Asymptotic Normality of Robust Risk Minimizers

University of Southern California

Supplementary Material

This documents contains additional technical details and numerical results that were omitted from the main text of the manuscript.

S1 Auxiliary results.

In the exposition below, we will often refer to the lemmas stated in section 3.1 of the main document.

1.3 Existence of solutions.

In this section, we discuss simple sufficient conditions for existence of the estimator $\widehat{\theta}_{n,k}$ defined in display (1.5) of the main document.

Proposition 1. *Assume that $\Theta \subset \mathbb{R}^d$ is compact and that $\ell(\theta, x)$ is continuous with respect to the first variable for P-almost all x. Moreover, let ρ be a convex function such that $\rho''(x) > 0$ for all $x \in \mathbb{R}$. Then $\widehat{\theta}_{n,k}$ exists.*

Proof. It suffices to show that $\widehat{L}(\theta, \theta')$ is continuous. The existence claim then easily follows as $\widehat{L}(\theta,\theta')$ must be uniformly continuous on $\Theta \times \Theta$ due to compactness, which in turn implies, via a standard argument, continuity of the function $\theta \mapsto \max_{\theta' \in \Theta} \widehat{L}(\theta,\theta')$, hence the existence of $\widehat{\theta}_{n,k}$, again in view of compactness. To establish the continuity of $\theta \mapsto \max_{\theta' \in \Theta} \widehat{L}(\theta,\theta')$ when $\widehat{L}(\theta,\theta')$ is uniformly continuous, note that for any $\theta \in \Theta$ and any $\varepsilon > 0$, $\widehat{L}(\theta,\theta') - \varepsilon \leqslant \widehat{L}(\tilde{\theta},\theta') \leqslant \widehat{L}(\theta,\theta') + \varepsilon$ for all $\theta' \in \Theta$ as long as $\|\tilde{\theta}-\theta\| \leqslant \delta(\varepsilon)$. It easily implies that $\max_{\theta' \in \Theta} \widehat{L}(\theta,\theta') - \varepsilon \leqslant \max_{\theta' \in \Theta} \widehat{L}(\tilde{\theta},\theta') \leqslant \max_{\theta' \in \Theta} \widehat{L}(\theta,\theta') + \varepsilon$, and the conclusion follows.

All that remains is to establish the continuity of $\widehat{L}(\theta,\theta')$. To this end, fix $\varepsilon > 0$ and let

$$R(z;\theta,\theta') = \frac{1}{k}\sum_{j=1}^{k} \rho\left(\sqrt{n}\,\frac{\bar{L}_j(\theta) - \bar{L}_j(\theta') - z}{\Delta_n}\right).$$

Since $R'(z;\theta,\theta')$ is strictly increasing in z, there exist $z_+(\varepsilon)$ and $z_-(\varepsilon)$ such that $R'(z_+(\varepsilon);\theta,\theta') = \varepsilon$ and $R'(z_-(\varepsilon);\theta,\theta') = -\varepsilon$. In particular, $\widehat{L}(\theta,\theta') \in (z_-(\varepsilon), z_+(\varepsilon))$. As $R''(\widehat{L}(\theta,\theta');\theta,\theta') > 0$ in view of the assumption $\rho'' > 0$, $|z_+(\varepsilon) - z_-(\varepsilon)| \to 0$ as $\varepsilon \to 0$. Since $\bar{L}_j(\theta) - \bar{L}_j(\theta')$ is continuous in θ, θ' by assumption, R is continuous in θ, θ' as well, hence $\left|R(z_+(\varepsilon);\tilde{\theta},\tilde{\theta}') - R(z_+(\varepsilon);\theta,\theta')\right| < \varepsilon$ and $\left|R(z_-(\varepsilon);\tilde{\theta},\tilde{\theta}') - R(z_-(\varepsilon);\theta,\theta')\right| < \varepsilon$ whenever $\|(\theta,\theta') - (\tilde{\theta},\tilde{\theta}')\| \leqslant \delta(\varepsilon)$ for some $\delta(\varepsilon)$ small enough. In this case, we see that the inequalities $R(z_+(\varepsilon);\tilde{\theta},\tilde{\theta}') > 0$ and $R(z_-(\varepsilon);\tilde{\theta},\tilde{\theta}') < 0$ hold, hence $\widehat{L}(\tilde{\theta},\tilde{\theta}') \in (z_-(\varepsilon), z_+(\varepsilon))$, implying that $\left|\widehat{L}(\tilde{\theta},\tilde{\theta}') - \widehat{L}(\theta,\theta')\right| \leqslant$

$|z_+(\varepsilon) - z_-(\varepsilon)| \to 0$ as $\varepsilon \to 0$, yielding the desired conclusion. □

We remark that elsewhere in this work, we choose ρ with the second derivative vanishing outside of a neighborhood of 0. However, $R''(\widehat{L}(\theta,\theta');\theta,\theta') > 0$ holds with high probability uniformly over $\theta, \theta' \in \Theta$ when Θ is compact and the class $\{\ell(\theta,\cdot),\ \theta \in \Theta\}$ satisfies the assumptions made. We sketch the steps needed to show this fact; all the required tools have already been established in the paper. First, note that in view of Lemma A.1 and the triangle inequality, $\sup_{\theta,\theta' \in \Theta} \left|\widehat{L}(\theta,\theta') - L(\theta,\theta')\right| = O_P(n^{-1/2})$ as $n, k \to \infty$ with high probability, hence

$$\inf_{\theta,\theta'} R''(\widehat{L}(\theta,\theta');\theta,\theta') \geqslant \inf_{\theta,\theta',|z|\leqslant D/\sqrt{n}} R''(L(\theta,\theta')+z;\theta,\theta')$$

for a large constant D, again with high probability. Next, the relation

$$\frac{1}{n} \sup_{\theta,\theta',|z|\leqslant D/\sqrt{n}} |R''(L(\theta,\theta')+z;\theta,\theta') - \mathbb{E}R''(L(\theta,\theta')+z;\theta,\theta')| = o_P(1)$$

as $n, k \to \infty$ follows from an argument identical to the one used to prove Lemma A.2 and Lemma 2. Finally,

$$\mathbb{E}\rho''\left(\sqrt{n}\,\frac{\bar{L}_j(\theta) - \bar{L}_j(\theta') - L(\theta,\theta') - z}{\Delta_n}\right) = \mathbb{E}\rho''\left(\frac{Z(\theta,\theta') - z\sqrt{n}}{\Delta_n}\right) + o(1)$$

in view of Lemma 1, where $Z(\theta,\theta')$ is a centered and normally distributed random variable with variance $\sigma^2(\theta,\theta')$. As $\rho''(x) \geqslant I\{|x| \leqslant 1\}$, we see that $\inf_{\theta,\theta',|z|\leqslant D/\sqrt{n}} \mathbb{E}\rho''\left(\frac{Z(\theta,\theta')-z\sqrt{n}}{\Delta_n}\right) > 0$, yielding the result.

S2 Proof of Theorem 1 (main text).

2.4 Preliminaries.

Let us recall some basic facts and existing results required in the proof. Given a metric space (T, ρ), the covering number $N(T, \rho, \varepsilon)$ is defined as the smallest $N \in \mathbb{N}$ such that there exists a subset $F \subseteq T$ of cardinality N with the property that for all $z \in T$, $\rho(z, F) \leqslant \varepsilon$. Let $\{Y(t),\ t \in T\}$ be a stochastic process indexed by T. We will say that it has sub-Gaussian increments with respect to some metric ρ if for all $t_1, t_2 \in \mathbb{T}$ and $s \in \mathbb{R}$,

$$\mathbb{E}e^{s(Y_{t_1} - Y_{t_2})} \leqslant e^{\frac{s^2 \rho^2(t_1, t_2)}{2}}.$$

Theorem (Dudley's entropy bound)**.** Let $\{Y(t),\ t \in T\}$ be a centered stochastic process with sub-Gaussian increments. Then the following inequality holds:

$$\mathbb{E} \sup_{t \in T} |Y(t) - Y(t_0)| \leqslant 12 \int\limits_0^{D(T)} \sqrt{\log N(T, \rho, \varepsilon)} d\varepsilon,$$

where $D(T)$ is the diameter of the space T with respect to ρ.

