1-\delta}\sqrt{\frac{e}{\alpha}}\frac{4(1+\alpha\sigma^2)^{3/2}}{\tau} e^{\frac{32\alpha r^2\tau^2}{3(1+\alpha\sigma^2)}}
\end{aligned}
$$

Note $|\psi'_\alpha(t)| \leq 1, |\psi''_\alpha(t)| \leq 1.5\sqrt{\alpha}$, by equation (5) in the proof of Theorem 1 yields

$$
\begin{aligned}
\eta_1(\delta, \alpha) &= \frac{(1-\delta)h'(0)\gamma - \delta}{2\sqrt{3} \times 1.5\sqrt{\alpha}\tau} \\
&= \frac{1}{9\sqrt{3\alpha}(1+\alpha\sigma^2)^{3/2}\tau}\left[1 - \delta(1 + 3(1+\alpha\sigma^2)^{3/2})\right].
\end{aligned}
$$

□

Proof of Corollary 2: Tukey's bisquare loss function is defined by

$$\rho_\alpha(t) = \begin{cases} \frac{1}{6}\alpha^2\left[1-(1-(t/\alpha)^2)^3\right], & \text{if } |t| \le \alpha \\ 0, & \text{if } |t| > \alpha. \end{cases} \tag{70}$$

The corresponding score function is

$$\psi_\alpha(t) = \rho'_\alpha(t) = \begin{cases} t(1-t^2/\alpha^2)^2, & \text{if } |t| \le \alpha \\ 0, & \text{if } |t| > \alpha. \end{cases} \tag{71}$$

Moreover, for any $\alpha > 0$, all of $\psi(t)$, $\psi'(t)$ and $\psi''(t)$ are bounded. Specifically, we have $|\psi_\alpha(t)| < \alpha$, $|\psi'(t)| < 4, |\psi''(t)| = 1/\alpha$. Therefore, the assumptions in Theorem 1 and Theorem 2 are satisfied. It is suffice to find the explicit expression of η_0 and η_1 in equation (4) and (5). Specifically, according to Assumption 4, x_i is τ^2-sub-Gaussian, $c_2 = 3, \gamma = 1/3$. Thus, we can calculate

$$\begin{aligned} h(z) &= \int_{-\infty}^{+\infty} \psi_\alpha(z+\epsilon) f_0(\epsilon) d\epsilon = \int_0^\alpha \psi_\alpha(t)[f_0(t-z) - f_0(t+z)]dt \\ &\ge \frac{2}{\sqrt{2\pi}\sigma^3} \int_0^\alpha e^{-\frac{(t+z)^2}{2\sigma^2}} tz\psi_\alpha(t)dt \ge \frac{2}{\sqrt{2\pi}\sigma^3} e^{-\frac{(z+\alpha)^2}{2\sigma^2}} z \int_0^\alpha t\psi_\alpha(t)dt \\ &> \frac{1}{7\sqrt{2\pi}\sigma^3} e^{-\frac{(z^2+\alpha^2)}{\sigma^2}} z\alpha^3 \end{aligned}$$

Thus, $H(s) > \frac{1}{7\sqrt{2\pi}\sigma^3} e^{-\alpha^2/\sigma^2}\alpha^3 e^{-s^2/\sigma^2}$. By equation (4) in the proof of Theorem 1 yields

$$\begin{aligned} \eta_0(\delta,\alpha) &= \frac{\delta L_\psi}{(1-\delta)\frac{3}{4} H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} \\ &< \frac{\delta}{1-\delta} \frac{28\sqrt{2\pi}}{\tau\sigma^3\alpha^2} e^{\frac{\alpha^2+64\tau^2 r^2}{\sigma^2}} \end{aligned}$$

Similarly, we can calculate

$$\begin{aligned} h'(0) &= E_{f_0}\psi'_\alpha(\epsilon) = \frac{2}{\alpha^4}\int_0^\alpha (\alpha-t)(\alpha+t)(\alpha^2-5t^2) f_0(t)dt \\ &= 2\alpha \int_0^1 (1-t)(1+t)(1-5t^2) f_0(\alpha t)dt \\ &:= M(\alpha,\sigma). \end{aligned}$$

For fixed $\sigma > 0, \alpha > 0$, we have $M(\alpha, \sigma) > 0$. Note $|\psi'_\alpha(t)| \leq 4, |\psi''_\alpha(t)| \leq 1/\alpha$, by equation (5) in the proof of Theorem 1 yields

$$\eta_1(\delta, \alpha) = \frac{(1-\delta)M(\alpha, \sigma)\tau^2 - 4\delta}{2\sqrt{3}\tau}\alpha \tag{72}$$

Moreover, according to equation (48) in the proof of Theorem 3, we have with high probability, all stationary points of the empirical risk function $\hat{L}_n(\theta)$ in (10) are inside the ball $B_2^p(\theta_0, r_s)$, where

$$r_s = \eta_0 + \frac{12C_\pi \tau \sqrt{(s_0 \log p)/n} + 2\tau\delta L_\psi}{(1-\delta)\frac{3}{4}H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma} \tag{73}$$

$$= (1+2\tau)\eta_0 + \frac{16C_\pi \tau \sqrt{(s_0 \log p)/n}}{(1-\delta)H(\frac{8\tau r}{3}\sqrt{\frac{c_2}{\gamma}})\tau\gamma}. \tag{74}$$

Therefore, as $n \gg s_0 \log p$, we have $r_s \approx (1+2\tau)\eta_0$, which completes the proof. □

References

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Mei, S., Y. Bai, A. Montanari, et al. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics 46*(6A), 2747–2774.