Proof. See the book by Talagrand (2005). □

The following bound allows one to control the error $\left|\widehat{L}(\theta, \theta_0) - L(\theta, \theta_0)\right|$ uniformly over compact subsets $\Theta' \subseteq \Theta$. Recall the adversarial contamination framework introduced in section 1, and define

$$\widetilde{\Delta} := \max\left(\Delta_n, \sup_{\theta\in\Theta'} \sigma(\theta,\theta_0)\right).$$

Lemma A.1. Let $\mathcal{L} = \{\ell(\theta,\cdot),\ \theta\in\Theta\}$ be a class of functions mapping S to $\mathbb{R}$, and assume that $\sup_{\theta\in\Theta'} \mathbb{E}\left|\ell(\theta,X) - \ell(\theta_0,X) - L(\theta,\theta_0)\right|^{2+\tau} < \infty$ for some $\tau \in [0,1]$. Then there exist absolute constants c, $C > 0$ and a function $g_\tau(x,\theta)$ satisfying $g_\tau(x,\theta) \overset{x\to\infty}{=} \begin{cases} o(1), & \tau = 0, \\ O(1), & \tau > 0 \end{cases}$ such that for all $s > 0$, n and k satisfying

$$\frac{s}{\sqrt{k}\Delta_n}\mathbb{E}\sup_{\theta\in\Theta'}\frac{1}{\sqrt{N}}\left|\sum_{j=1}^N \left(\ell(\theta,X_j) - \ell(\theta_0,X_j) - L(\theta,\theta_0)\right)\right| \\ + \sup_{\theta\in\Theta'}\left[g_\tau(n,\theta)\frac{\mathbb{E}\left|\ell(\theta,X) - \ell(\theta_0,X) - L(\theta,\theta_0)\right|^{2+\tau}}{\Delta_n^{2+\tau} n^{\tau/2}}\right] + \frac{\mathcal{O}}{k} \leqslant c,$$

the following inequality holds with probability at least $1 - \frac{1}{s}$:

$$\sup_{\theta\in\Theta'}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \\ \leqslant C\Bigg[s\cdot\frac{\widetilde{\Delta}}{\Delta_n}\mathbb{E}\sup_{\theta\in\Theta'}\left|\frac{1}{N}\sum_{j=1}^N\Big(\ell(\theta,X_j) - \ell(\theta_0,X_j) - L(\theta,\theta_0)\Big)\right| \\ + \widetilde{\Delta}\left(\frac{1}{\sqrt{n}}\frac{\mathcal{O}}{k} + \frac{1}{\sqrt{n}}\sup_{\theta\in\Theta'}\left[g_\tau(n,\theta)\frac{\mathbb{E}\left|\ell(\theta,X) - \ell(\theta_0,X) - L(\theta,\theta_0)\right|^{2+\tau}}{\Delta_n^{2+\tau} n^{\tau/2}}\right]\right)\Bigg].$$

We will only use the bound of the lemma with $\tau = 0$. The proof of this bound is similar to the argument behind Theorem 3.1 in (Minsker, 2019b); for the readers' convenience, we present the details in section 2.4 below.

For the illustration purposes, assume that $\mathcal{O} = 0$, whence the result above implies that as long as

$$\mathbb{E} \sup_{\theta\in\Theta'} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |\ell(\theta, X_j)) - \ell(\theta_0, X_j) - L(\theta,\theta_0)| = O(1)$$

and $\sigma(\Theta') \lesssim \Delta_n = O(1)$,

$$\sup_{\theta\in\Theta'} \left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| = O_p\left(N^{-1/2} + n^{-(1+\tau)/2}\Delta_n^{-(2+\tau)}\right).$$

Moreover, if $\mathcal{O} = \kappa N$ and $\Delta_n = O(1)$, then, setting $k \asymp N\kappa^{\frac{2}{2+\tau}}$, we see that

$$\sup_{\theta\in\Theta'} \left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| = O_p\left(N^{-1/2} + \kappa^{\frac{1+\tau}{2+\tau}}\right).$$

Lemma A.2. Assume that $X_1, \ldots, X_n$ are i.i.d. Let $\theta \in \Theta$, and set $\delta_0 := r(\theta)$, where $r(\theta)$ is defined in Assumption 3. Then for all $0 < \delta \leqslant \delta_0$,

$$\begin{aligned}
\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} &\left|\frac{1}{k}\sum_{j=1}^{k} \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right)\right. \\
&\qquad\qquad \left. - \mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta',\theta_0) - L(\theta',\theta_0)\right)\right)\right| \\
&\qquad \leqslant \frac{8}{\Delta_n\sqrt{k}} \mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} \left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(\theta',X_j) - \ell(\theta_0,X_j) - L(\theta',\theta_0)\right)\right|
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\sup_{\|\theta'-\theta\|\leqslant\delta} &\left|\frac{1}{k}\sum_{j=1}^{k} \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right)\right. \\
&\qquad\qquad \left. - \mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta',\theta_0) - L(\theta',\theta_0)\right)\right)\right| \\
&\qquad \leqslant \frac{8s}{\Delta_n\sqrt{k}} \mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} \left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(\theta',X_j) - \ell(\theta_0,X_j) - L(\theta',\theta_0)\right)\right|
\end{aligned}$$

with probability at least $1-\frac{1}{s}$, where $C>0$ is an absolute constant. Moreover, the bound still holds if ρ'' is replaced by ρ''', up to the change in constants.

The proof is given in section 2.4.

Lemma A.3. Let $\{A_n(\theta),\ \theta\in\Theta\}, \{B_n(\theta),\ \theta\in\Theta\subseteq\mathbb{R}^d\}$ be sequences of stochastic processes such that for every $\theta\in\Theta$, the sequences of random variables $\{A_n(\theta)\}_{n\geqslant 1}$ and $\{B_n(\theta)\}_{n\geqslant 1}$ are stochastically bounded, and for any $\varepsilon>0$,

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\delta}|A_n(\theta)-A_n(\theta_0)|\geqslant\varepsilon\right)\to 0 \text{ as } \delta\to 0,$$

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\delta}|B_n(\theta)-B_n(\theta_0)|\geqslant\varepsilon\right)\to 0 \text{ as } \delta\to 0.$$

Then

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\delta}|A_n(\theta)B_n(\theta)-A_n(\theta_0)B_n(\theta_0)|\geqslant\varepsilon\right)\to 0 \text{ as } \delta\to 0.$$

Moreover, if there exists $c>0$ such that

$$\liminf_{n\to\infty}\mathbb{P}(|B_n(\theta_0)|\geqslant c)=1,$$

then the following also holds:

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{A_n(\theta)}{B_n(\theta)}-\frac{A_n(\theta_0)}{B_n(\theta_0)}\right|\geqslant\varepsilon\right)\to 0 \text{ as } \delta\to 0.$$

Proof. The result follows in a straightforward manner from the triangle inequality hence the details are omitted. □

Let us commence the proof of the theorem. To simplify and clarify the notation, we will omit subscript j in most cases and simply write "k, n" instead of "k_j, n_j" to denote the increasing sequences of the number of subgroups and their cardinalities. For every $\theta' \in \Theta$, define

$$\widehat{\theta}(\theta') := \underset{\theta\in\Theta}{\operatorname{argmax}}\, \widehat{L}(\theta', \theta) = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \widehat{L}(\theta, \theta')$$

Above, we assumed that the maximum is attained so that $\widehat{\theta}(\theta')$ is well defined; however, the argument also holds with $\widehat{\theta}(\theta')$ replaced by a near-maximizer. We will set $\widehat{\theta}^{(1)}_{n,k} := \widehat{\theta}_{n,k}$ and $\widehat{\theta}^{(2)}_{n,k} := \widehat{\theta}(\widehat{\theta}^{(1)}_{n,k})$. Observe that $\widehat{L}\left(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}\right) \leqslant \widehat{L}\left(\theta_0, \widehat{\theta}(\theta_0)\right)$, hence whenever $\|\widehat{\theta}^{(j)}_{n,k} - \theta_0\| \leqslant R,\ j = 1, 2$,

$$\begin{aligned}
L(\widehat{\theta}^{(1)}_{n,k}) - L(\widehat{\theta}^{(2)}_{n,k}) &= L(\widehat{\theta}^{(1)}_{n,k}) - L(\widehat{\theta}^{(2)}_{n,k}) \pm \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}) \\
&\leqslant \widehat{L}\left(\theta_0, \widehat{\theta}(\theta_0)\right) + \sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1, \theta_2) - L(\theta_1, \theta_2)\right| \\
&\leqslant L(\theta_0) - L(\widehat{\theta}(\theta_0)) + 2 \sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1, \theta_2) - L(\theta_1, \theta_2)\right| \\
&\leqslant 2 \sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1, \theta_2) - L(\theta_1, \theta_2)\right|,
\end{aligned}$$

where we used the fact that $L(\theta_0) - L(\widehat{\theta}(\theta_0)) \leqslant 0$ in the last step. On the other hand, for any $\varepsilon > 0$,

$$\inf_{\|\theta_1-\theta_0\|\geqslant\varepsilon} \sup_{\theta_2} \left(L(\theta_1) - L(\theta_2)\right) > L(\theta_0) + \delta - L(\theta_0) = \delta$$

where $\delta := \delta(\varepsilon) > 0$ exists in view of Assumption 2. Therefore,

$$\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| \geqslant \varepsilon\right) \leqslant \mathbb{P}\left(\sup_{\|\theta_j - \theta_0\| \leqslant R, j=1,2} \left|\widehat{L}(\theta_1,\theta_2) - L(\theta_1,\theta_2)\right| > \delta/2\right) \\ + \mathbb{P}\left(\left\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\right\| > R \text{ or } \left\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\right\| > R\right).$$

It follows from Lemma A.1 that

$$\sup_{\|\theta_j - \theta_0\| \leqslant R, j=1,2} \left|\widehat{L}(\theta_1,\theta_2) - L(\theta_1,\theta_2)\right| \to 0 \text{ in probability}$$

as long as $\limsup_{k,n\to\infty} \frac{\mathcal{O}(k,n)}{k} \leqslant c$ as $n, k \to \infty$. Indeed, to verify this, it suffices to show that

$$\limsup_{N\to\infty} \mathbb{E} \sup_{\|\theta_j - \theta_0\| \leqslant R, j=1,2} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \left(\ell(\theta_1, X_j) - \ell(\theta_2, X_j) - L(\theta_1,\theta_2)\right)\right| < \infty,$$

which follows from the triangle inequality and the relation

$$\limsup_{N\to\infty} \mathbb{E} \sup_{\|\theta_1 - \theta_0\| \leqslant R} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \left(\ell(\theta_1, X_j) - \ell(\theta_0, X_j) - L(\theta_1,\theta_0)\right)\right| < \infty. \quad (2.1)$$

To establish the latter, we use a well-known argument based on symmetrization inequality and Dudley's entropy integral bound (see section 2.4). Let $\varepsilon_1, \ldots, \varepsilon_N$ be i.i.d. random signs, independent of the data $X_1, \ldots, X_N$. Then symmetrization inequality (van der Vaart and Wellner, 1996) yields that

$$\mathbb{E} \sup_{\theta\in\Theta: \|\theta - \theta_0\| \leqslant R} \frac{1}{\sqrt{N}} \left| \sum_{j=1}^{N} \left(\ell(\theta, X_j) - \ell(\theta_0, X_j) - L(\theta,\theta_0)\right)\right|$$

$$\leqslant 2\mathbb{E} \sup_{\theta\in\Theta:\|\theta-\theta_0\|\leqslant R} \frac{1}{\sqrt{N}}\left|\sum_{j=1}^N \varepsilon_j\left(\ell(\theta,X_j)-\ell(\theta_0,X_j)\right)\right|.$$

Conditionally on $X_1,\ldots,X_N$, the process

$$\ell(\theta,\cdot)\mapsto \frac{1}{\sqrt{N}}\sum_{j=1}^N \varepsilon_j\left(\ell(\theta,X_j)-\ell(\theta_0,X_j)\right)$$

has sub-Gaussian increments with respect to the semi-metric $d_N^2(\theta_1,\theta_2) := \frac{1}{N}\sum_{j=1}^N\left(\ell(\theta_1,X_j)-\ell(\theta_2,X_j)\right)^2$. It follows from compactness of the set $B(\theta_0,R) = \{\theta : \|\theta-\theta_0\|\leqslant R\}$ and Assumption 3 that there exist $\theta_1,\ldots,\theta_{N(R)}$ such that $\bigcup_{j=1}^{N(R)} B(\theta_j,r(\theta_j))\supseteq B(\theta_0,R)$ and

$$|\ell(\theta',x)-\ell(\theta'',x)|\leqslant \mathcal{V}(x;r(\theta_j))\|\theta'-\theta''\|$$

for all $\theta',\theta''\in B(\theta_j,r(\theta_j))$. To cover $B(\theta_0,R)$ by the balls of d_N-radius τ, it suffices to cover each of the $N(R)$ balls $B(\theta_j,r(\theta_j))$. It is easy to see that the latter requires at most $\left(\frac{6r(\theta_j)\|\mathcal{V}(\cdot;r(\theta_j))\|_{L_2(P_N)}}{\tau}\right)^d$ balls of radius τ. Therefore,

$$\log^{1/2} N(B(\theta_0,R),d_N,\tau)\leqslant \log^{1/2}\left(\sum_{j=1}^{N(R)}\left[\left(\frac{6r(\theta_j)\|\mathcal{V}(\cdot;r(\theta_j))\|_{L_2(P_N)}}{\tau}\right)^d\vee 1\right]\right).$$

Note that for any $x_1,\ldots,x_m\geqslant 1$, $\sum_{j=1}^m x_j\leqslant m\prod_{j=1}^m x_j$, or $\log\left(\sum_{j=1}^m x_j\right)\leqslant \log m+\sum_{j=1}^m \log x_j$, so that

$$\log^{1/2}\left(\sum_{j=1}^{N(R)}\left[\left(\frac{6r(\theta_j)\|\mathcal{V}(\cdot;r(\theta_j))\|_{L_2(P_N)}}{\tau}\right)^d\vee 1\right]\right)$$
$$\leqslant \log^{1/2} N(R)+\sum_{j=1}^{N(R)}\sqrt{d}\log_+^{1/2}\left(\frac{6r(\theta_j)\|\mathcal{V}(\cdot;r(\theta_j))\|_{L_2(P_N)}}{\tau}\right),$$

where $\log_+(x) := \max(\log x, 0)$. Moreover, the diameter D_N of the set $B(\theta_0, R)$ is at most $2\sum_{j=1}^{N(R)} r(\theta_j)\|\mathcal{V}(\cdot; r(\theta_j))\|_{L_2(P_N)}$. Therefore,

$$\begin{aligned}
&\int_0^{D_N} \log^{1/2} N(B(\theta_0, R), d_N, \tau) d\tau \\
&\leqslant C\left(D_N \log^{1/2} N(R) + \sqrt{d}\sum_{j=1}^{N(R)} r(\theta_j)\|V(\cdot; r(\theta_j))\|_{L_2(P_N)} \int_0^1 \log^{1/2}(1/\tau) d\tau\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E} \sup_{\theta\in\Theta:\|\theta-\theta_0\|\leqslant R} \frac{1}{\sqrt{N}}\left|\sum_{j=1}^N \varepsilon_j(\ell(\theta, X_j) - \ell(\theta_0, X_j))\right| \\
\leqslant C\log^{1/2}(N(R))\sum_{j=1}^{N(R)} r(\theta_j)\|\mathcal{V}(\cdot; r(\theta_j))\|_{L_2(P)} < \infty.
\end{aligned}$$

It remains to establish that $\mathbb{P}\Big(\Big\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\Big\| > R \text{ or } \Big\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\Big\| > R\Big) \to 0$. To this end, notice that by the definition of $\widehat{\theta}^{(1)}_{n,k}$,

$$\begin{aligned}
0 \leqslant \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}) &\leqslant \widehat{L}\left(\theta_0, \widehat{\theta}(\theta_0)\right) \\
&\leqslant \underbrace{L(\theta_0) - L\left(\widehat{\theta}(\theta_0)\right)}_{\leqslant 0} + \sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta_0, \theta) - L(\theta_0, \theta)\right|
\end{aligned}$$

on the event $\left\{\|\widehat{\theta}(\theta_0) - \theta_0\| \leqslant R\right\}$. It has already been established that

$$\sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta_0, \theta) - L(\theta_0, \theta)\right| \to 0 \text{ in probability.}$$

To show that $\mathbb{P}\Big(\|\widehat{\theta}(\theta_0) - \theta_0\| > R\Big) \to 0$ for R large enough and as $n, k \to \infty$, recall that

$$B(n, R, t) = \mathbb{P}\left(\inf_{\|\theta-\theta_0\|\geqslant R} \frac{1}{n}\sum_{j=1}^n \ell(\theta, X_j) < L(\theta_0) + t\right)$$

and that $\lim_{R\to\infty}\limsup_{n\to\infty} B(n,R,t) = 0$ for some $t > 0$ in view of Assumption 4. As moreover $\frac{1}{n}\sum_{j=1}^n \ell(\theta_0, X_j) \to L(\theta_0)$ in probability, one can choose R_0 and n_0 such that

$$\tilde{B}(n,R,t) = \mathbb{P}\left(\inf_{\|\theta-\theta_0\|\geqslant R} \frac{1}{n}\sum_{j=1}^n \ell(\theta, X_j) - \frac{1}{n}\sum_{j=1}^n \ell(\theta_0, X_j) < t/2\right) < \gamma$$

for all $n \geqslant n_0(\gamma)$ and $R \geqslant R_0(\gamma)$ for any $\gamma > 0$. As

$$\widehat{L}(\theta,\theta_0) = \underset{z\in\mathbb{R}}{\operatorname{argmin}} \sum_{j=1}^k \rho\left(\frac{\sqrt{n}}{\Delta_n}(\bar{L}_j(\theta) - \bar{L}_j(\theta_0) - z)\right),$$

it solves the equation $\sum_{j=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}(\bar{L}_j(\theta) - \bar{L}_j(\theta_0) - \widehat{L}(\theta,\theta_0))\right) = 0$. Assumption 1 implies that $\rho'(x) = \|\rho'\|_\infty$ for $x \geqslant 2$. Therefore, $\widehat{L}(\theta,\theta_0) < t/4$ only if $\bar{L}_j(\theta) - \bar{L}_j(\theta_0) < t/4 + 2\frac{\Delta_n}{\sqrt{n}}$ for $j \in J$ such that $|J| \geqslant k/2$. To see this, suppose that there exists a subset $J' \subseteq \{1,\dots,k\}$ of cardinality $|J'| > k/2$ such that $\bar{L}_j(\theta) - \bar{L}_j(\theta_0) \geqslant t/4 + 2\frac{\Delta_n}{\sqrt{n}}$ for $j \in J'$ while $\widehat{L}(\theta,\theta_0) < t/4$. In turn, it implies that $\bar{L}_j(\theta) - \bar{L}_j(\theta_0) > 2\frac{\Delta_n}{\sqrt{n}},\ j \in J'$, whence

$$\sum_{j=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}(L_j(\theta) - \bar{L}_j(\theta_0) - \widehat{L}(\theta,\theta_0))\right)$$
$$> \frac{k}{2}\|\rho\|_\infty + \sum_{j\notin J'} \rho'\left(\frac{\sqrt{n}}{\Delta_n}(L_j(\theta) - \bar{L}_j(\theta_0) - \widehat{L}(\theta,\theta_0))\right) > 0,$$

leading to a contradiction. Therefore,

$$\mathbb{P}\left(\inf_{\|\theta-\theta_0\|\geqslant R} \widehat{L}(\theta,\theta_0) < t/4\right)$$
$$\leqslant \mathbb{P}\left(\exists J \subseteq \{1,\dots,k\},\ |J| \geqslant k/2 :\ \inf_{\|\theta-\theta_0\|\geqslant R} \bar{L}_j(\theta) - \bar{L}_j(\theta_0) < t/4 + 2\frac{\Delta_n}{\sqrt{n}},\ j \in J\right). \tag{2.2}$$

Let $\mathcal{E}$ be the event

$$\mathcal{E} = \left\{ \exists J \subseteq \{1,\ldots,k\},\ |J| \geqslant k/2 : \inf_{\|\theta-\theta_0\| \geqslant R} \bar{L}_j(\theta) - \bar{L}_j(\theta_0) < t/4 + 2\frac{\Delta_n}{\sqrt{n}},\ j \in J \right\}.$$

Since at most $\mathcal{O}$ out of k blocks of data may contain outliers, for $\mathcal{E}$ to hold there must be a set of indices J' among the contamination-free blocks of data such that the cardinality of J' satisfies $|J'| \geqslant k/2 - \mathcal{O}$ and such that for all $j \in J'$,

$$\inf_{\|\theta-\theta_0\| \geqslant R} \bar{L}_j(\theta) - \bar{L}_j(\theta_0) < t/4 + 2\frac{\Delta_n}{\sqrt{n}}.$$

Probability of the latter is bounded by, in view of the union bound, by

$$\binom{k-\mathcal{O}}{\lfloor k/2 \rfloor - \mathcal{O}} \left(\tilde{B}(n,R,t) \right)^{\lfloor k/2 \rfloor - \mathcal{O}} \leqslant \tilde{C}^{\lfloor k/2 \rfloor - \mathcal{O}} \left(\tilde{B}(n,R,t) \right)^{\lfloor k/2 \rfloor - \mathcal{O}}$$

whenever $2\frac{\Delta_n}{\sqrt{n}} \leqslant t/2$ and where we used the inequality $\binom{M}{l} \leqslant (Me/l)^l$ together with the fact that $\frac{\mathcal{O}}{k} \leqslant c$ for a sufficiently small absolute constant $c > 0$ and n, k large enough. Here, $\tilde{C} \geqslant \frac{(k-\mathcal{O})e}{\lfloor k/2 \rfloor - \mathcal{O}}$ is another absolute constant whose value depends on c. Moreover, if $n \geqslant n_0(0.25/\tilde{C})$ and $R \geqslant R_0(0.25/\tilde{C})$, we deduce that $\mathbb{P}(\mathcal{E}) < 0.25^{k(1/2-c)-1} \to 0$ as $k \to \infty$ since c is chosen to be small.

As $\widehat{L}(\theta_0, \theta_0) = 0$ a.s., preceding discussion implies that $\mathbb{P}\Big(\|\widehat{\theta}(\theta_0) - \theta_0\| < R \Big) \to 1$ as $n, k, R \to \infty$. We have thus shown that

$$\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}) \to 0 \text{ in probability.} \tag{2.3}$$

On the other hand, by the definition of $\widehat{\theta}^{(2)}_{n,k}$, it holds that $\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}) \geqslant \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0)$. Now, assume that $\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R$ while $\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) < L(\theta_0) + t/2 - L(\theta_0) = t/2$. Arguing as before, we see that there exists $J' \subset \{1, \ldots, k\}$ such that $|J'| > k/2$ and $\bar{L}_j(\widehat{\theta}^{(1)}_{n,k}) - \bar{L}_j(\theta_0) < L(\theta_0) + t/2 - L(\theta_0) + 2\frac{\Delta_n}{\sqrt{n}}$ for $j \in J'$, which implies the inequalities

$$\inf_{\|\theta-\theta_0\|>R} \bar{L}_j(\theta) < L(\theta_0) + t/2 + 2\frac{\Delta_n}{\sqrt{n}} + \left(\bar{L}_j(\theta_0) - L(\theta_0)\right), \ j \in J'.$$

Clearly, $\mathbb{P}\left(\left|\left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right| \geqslant t/4\right) \leqslant \frac{16}{nt^2} \operatorname{Var}\left(\ell(\theta_0, X)\right)$, therefore, for n and R large enough,

$$\mathbb{P}\left(\inf_{\|\theta-\theta_0\|>R} \bar{L}_j(\theta) < L(\theta_0) + t/2 + 2\frac{\Delta_n}{\sqrt{n}} + \left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right) < 0.01$$

for any j. Reasoning as in (2.2), we see that

$$\mathbb{P}\left(\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) < t/2 \text{ and } \|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \to 0 \text{ as } k, n \to \infty.$$

We deduce that on the one hand,

$$\mathbb{P}\left(\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) \geqslant t/2 \bigcap \|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \to \mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right).$$

In view of (2.3), we see that on the other hand,

$$\mathbb{P}\left(\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) \geqslant t/2 \bigcap \|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \leqslant \mathbb{P}\left(\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta_0) \geqslant t/2\right) \to 0,$$

implying that $\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \to 0$ for R large enough as $n, k \to \infty$.

Finally, assume that $\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| > R$ and that $\widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}) > L(\widehat{\theta}^{(1)}_{n,k}) - L(\theta_0) - t/2$. Repeating the reasoning behind (2.2), we see that the latter implies that there exists $J' \subset \{1, \ldots, k\}$ such that $|J'| > k/2$ and $\bar{L}_j(\widehat{\theta}^{(1)}_{n,k}) - \bar{L}_j(\widehat{\theta}^{(2)}_{n,k}) > L(\widehat{\theta}^{(1)}_{n,k}) - \left(L(\theta_0) + t/2 + 2\frac{\Delta_n}{\sqrt{n}}\right)$ for $j \in J'$, yielding that on the event $\left\{\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| \leqslant R\right\}$,

$$\inf_{\|\theta-\theta_0\|>R} \bar{L}_j(\theta) < L(\theta_0) + t/2 + 2\frac{\Delta_n}{\sqrt{n}} + \left(\bar{L}_j(\widehat{\theta}^{(1)}_{n,k}) - L(\widehat{\theta}^{(1)}_{n,k})\right)$$
$$\leqslant L(\theta_0) + t/2 + 2\frac{\Delta_n}{\sqrt{n}} + \sup_{\|\theta'-\theta_0\|\leqslant R} \left|\bar{L}_j(\theta') - L(\theta')\right|$$

for $j \in J'$. We have shown before that $\mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \to 0$ for R large enough as $n, k \to \infty$. As $\mathbb{E}\sup_{\|\theta'-\theta_0\|\leqslant R}\left|\bar{L}_j(\theta') - L(\theta')\right| \to 0$ for any $R > 0$ as $n \to \infty$ (indeed, this follows from (2.1) and the triangle inequality), for n and R large enough, the argument similar to (2.2) implies that

$$\mathbb{P}\left(\sup_{\|\theta-\theta_0\|>R} \widehat{L}(\widehat{\theta}^{(1)}_{n,k}, \theta) > L(\widehat{\theta}^{(1)}_{n,k}) - L(\theta_0) - t/2\right) \to 0 \text{ as } k \to \infty,$$

therefore $\mathbb{P}\left(\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| > R \bigcap \widehat{L}\left(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}\right) \leqslant L(\widehat{\theta}^{(1)}_{n,k}) - (L(\theta_0) + t/2)\right) \to \mathbb{P}\left(\|\widehat{\theta}^{(2)}_{n,k} - \theta_0\| > R\right)$. On the other hand,

$$\mathbb{P}\left(\widehat{L}\left(\widehat{\theta}^{(1)}_{n,k}, \widehat{\theta}^{(2)}_{n,k}\right) \leqslant L(\widehat{\theta}^{(1)}_{n,k}) - (L(\theta_0) + t/2)\right)$$
$$\leqslant \mathbb{P}\left(\widehat{L}\left(\widehat{\theta}^{(1)}_{n,k}, \theta_0\right) \leqslant L(\widehat{\theta}^{(1)}_{n,k}) - (L(\theta_0) + t/2)\right) \leqslant \mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right)$$
$$+\mathbb{P}\left(L(\widehat{\theta}^{(1)}_{n,k}) - L(\theta_0) - \sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta, \theta_0) - (L(\theta) - L(\theta_0))\right| \leqslant L(\widehat{\theta}^{(1)}_{n,k}) - (L(\theta_0) + t/2)\right)$$
$$= \mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta, \theta_0) - (L(\theta) - L(\theta_0))\right| \geqslant t/2\right) + \mathbb{P}\left(\|\widehat{\theta}^{(1)}_{n,k} - \theta_0\| > R\right) \to 0$$

for R large enough as $n, k \to \infty$, therefore completing the proof of consistency.

S3 Proof of Lemma 1 (main text).

We will apply the standard Lindeberg's replacement method (see for example O'Donnell, 2014, chapter 11). For $1 \leqslant j \leqslant n+1$, define $T_j := F\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j}^{n} Z_j\right)$. Then

$$\left|\mathbb{E}F\left(\sum_{j=1}^{n} \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^{n} Z_j\right)\right| = |\mathbb{E}T_{n+1} - \mathbb{E}T_1| \leqslant \sum_{j=1}^{n} |\mathbb{E}T_{j+1} - \mathbb{E}T_j|.$$

Moreover, Taylor's expansion formula gives that there exists (random) $\mu \in [0,1]$ such that

$$\begin{aligned} T_{j+1} = F\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) + F'\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) \xi_j \\ + F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) \frac{\xi_j^2}{2} \\ + \left(F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j + \mu\xi_j\right) - F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right)\right) \frac{\xi_j^2}{2}. \end{aligned}$$

Similarly,

$$\begin{aligned} T_j = F\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) + F'\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) Z_j \\ + F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right) \frac{Z_j^2}{2} \\ + \left(F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j + \mu' Z_j\right) - F''\left(\sum_{i=1}^{j-1} \xi_i + \sum_{i=j+1}^{n} Z_j\right)\right) \frac{Z_j^2}{2}. \end{aligned}$$

Lipschitz continuity and boundedness of F'' imply that

$$|F''(x) - F''(y)| \leqslant C(F) \min(1, |x-y|)$$

with $C(F) = \max\left(2\|F\|_\infty, L(F'')\right)$. Therefore,

$$\begin{aligned}
&|\mathbb{E}T_{j+1} - \mathbb{E}T_j| \\
&\quad \leqslant \left|\mathbb{E}\left(F''\left(\sum_{i=1}^{j-1}\xi_i + \sum_{i=j+1}^{n} Z_j + \mu\xi_j\right) - F''\left(\sum_{i=1}^{j-1}\xi_i + \sum_{i=j+1}^{n} Z_j\right)\right)\frac{\xi_j^2}{2}\right| \\
&\quad + \left|\mathbb{E}\left(F''\left(\sum_{i=1}^{j-1}\xi_i + \sum_{i=j+1}^{n} Z_j + \mu' Z_j\right) - F''\left(\sum_{i=1}^{j-1}\xi_i + \sum_{i=j+1}^{n} Z_j\right)\right)\frac{Z_j^2}{2}\right| \\
&\qquad\qquad\qquad\qquad\qquad\qquad \leqslant C_1(F)\mathbb{E}\left[\xi_j^2 \min(|\xi_j|, 1)\right],
\end{aligned}$$

and the first claim follows. To establish the second inequality, it suffices to observe that for all j, $\mathbb{E}\left[\xi_j^2 \min(|\xi_j|, 1)\right] = \mathbb{E}|\xi_j|^3 I\{|\xi_j| \leqslant 1\} + \mathbb{E}|\xi_j|^2 I\{|\xi_j| > 1\}$. Clearly, $|\xi_j|^3 \leqslant |\xi_j|^{2+\tau}$ on the event $\{|\xi_j| \leqslant 1\}$, whereas $|\xi_j|^2 \leqslant |\xi_j|^{2+\tau}$ on the event $\{|\xi_j| > 1\}$.

S4 Proof of Lemma 2 (main text).

Symmetrization inequality yields that

$$\begin{aligned}
&\mathbb{E}\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^{n}\left(f_{\theta_1}(X_j) - f_{\theta_2}(X_j) - P(f_{\theta_1} - f_{\theta_2})\right)\right|\right)^p \\
&\qquad \leqslant C(p)\mathbb{E}\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^{n}\varepsilon_j\left(f_{\theta_1}(X_j) - f_{\theta_2}(X_j)\right)\right|\right)^p \\
&\qquad\quad = C(p)\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^{n}\varepsilon_j\left(f_{\theta_1}(X_j) - f_{\theta_2}(X_j)\right)\right|\right)^p.
\end{aligned}$$

As the process $f \mapsto \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\varepsilon_j\left(f_{\theta_1}(X_j) - f_{\theta_2}(X_j)\right)$ is sub-Gaussian conditionally on $X_1, \ldots, X_n$, its (conditional) L_p-norms are equivalent to L_1

norm. Hence, Dudley's entropy bound (see Theorem 2.2.4 in van der Vaart and Wellner (1996)) implies that

$$\begin{aligned}\mathbb{E}_\varepsilon \sup_{\theta_1,\theta_2\in\Theta'} &\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n \varepsilon_j\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)\right|\right)^p \\ &\leqslant C(p)\left(\mathbb{E}_\varepsilon \sup_{\theta_1,\theta_2\in\Theta'}\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n \varepsilon_j\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)\right|\right)^p \\ &\qquad\qquad \leqslant C(p)\left(\int_0^{D_n(\Theta')}\log^{1/2} N(z,T_n,d_n)dz\right)^p,\end{aligned}$$

where

$$d_n^2(f_{\theta_1},f_{\theta_2})=\frac{1}{n}\sum_{j=1}^n\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)^2,$$

$$T_n=\{(f_\theta(X_1),\ldots,f_\theta(X_n)),\ \theta\in\Theta'\}\subseteq\mathbb{R}^n$$

and $D_n(\Theta')$ is the diameter of Θ with respect to the distance $d_n(\cdot,\cdot)$. As $f_\theta(\cdot)$ is Lipschitz in θ, we have that $d_n^2(f_{\theta_1},f_{\theta_2})\leqslant\frac{1}{n}\sum_{j=1}^n M^2(X_j)\|\theta_1-\theta_2\|^2$, implying that $D_n(\Theta')\leqslant\|M\|_{L_2(\Pi_n)}\mathrm{diam}(\Theta',\|\cdot\|)$ and

$$\log N(z,T_n,d_n)\leqslant\log N\left(z/\|M\|_{L_2(\Pi_n)},\Theta',\|\cdot\|\right)\leqslant\log\left(C\frac{\mathrm{diam}(\Theta',\|\cdot\|)\,\|M\|_{L_2(\Pi_n)}}{z}\right)^d.$$

Therefore,

$$\left(\int_0^{D_n(\Theta')}\log^{1/2}N(z,T_n,d_n)dz\right)^p\leqslant Cd^{p/2}\left(\mathrm{diam}(\Theta',\|\cdot\|)\cdot\|M\|_{L_2(\Pi_n)}\right)^p$$

and

$$\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n\varepsilon_j\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)\right|\right)^p\leqslant Cd^{p/2}\mathrm{diam}^p(\Theta',\|\cdot\|)\mathbb{E}\|M\|^p_{L_2(\Pi_n)}.$$

Proof of the second bound follows from the triangle inequality

$$\mathbb{E}\sup_{\theta\in\Theta'}\left(\left|\sum_{j=1}^{n}\frac{1}{\sqrt{n}}\left(f_\theta(X_j)-Pf_{\theta_1}\right)\right|\right)^p \leqslant C(p)\left(\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(f_{\theta_0}(X_j)-Pf_{\theta_0})\right|^p\right.$$
$$\left.+\,\mathbb{E}\sup_{\theta\in\Theta'}\left(\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(f_\theta(X_j)-f_{\theta_0}(X_j)-P(f_\theta-f_{\theta_0})\right)\right|\right)^p\right),$$

and Rosenthal's inequality (Ibragimov and Sharakhmetov, 2001) applied to the term $\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(f_{\theta_0}(X_j)-Pf_{\theta_0})\right|^p$.

S5 Proof of Lemma 3 (main text).

First, observe that in view of Assumption 3,

$$\sigma^2(\delta)\leqslant\sup_{\|\theta-\theta_0\|\leqslant\delta}\mathbb{E}|\ell(\theta,X)-\ell(\theta_0,X)|^2\leqslant\mathbb{E}\mathcal{V}^2(X;r(\theta_0))\,\delta^2.$$

Next, define

$$\widehat{G}_k(z;\theta):=\frac{1}{k}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)-z\right)\right)$$

so that $\widehat{G}_k(\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0);\theta)=0$, and let

$$G_k(z;\theta):=\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta,\theta_0)-L(\theta,\theta_0)-z\right)\right)$$

In the definition of $G_k(z;\theta)$, we also assumed that $\bar{L}_1(\theta,\theta_0)$ is based on the contamination-free sample. Next, consider the stochastic process

$$R_k(\theta)=\widehat{G}_k(0;\theta)+\partial_z G_k(z;\theta)\big|_{z=0}\left(\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right).$$

We claim that for any $\theta \in \Theta$,

$$\sqrt{N}\frac{R_k(\theta')}{\partial_z G_k(z;\theta')|_{z=0}} = O_P\left(\frac{\delta^2}{\sqrt{k}} + \sqrt{k}\delta^3 + \frac{\mathcal{O}^2}{k^{3/2}}\right) \tag{2.4}$$

uniformly over θ' in the neighborhood of θ_0. Taking this claim for granted for now, we see that

$$\sqrt{N}\left(\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right) = -\sqrt{N}\frac{\widehat{G}_k(0;\theta)}{\partial_z G_k(z;\theta)|_{z=0}} + \sqrt{N}\frac{R_k(\theta)}{\partial_z G_k(z;\theta)|_{z=0}},$$

and in particular it follows from the claim above that the weak limits of $\sqrt{N}(\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0))$ and

$$-\sqrt{N}\frac{\widehat{G}_k(0;\theta)}{\partial_z G_k(z;\theta)|_{z=0}} = \frac{\Delta_n}{\sqrt{k}}\frac{\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0) - L(\theta,\theta_0)\right)\right)}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta,\theta_0) - L(\theta,\theta_0)\right)\right)}.$$

coincide whenever δ is sufficiently small (note that we can change the order of differentiation and expectation in the denominator as ρ'' is bounded). It remains to establish the relation (2.4) that implies the bound for $\sup_{\|\theta-\theta_0\|\leqslant\delta}|\mathcal{R}_{n,k}(\theta)|$ in the statement of the lemma. To this end, define

$$\widehat{e}_N(\theta) := \widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)$$

so that $\hat{G}_k(\widehat{e}_N(\theta);\theta) = 0$. Recall the definition of $R_k(\theta)$ and observe that the following identity is immediate via Taylor's expansion:

$$R_k(\theta) = \underbrace{\widehat{G}_k\left(\widehat{e}_N(\theta);\theta\right)}_{=0} + \partial_z G_k(z;\theta)\Big|_{z=0}\widehat{e}_N(\theta) - \left(\widehat{G}_k\left(\widehat{e}_N(\theta);\theta\right) - \widehat{G}_k(0;\theta)\right).$$

For any $\theta \in \Theta$ and $j = 1,\ldots,k$, there exists $\tau_j = \tau_j(\theta) \in [0,1]$ such that

$$\begin{aligned}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)-\widehat{e}_N(\theta)\right)\right) &= \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\\ &\quad-\frac{\sqrt{n}}{\Delta_n}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\cdot\widehat{e}_N(\theta)\\ &\quad+\frac{n}{\Delta_n^2}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)-\tau_j\widehat{e}_N(\theta)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{G}_k\left(\widehat{e}_N(\theta);\theta\right)-\widehat{G}_k(0;\theta) &= -\frac{\sqrt{n}}{k\Delta_n}\sum_{j=1}^{k}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\cdot\widehat{e}_N(\theta)\\ &\quad+\frac{n}{k\Delta_n^2}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2\\ &\quad+\frac{n}{k\Delta_n^2}\sum_{j=1}^{k}\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)-\tau_j\widehat{e}_N(\theta)\right)\right)\right.\\ &\qquad\left.-\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2\end{aligned}$$

and

$$\begin{aligned}R_k(\theta) &= \frac{\sqrt{n}}{\Delta_n}\frac{1}{k}\sum_{j=1}^{k}\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right.\\ &\qquad\left.-\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right)\cdot\widehat{e}_N(\theta)\\ &\quad-\frac{n}{\Delta_n^2}\frac{1}{k}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2\\ &\quad-\frac{n}{\Delta_n^2}\frac{1}{k}\sum_{j=1}^{k}\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)-\tau_j\widehat{e}_N(\theta)\right)\right)\right.\\ &\left.-\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2 = R'(\theta)+R''(\theta)+R'''(\theta).\end{aligned} \tag{2.5}$$

It follows from Lemma A.1 (with $\mathcal{O}=0$) and Lemma 2 (see the main paper) that

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}|\widehat{e}_N(\theta)|\leqslant C(d,\theta_0)\left(\frac{\delta}{\sqrt{N}}s+\frac{\delta^2}{\sqrt{n}}+\frac{\mathcal{O}}{k\sqrt{n}}\right)$$

with probability at least $1-s^{-1}$ whenever $s\lesssim\sqrt{k}\wedge\sqrt{n}$. Moreover, Lemma A.2 combined with Lemma 2 yields that

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{k}\sum_{j=1}^{k}\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right.\right.$$
$$\left.\left.-\,\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right)\right|\leqslant C(d,\theta_0)\left(\frac{\delta}{\sqrt{k}}s+\frac{\mathcal{O}}{k}\right)$$

with probability at least $1-s^{-1}$ (here, we also used the fact that at most $\mathcal{O}$ out of k blocks may contain outliers). Therefore, the first term $R'(\theta)$ in (2.5) satisfies

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}|R'(\theta)|\leqslant C(d,\theta_0)\left(\frac{\delta^2}{k}s^2+\frac{\delta^3}{\sqrt{k}}s+\delta^2\frac{\mathcal{O}}{k}+\frac{\mathcal{O}^2}{k^2}\right)$$

on event $\mathcal{E}$ of probability at least $1-\frac{2}{s}$. Observe that

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{k}\sum_{j=1}^{k}\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right.\right.$$
$$\left.\left.-\,\mathbb{E}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right)\right|\leqslant C(d,\theta_0)\left(\frac{\delta}{\sqrt{k}}s+\frac{\mathcal{O}}{k}\right)$$

with probability at least $1-s^{-1}$, again by Lemma A.2, and

$$\left|\mathbb{E}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta,\theta_0)-L(\theta,\theta_0)\right)\right)\right|\leqslant C\delta^2$$

by Lemma 1. Therefore, the term $R''(\theta)$ admits an upper bound

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}|R''(\theta)| \leqslant C(d,\theta_0)\left(\delta^3+\frac{\mathcal{O}^3}{k^3}\right)$$

which holds with probability at least $1-s^{-1}$ (here, we again used the inequality $s\lesssim\sqrt{k}$) to simplify the expression). Finally, as ρ''' is Lipschitz continuous by assumption, the third term $R'''(\theta)$ can be estimated via

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}|R'''(\theta)| \leqslant C(d,\theta_0)\frac{n}{\Delta_n^2}\left|\widehat{e}_N(\theta)\right|^3 \leqslant \frac{C(d,\theta_0)}{\sqrt{n}}\left(\delta^3+\frac{\mathcal{O}^3}{k^3}\right)$$

on event $\mathcal{E}$ (note that this upper bound is smaller than the upper bound for $\sup_{\|\theta-\theta_0\|\leqslant\delta}|R''(\theta)|$ by the multiplicative factor of $\sqrt{n}$). Combining the estimates above and excluding all the higher order terms, it is easy to conclude that

$$\sqrt{N}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{R_k(\theta)}{\partial_z G_k(z;\theta)|_{z=0}}\right| \leqslant C(d,\theta_0)\left(\delta^2\frac{s^2}{\sqrt{k}}+\sqrt{k}\delta^3+\frac{\mathcal{O}^2}{k^{3/2}}\right)$$

with probability at least $1-\frac{3}{s}$.

S6 Proof of Lemma A.1.

Define

$$\hat{G}_k(z;\theta) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-\bar{L}_j(\theta_0)-L(\theta,\theta_0)-z}{\Delta_n}\right),$$

and recall that the contaminated sample $X_1,\ldots,X_N$ contains $\mathcal{O}$ outliers; let $I\subset\{1,\ldots,N\}$ denote the index set of the outliers. Moreover, let

$\tilde{X}_1, \ldots, \tilde{X}_N$ be an i.i.d. sample from P such that $\tilde{X}_j \equiv X_j$ for $j \notin I$, and let $\tilde{G}_k(z;\theta)$ be a version of $\hat{G}_k(z;\theta)$ based on the uncontaminated sample. Clearly, $\left|\hat{G}_k(z;\theta) - \tilde{G}_k(z;\theta)\right| \leqslant 2\|\rho\|_\infty \frac{\mathcal{O}}{\sqrt{k}}$ almost surely, for all $z \in \mathbb{R}$.

Suppose that $z_1, z_2 \in \mathbb{R}$ are such that on an event of probability close to 1, $\hat{G}_k(z_1;\theta) > 0$ and $\hat{G}_k(z_2;\theta) < 0$ for all $\theta \in \Theta$ simultaneously. Since $\hat{G}_k$ is non-increasing in z, it is easy to see that on this event, $\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0) \in (z_1, z_2)$ for all $\theta \in \Theta$, implying that

$$\sup_{\theta\in\Theta'} \left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \leqslant \max(|z_1|, |z_2|). \tag{2.6}$$

Our goal is to find z_1, z_2 satisfying conditions above and such that $|z_1|$, $|z_2|$ are as small as possible. Let $W(\theta)$ stand for a centered normally distributed random variable with variance $\sigma^2(\theta,\theta_0)$, and observe that

$$\hat{G}_k(z;\theta) = A_0 + A_1 + A_2 + A_3,$$

where

$$\begin{aligned}
A_0(\theta) &= \hat{G}_k(z;\theta) - \tilde{G}_k(z;\theta),\\
A_1(\theta) &= \frac{1}{\sqrt{k}} \sum_{j=1}^{k} \left(\rho'\left(\sqrt{n}\, \frac{\bar{L}_j(\theta) - \bar{L}_j(\theta_0) - L(\theta,\theta_0) - z}{\Delta_n} \right) \right.\\
&\qquad \left. - \mathbb{E}\rho'\left(\sqrt{n}\, \frac{\bar{L}_j(\theta) - \bar{L}_j(\theta_0) - L(\theta,\theta_0) - z}{\Delta_n} \right) \right),\\
A_2(\theta) &= \sqrt{k}\left(\mathbb{E}\rho'\left(\sqrt{n}\, \frac{\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0) - z}{\Delta_n} \right) - \mathbb{E}\rho'\left(\frac{W(\theta) - \sqrt{n}z}{\Delta_n} \right) \right),\\
A_3(\theta) &= \sqrt{k}\mathbb{E}\rho'\left(\frac{W(\theta) - \sqrt{n}z}{\Delta_n} \right).
\end{aligned}$$

With some abuse of notation, we assume that $A_1(\theta)$ and $A_2(\theta)$ are evaluated based on the contamination-free sample $\tilde{X}_1, \ldots, \tilde{X}_N$. Next, suppose that $\varepsilon_0, \varepsilon_1, \varepsilon_2$ are positive and such that

$$\inf_{\theta\in\Theta'} A_0(\theta) > -\varepsilon_0, \quad \inf_{\theta\in\Theta'} A_1(\theta) > -\varepsilon_1$$

with high probability and $\inf_{\theta\in\Theta'} A_2(\theta) > -\varepsilon_2$. Then z_1 satisfying

$$\inf_{\theta\in\Theta'} \mathbb{E}\rho'\left(\frac{W(\theta)-\sqrt{n}z_1}{\Delta_n}\right) \geqslant \frac{\varepsilon_0+\varepsilon_1+\varepsilon_2}{\sqrt{k}}$$

will conform to our requirements. Since

$$\mathbb{E}\rho'\left(\frac{W(\theta)-\sqrt{n}z_1}{\Delta_n}\right) \approx \underbrace{\mathbb{E}\rho'\left(\frac{W(\theta)}{\Delta_n}\right)}_{=0} - \mathbb{E}\rho''\left(\frac{W(\theta)}{\Delta_n}\right)\frac{\sqrt{n}z_1}{\Delta_n}$$

for small z_1, a natural choice is $z_1 \approx \frac{\Delta_n}{\inf_{\theta\in\Theta'}\mathbb{E}\rho''\left(\frac{W(\theta)}{\Delta_n}\right)}\frac{\varepsilon_0+\varepsilon_1+\varepsilon_2}{\sqrt{nk}}$. This argument is made precise in (Minsker, 2019b, Lemma 4.3) which shows that the choice

$$z_1 = -\frac{\varepsilon_0+\varepsilon_1+\varepsilon_2}{0.09}\frac{\widetilde{\Delta}}{\sqrt{nk}}$$

is sufficient whenever $\varepsilon_j,\ j = 0,1,2$ are not too large (specifically, when $\frac{\varepsilon_0+\varepsilon_1+\varepsilon_2}{\sqrt{k}} \leqslant 0.045$ - this is precisely the main condition needed for the bound of lemma to hold). It remains to provide the values for $\varepsilon_j,\ j = 0,1,2$. We have already shown above that ε_0 can be chosen as $\varepsilon_0 = 2\|\rho\|_\infty \frac{\mathcal{O}}{\sqrt{k}}$. To find a feasible value of ε_1, we will apply Markov's inequality stating that with probability at least $1-1/s$,

$$\begin{aligned}\sup_{\theta\in\Theta'}|A_1(\theta)| \leqslant s\,\mathbb{E}\sup_{\theta\in\Theta'}&\left|\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-\bar{L}_j(\theta_0)-L(\theta,\theta_0)-z\right)\right)\right.\\ &\left.-\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta)-\bar{L}_1(\theta_0)-L(\theta,\theta_0)-z\right)\right)\right|.\end{aligned}$$

The expected supremum can be estimated in a standard way using the symmetrization, contraction and desymmetrization inequalities (e.g. see the proof of Lemma A.2), yielding that

$$\begin{aligned}\mathbb{E}\sup_{\theta\in\Theta'}&\left|\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L_j(\theta_0)-L(\theta,\theta_0)-z\right)\right)\right.\\ &\qquad\left.-\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta)-\bar{L}_1(\theta_0)-L(\theta,\theta_0)-z\right)\right)\right|\\ &\qquad\leqslant\frac{8L(\rho')}{\Delta_n}\mathbb{E}\sup_{\theta\in\Theta'}\frac{1}{\sqrt{N}}\left|\sum_{j=1}^{N}\left(\ell(\theta,X_j)-\ell(\theta_0,X_j)-L(\theta,\theta_0)\right)\right|.\end{aligned}$$

It remains to obtain an appropriate value for ε_2. Note that for any bounded non-negative function $g:\mathbb{R}\mapsto\mathbb{R}_+$ and any signed measure Q,

$$\left|\int_{\mathbb{R}}g(x)dQ\right|=\left|\int_0^{\|f\|_\infty}Q\left(x:\,g(x)\geqslant t\right)dt\right|\leqslant\|g\|_\infty\max_{t\geqslant0}\left|Q\left(x:\,g(x)\geqslant t\right)\right|.$$

Moreover, if g is monotone, the sets $\{x:\,g(x)\geqslant t\}$ and $\{x:\,g(x)\leqslant t\}$ are half-intervals. Note that $\rho'=\max(\rho',0)-\max(-\rho',0)$ is a difference of two non-negative monotone functions. Therefore,

$$\begin{aligned}&\left|\int_{\mathbb{R}}\rho'\left(\frac{x-\sqrt{n}z}{\Delta_n}\right)dQ(x)\right|\\ &\qquad\leqslant\|\rho'\|_\infty\left(\max_{t\geqslant0}\left|Q\left(x:\,\rho'(x)\geqslant t\right)\right|+\max_{t\leqslant0}\left|Q\left(x:\,\rho'(x)\leqslant t\right)\right|\right).\end{aligned}$$

Take Q to be the difference of the distributions of $\sqrt{n}\,\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)$ and $W(\theta)$, denoted $\Phi_\theta^{(n,k)}$ and Φ_θ respectively, so that

$$\sqrt{k}\left(\mathbb{E}\rho'\left(\sqrt{n}\,\frac{\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0) - z}{\Delta_n}\right) - \mathbb{E}\rho'\left(\frac{W(\theta) - \sqrt{n}z}{\Delta_n}\right)\right) \\ \leqslant 2\sqrt{k}\|\rho'\|_\infty \sup_{t\in\mathbb{R}}\left|\Phi_\theta^{(n,k)}(t) - \Phi_\theta(t)\right|.$$

A well-known result by Feller (1968) states that $\sup_{t\in\mathbb{R}}\left|\Phi_\theta^{(n,k)}(t) - \Phi_\theta(t)\right| \leqslant 6g_\theta(n)$, where

$$g_\theta(n) := \frac{1}{\sqrt{n}}\mathbb{E}\Bigg[\left(\frac{\ell(\theta,X) - \ell(\theta_0,X) - L(\theta,\theta_0)}{\sigma(\theta,\theta_0)}\right)^2 \\ \times \min\left(\left|\frac{\ell(\theta,X) - \ell(\theta_0,X) - L(\theta,\theta_0)}{\sigma(\theta,\theta_0)}\right|, \sqrt{n}\right)\Bigg],$$

It is easy to see that $g_\theta(n) \to 0$ as $n \to \infty$ if $\mathrm{Var}(\ell(\theta,X)) < \infty$, and distributions with finite variance, and moreover $g_\theta(n) \leqslant C\mathbb{E}\left|\frac{\ell(\theta,X)-\ell(\theta_0,X)-L(\theta,\theta_0)}{\sigma(\theta,\theta_0)}\right|^\tau n^{-\tau/2}$ if $\mathbb{E}\left|\frac{\ell(\theta,X)-\ell(\theta_0,X)-L(\theta,\theta_0)}{\sigma(\theta,\theta_0)}\right|^{2+\tau} < \infty$ for some $\tau \in (0,1]$. Therefore, the function $g_\tau(n,\theta)$ in the statement of the lemma can be chosen as $g_\tau(n,\theta) = g_\theta(n)$ when $\tau = 0$ and $g_\tau(n,\theta) = C$ when $\tau > 0$. We conclude that the choice $\varepsilon_2 = 12\sqrt{k}\|\rho'\|_\infty \sup_{\theta\in\Theta'} g_\theta(n)$ satisfies the desired requirements.

It remains to recall the bound (2.6) and that $z_1 = -\frac{\varepsilon_0+\varepsilon_1+\varepsilon_2}{0.09}\frac{\tilde{\Delta}}{\sqrt{nk}}$. The matching bound for z_2 is obtained in an identical fashion.

Remark 2. The bound for ε_2 that we established above is slightly weaker than the one used in the statement of the lemma; an improved version can be obtained using the non-uniform version of the Berry-Esseen bound with additional effort, and we refer the reader to (Minsker, 2019b, Lemma 4.2) for the technical details.

S7 Proof of Lemma A.2.

Let $\varepsilon_1, \ldots, \varepsilon_k$ be i.i.d. Rademacher random variables independent of $X_1, \ldots, X_N$, and note that by symmetrization and contraction inequalities for the Rademacher sums (Ledoux and Talagrand, 1991),

$$\begin{aligned}
\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} \Bigg| \frac{1}{k}\sum_{j=1}^{k} \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right) \\
- \mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta',\theta_0) - L(\theta',\theta_0)\right)\right)\Bigg| \\
\leqslant 2\,\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} \frac{1}{k}\left|\sum_{j=1}^{k} \varepsilon_j \left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right) - \rho''(0)\right)\right| \\
\leqslant \frac{4L(\rho'')}{\Delta_n\sqrt{k}}\,\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta}\left|\sum_{j=1}^{k}\varepsilon_j \frac{\sqrt{n}}{\sqrt{k}}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right|,
\end{aligned}$$

where we used the fact that $\phi(x) := \rho''\left(\frac{\sqrt{n}}{\Delta_n}x\right) - \rho''(0)$ is Lipschitz continuous (in fact, Assumption 1 implies that the Lipschitz constant is equal to 1) and satisfies $\phi(0) = 0$. Now, desymmetrization inequality (Lemma 2.3.6 in van der Vaart and Wellner, 1996) implies that

$$\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta}\left|\sum_{j=1}^{k}\varepsilon_j \frac{\sqrt{n}}{\sqrt{k}}\left(\bar{L}_j(\theta',\theta_0) - L(\theta',\theta_0)\right)\right|$$

$$\leqslant \frac{2}{\sqrt{N}}\mathbb{E} \sup_{\|\theta'-\theta\|\leqslant\delta} \left| \sum_{j=1}^{N} \left(\ell(\theta', X_j) - \ell(\theta_0, X_j) - L(\theta', \theta_0)\right) \right|,$$

hence the claim follows.

The fact that ρ'' can be replaced by ρ''' follows along the same lines as ρ''' is Lipschitz continuous and $\|\rho'''\|_\infty < \infty$ by Assumption 1.

S8 Numerical experiment: logistic regression.

As a simple proof of concept, we implemented the gradient descent-ascent algorithm mentioned in section 2.1 for the problem of logistic regression; for a detailed discussion of closely related methods, we refer the reader to (Lecué and Lerasle, 2020; Mathieu and Minsker, 2021). In the present setup, the dataset consists of pairs $(Z_j, Y_j) \in \mathbb{R}^2 \times \{\pm 1\}$, where the marginal distribution of the labels is uniform on $\{\pm 1\}$, while the conditional distributions of Z_j's are normal, that is, $\text{Law}\,(Z_1 \,|\, Y_1 = 1) = \mathcal{N}\left((-1,-1)^T, 4I_2\right)$, $\text{Law}\,(Z \,|\, Y = -1) \sim \mathcal{N}\left((1,1), 4I_2\right)$, and $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$; here, I_2 stands for the 2×2 identity matrix. The loss function is defined as $\ell(\theta, Z, Y) = \log\left(1 + e^{-Y\langle\theta, Z\rangle}\right)$, $\theta \in \mathbb{R}^2$. The dataset includes 40 outliers for which $Y_j \equiv 1$ and $Z \sim \mathcal{N}\left((25,10), 0.25 I_2\right)$. The sample of 500 "informative" observations was generated, along with 40 outliers, and we compared the performance of robust method proposed in this paper with the standard logistic regression, as implemented in the Scikit-learn package (Pedregosa et al., 2011), that is known to be sensitive to outliers. Results of the experiment are presented in figure 1 and illustrate the robustness of proposed approach.

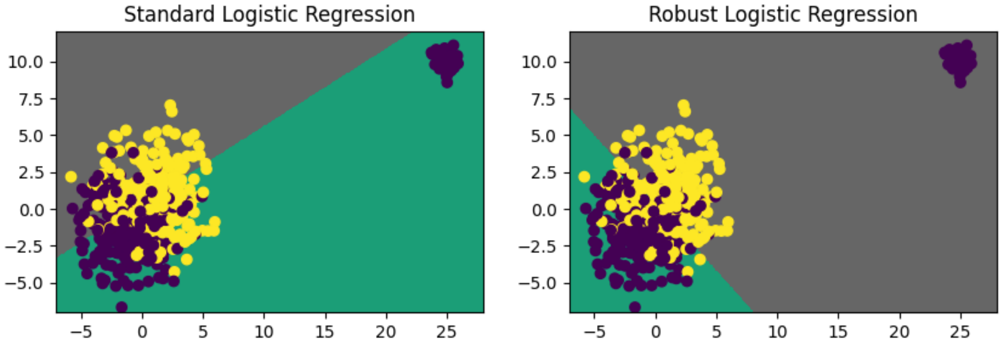

Figure 1: Scatter plot of $N = 540$ samples from the training dataset (500 informative observations and 40 outliers). The color of the points correspond to their labels and the background color – to the predicted labels (gray region corresponds to yellow labels and green – to purple labels).

References

Alistarh, D., Z. Allen-Zhu, and J. Li (2018). Byzantine stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4613–4623.

Alon, N., Y. Matias, and M. Szegedy (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 20–29. ACM.

REFERENCES

Audibert, J.-Y., O. Catoni, et al. (2011). Robust linear least squares regression. *The Annals of Statistics 39*(5), 2766–2794.

Brownlees, C., E. Joly, G. Lugosi, et al. (2015). Empirical risk minimization for heavy-tailed losses. *The Annals of Statistics 43*(6), 2507–2536.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 48, pp. 1148–1185. Institut Henri Poincaré.

Chen, Y., L. Su, and J. Xu (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 1*(2), 1–25.

Cherapanamjeri, Y., S. B. Hopkins, T. Kathuria, P. Raghavendra, and N. Tripuraneni (2019). Algorithms for heavy-tailed statistics: Regression, covariance estimation, and beyond. *arXiv preprint arXiv:1912.11071*.

Devroye, L., M. Lerasle, G. Lugosi, and R. I. Oliveira (2016). Sub-Gaussian mean estimators. *The Annals of Statistics 44*(6), 2695–2725.

Feller, W. (1968). On the Berry-Esseen theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 10*(3), 261–268.

Holland, M. J. and K. Ikeda (2017). Robust regression using biased objectives. *Machine Learning 106*(9-10), 1643–1679.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical*

Statistics 35(1), 73–101.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer.

Ibragimov, R. and S. Sharakhmetov (2001). The best constant in the Rosenthal inequality for nonnegative random variables. *Statistics & probability letters 55*(4), 367–376.

Lecué, G. and M. Lerasle (2020). Robust machine learning by median-of-means: theory and practice. *The Annals of Statistics 48*(2), 906–931.

Lecué, G., M. Lerasle, and T. Mathieu (2020). Robust classification via MOM minimization. *Machine learning 109*, 1635–1665.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: isoperimetry and processes.* Berlin: Springer-Verlag.

Lerasle, M. and R. I. Oliveira (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Lugosi, G. and S. Mendelson (2019a). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics 19*(5), 1145–1190.

Lugosi, G. and S. Mendelson (2019b). Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society 22*(3), 925–965.

Mathieu, T. and S. Minsker (2021). Excess risk bounds in robust empirical risk minimization. *Information and Inference: A Journal of the IMA 10*(4), 1423–1490.

Minsker, S. (2019a). Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics 13*(2), 5213–5252.

Minsker, S. (2019b). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

Minsker, S. and S. Yao (2025). Generalized median of means principle for Bayesian inference. *Machine Learning 114*(4), 115.

Nemirovski, A. and D. Yudin (1983). *Problem complexity and method efficiency in optimization.* John Wiley & Sons Inc.

O'Donnell, R. (2014). *Analysis of boolean functions.* Cambridge University Press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*(Oct), 2825–2830.

Prasad, A., A. S. Suggala, S. Balakrishnan, P. Ravikumar, et al. (2020). Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B 82*(3), 601–627.

Talagrand, M. (2005). *The generic chaining.* Springer.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. New York: Springer-Verlag.

Yin, D., Y. Chen, R. Kannan, and P. Bartlett (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR.