Asymptotic normality of robust risk minimizers

Stanislav Minsker [*,1]

[1] *Department of Mathematics, University of Southern California*
e-mail: [*] minsker@usc.edu

Abstract: This paper investigates asymptotic properties of a class of algorithms that can be viewed as robust analogues of the classical empirical risk minimization. These strategies are based on replacing the usual empirical average by a robust proxy of the mean, such as the median-of-means estimator. It is well known by now that the excess risk of resulting estimators often converges to 0 at the optimal rates under much weaker assumptions than those required by their "classical" counterparts. However, much less is known about asymptotic properties of the estimators themselves, for instance, whether robust analogues of the maximum likelihood estimators are asymptotically efficient. We make a step towards answering these questions and show that for a wide class of parametric problems, minimizers of the appropriately defined robust proxy of the risk converge to the minimizers of the true risk at the same rate, and often have the same asymptotic variance, as the classical M-estimators. Moreover, our results show that the algorithms based on "tournaments" or "min-max" type procedures asymptotically outperform algorithms based on direct risk minimization.

Keywords and phrases: robust estimation, median-of-means estimator, adversarial contamination, asymptotic normality, consistency.

1. Introduction.

The concept of robustness addresses stability of statistical estimators under various forms of perturbations, such as the presence of corrupted observations ("outliers") in the data. The questions related to robustness in the framework of statistical learning theory have recently seen a surge in interest, both from the theoretical and practical perspectives, and resulted in the development of novel algorithms. These new robust algorithms are characterized by the fact that they provably work under minimal assumptions on the underlying data-generating mechanism, often requiring the existence of moments of low order only. Majority of existing works have focused on the bounds for the risk of the estimators (e.g., the classification or prediction error) produced by the algorithms, while in this paper we are interested in the asymptotic properties of the estimators themselves.

Next, we introduce the mathematical framework used in the paper. Let $(S,\mathcal{S})$ be a measurable space, and let $X \in S$ be a random variable with distribution P. Suppose that $X_1,\ldots,X_N$ are i.i.d. copies of X. Moreover, assume that $\mathcal{L} = \left\{\ell(\theta,\cdot),\ \theta \in \Theta \subseteq \mathbb{R}^d\right\}$ is a class of measurable functions from S to $\mathbb{R}$ indexed by an open subset of $\mathbb{R}^d$. "Population" versions of many estimation problems in statistics and statistical learning can be formulated as risk minimization of the form

$$\mathbb{E}\,\ell(\theta,X) \to \min_{\theta\in\Theta}. \tag{1}$$

In particular, when $\{p_\theta,\ \theta\in\Theta\}$ is a family of probability density functions with respect to some σ-finite measure μ and $\ell(\theta,\cdot) = -\log p_\theta(\cdot)$, the resulting problem corresponds to maximum likelihood estimation. In what follows, we will set $L(\theta) := \mathbb{E}\ell(\theta,X)$. Throughout the paper, we will assume that the minimum in problem (1) is attained at a unique point $\theta_0 \in \Theta$. The

[*] Author acknowledges support by the National Science Foundation grants DMS-1712956 and CIF-1908905.

true distribution P is typically unknown, and an estimator of θ_0 is obtained via minimizing the *empirical risk* , namely

$$\widetilde{\theta}_N := \operatorname*{argmin}_{\theta\in\Theta} L_N(\theta), \tag{2}$$

where $L_N(\theta) := \frac{1}{N}\sum_{j=1}^N \ell(\theta, X_j)$. If the marginal distributions of the process $\{\ell(\theta,\cdot),\ \theta\in\Theta\}$ are heavy-tailed, meaning that they possess finite moments of low order only, then the error $|L_N(\theta) - L(\theta)|$ can be large with non-negligible probability, motivating the need of alternative proxies for the risk $L(\theta)$. Another scenario of interest corresponds to *adversarial contamination* framework, where the initial dataset of cardinality N' is merged with a set of $\mathcal{O} < N$ outliers generated by an adversary who has an opportunity to inspect the data, and the combined dataset of cardinality $N = N' + \mathcal{O}$ is presented to the statistician. In what follows, the proportion of outliers will be denoted by $\kappa := \kappa_N = \frac{\mathcal{O}}{N}$. Similarly to the heavy-tailed scenario, the empirical loss $L_N(\theta)$ is not a robust proxy for $\mathbb{E}\ell(\theta, X)$, therefore estimation and inference results based on minimizing $L_N(\theta)$ may be unreliable.

One may approach the problem of estimating θ_0 robustly from different angles. One class of popular methods consists of robust versions of the gradient descent algorithm for the optimization problem (1), where the gradient $\nabla L(\theta_k)$ is estimated on each iteration k; for example, this approach has been explored by Prasad et al. (2018); Chen et al. (2017); Yin et al. (2018); Alistarh et al. (2018), among others. Another technique (the one that we investigate in this paper) is based on replacing the average $L_N(\cdot)$ by a robust proxy of $L(\theta)$. Its advantage is the fact that we only need to estimate a real-valued quantity $L(\theta)$, as opposed to the high-dimensional gradient $\nabla L(\theta)$. On the other hand, favorable properties, such as convexity, that are "inherited" by the formulation (2) from (1), are usually lost in this case. Several representative papers that explore this direction include the works by Audibert et al. (2011); Lerasle and Oliveira (2011); Brownlees et al. (2015); Lugosi and Mendelson (2016); Lecué and Lerasle (2017); Cherapanamjeri et al. (2019); Minsker and Mathieu (2019); also, see an excellent survey paper by Lugosi and Mendelson (2019). Instead of the empirical risk $L_N(\theta)$, these works employ robust estimators of the risk such as the median-of-means estimator Nemirovski and Yudin (1983); Alon et al. (1996); Devroye et al. (2016) or Catoni's estimator Catoni (2012).

In this paper, we study estimators based on the modifications of the median-of-means principle introduced in (Minsker, 2019a) and constructed as follows. Let $k \leqslant N/2$ be an integer, and assume that $G_1, \ldots, G_k$ are disjoint subsets of the index set $\{1, \ldots, N\}$ of cardinality $|G_j| = n \geqslant \lfloor N/k \rfloor$ each. For $\theta \in \Theta$, let

$$\bar{L}_j(\theta) := \frac{1}{n}\sum_{i\in G_j} \ell(\theta, X_i)$$

be the empirical risk evaluated over the subsample indexed by G_j. Assume that $\rho : \mathbb{R} \mapsto \mathbb{R}_+$ is a convex, even function that is increasing on $(0,\infty)$ and such that its (right) derivative is bounded. Let $\{\Delta_n\}_{n\geqslant 1}$ be a bounded sequence of positive scalars ("scaling factors") such that $\Delta_\infty := \lim_{n\to\infty}\Delta_n \in (0,\infty]$ exists, and define

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

$\widehat{L}(\theta)$ is the estimator that we referred to as a robust proxy of $L(\theta)$, where robustness is justified by the fact that the error $\left|\widehat{L}(\theta) - L(\theta)\right|$ satisfies non-asymptotic exponential deviation bounds under minimal assumptions on the tails of the random variable $\ell(\theta, X)$, and the ability of $\widehat{L}(\theta)$ to resist adversarial outliers. For example, Theorem 3 in (Minsker, 2019a) essentially states that for $s \lesssim k$, $\left|\widehat{L}(\theta) - L(\theta)\right| \leqslant C\,\mathrm{Var}^{1/2}\left(\sqrt{\frac{s}{N}} + \frac{k}{N} + \frac{\mathcal{O}\sqrt{n}}{N}\right)$ with probability at least $1 - e^{-s}$, assuming that $\mathbb{E}|\ell(\theta,X)|^3<\infty$. A natural analogue of the empirical risk minimizer $\widetilde{\theta}_N$ can be obtained by minimizing the robust risk estimator $\widehat{L}(\theta)$, that is,

$$\widehat{\theta}_{n,k} := \operatorname*{argmin}_{\theta\in\Theta} \widehat{L}(\theta), \tag{3}$$

where the minimum is assumed to be achieved (this is not crucial for our results, as they still hold if the sequence $\{\widehat{\theta}_{n_j,k_j}\}_{j\geqslant 1}$ is replaced by a sequence of near-minimizers). This approach has been previously investigated by Brownlees et al. (2015); Holland and Ikeda (2017); Lecué et al. (2018); Minsker and Mathieu (2019), where the main object of interest was the excess risk $\mathcal{E}(\widehat{\theta}_N) := L(\widehat{\theta}_N) - L(\theta_0)$. In the present work however, we will be interested in the asymptotic behavior of the estimator error $\widehat{\theta}_N - \theta_0$, rather than the risk: specifically, we will establish asymptotic normality of $\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right)$. The nonlinear nature of the estimator $\widehat{L}(\theta)$ makes the proofs more technical compared to the classical theory of empirical risk minimization.

Another fruitful approach is based on the so-called "median-of-means tournaments" (Lugosi and Mendelson, 2016) and the closely related "min-max" (Audibert et al., 2011; Lecué and Lerasle, 2017) estimators. The latter is based on a simple observation that $\theta_0 = \operatorname{argmin}_{\theta\in\Theta}\max_{\theta'\in\Theta}\left(L(\theta)-L(\theta')\right)$. Therefore, an estimator of θ_0 can be constructed by replacing the difference $L(\theta)-L(\theta')$ by its robust proxy. A natural candidate is defined via

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

where $\{\Delta_n\}_{n\geqslant 1}$ is same as before, and

$$\left(\widehat{\theta}^{(1)}_{n,k},\widehat{\theta}^{(2)}_{n,k}\right) = \operatorname*{argmin}_{\theta\in\Theta}\max_{\theta'\in\Theta} \widehat{L}(\theta,\theta'). \tag{4}$$

This approach, based on a seemingly "tautological" idea, often leads to stronger results as it allows to overcome the difficulty caused by the fact that $\widehat{L}(\theta,\theta')\neq\widehat{L}(\theta)-\widehat{L}(\theta)$ by directly estimating the difference of the risks. Our results demonstrate that in the classical parametric framework, the sequences $\sqrt{N}\left(\widehat{\theta}^{(i)}_{n,k}-\theta_0\right)$, $i=1,2$, are asymptotically normal under essentially the same set of sufficient conditions as required by the standard M-estimators (Van der Vaart, 2000). On the other hand, conditions required for asymptotic normality of $\sqrt{N}\left(\widehat{\theta}_N-\theta_0\right)$ are slightly more restrictive. A somewhat surprising fact is that is that the estimators $\widehat{\theta}^{(i)}_{n,k}$, $i=1,2$, are often asymptotically efficient, while $\widehat{\theta}_N$ is not.

Finally, let us remark that the "classical" median-of-means estimator corresponds to $\rho(x)=|x|$. In this paper, we will only deal with smooth loss functions ρ, and the possibility of extensions of our results to the case $\rho(x)=|x|$ is left as an open problem.

1.1. Notation.

Absolute constants will be denoted c, c_1, C, C_1, C', etc., and may take different values in different parts of the paper. Given $a,b\in\mathbb{R}$, we will write $a\wedge b$ for $\min(a,b)$ and $a\vee b$ for $\max(a,b)$. For a function $f:\mathbb{R}^d\mapsto\mathbb{R}$, define

$$\operatorname*{argmin}_{y\in\mathbb{R}^d} f(y) := \{y\in\mathbb{R}^d : f(y)\leqslant f(x) \text{ for all } x\in\mathbb{R}^d\},$$

and $\|f\|_\infty := \operatorname{ess\,sup}\{|f(y)| : y\in\mathbb{R}^d\}$. Moreover, $L(f)$ will stand for the Lipschitz constant of f; if $d=1$ and f is m times differentiable, $f^{(m)}$ will denote the m-th derivative of f. For a function

$g(\theta, x)$ mapping $\mathbb{R}^d \times \mathbb{R}$ to $\mathbb{R}$, $\partial_\theta g$ will denote the vector of partial derivatives with respect to the coordinates of θ; similarly, $\partial^2_\theta g$ will denote the matrix of second partial derivatives.

For $x \in \mathbb{R}^d$, $\|x\|$ will stand for the Euclidean norm of x, $\|x\|_\infty := \max_j |x_j|$, and for a matrix $A \in \mathbb{R}^{d\times d}$, $\|A\|$ will denote the spectral norm of A. We will frequently use the standard big-O and small-o notation, as well as their in-probability siblings o_P and O_P. For vector-valued sequences $\{x_j\}_{j\geqslant 1}$, $\{y_j\}_{j\geqslant 1} \subset \mathbb{R}^d$, expressions $x_j = o(y_j)$ and $x_j = O(y_j)$ are assumed to hold coordinate-wise. For a square matrix $A \in \mathbb{R}^{d\times d}$, $\operatorname{tr} A := \sum_{j=1}^d A_{j,j}$ denotes the trace of A.

Given a function $g : \mathbb{R} \mapsto \mathbb{R}$, measure Q and $1 \leqslant p < \infty$, we set $\|g\|^p_{L_p(Q)} := \int_{\mathbb{R}} |g(x)|^p dQ$. For i.i.d. random variables $X_1, \ldots, X_N$ distributed according to P, $P_N := \frac{1}{N}\sum_{j=1}^N \delta_{X_j}$ will stand for the empirical measure; here, $\delta_X(g) := g(X)$. The expectation with respect to a probability measure Q will be denoted $\mathbb{E}_Q$; if the measure is not specified, it will be assumed that the expectation is taken with respect to P, the distribution of X. Given $f : S \mapsto \mathbb{R}^d$, we will write Qf for $\int f dQ \in \mathbb{R}^d$, assuming that the last integral is calculated coordinate-wise. For $\theta \in \Theta$, let $\sigma^2(\theta) = \operatorname{Var}(\ell(\theta, X))$ and for $\Theta' \subseteq \Theta$, define $\sigma^2(\Theta') := \sup_{\theta\in\Theta'} \sigma^2(\theta)$.

Finally, we will adopt the convention that the infimum over the empty set is equal to $+\infty$. Additional notation and auxiliary results are introduced on demand.

2. Statements of the main results.

We begin by listing the assumptions on the model; these conditions are very similar to the standard assumptions made in the parametric framework (Van der Vaart, 2000; van der Vaart and Wellner, 1996). The first assumption lists the requirements for the loss function ρ (chosen by the statistician).

Assumption 1. *The function $\rho : \mathbb{R} \mapsto \mathbb{R}$ is convex, even, and such that*

(i) $\rho'(z) = z$ *for* $|z| \leqslant 1$ *and* $\rho'(z) = \text{const}$ *for* $z \geqslant 2$.
(ii) $z - \rho'(z)$ *is nondecreasing;*
(iii) $\rho^{(5)}$ *is bounded and Lipschitz continuous.*

An example of a function ρ satisfying required assumptions is given by "smoothed" Huber's loss defined as follows. Let

$$H(y) = \frac{y^2}{2} I\{|y| \leqslant 3/2\} + \frac{3}{2}\left(|y| - \frac{3}{4}\right) I\{|y| > 3/2\}$$

be the usual Huber's loss. Moreover, let ψ be the "bump function" $\psi(x) = C\exp\left(-\frac{4}{1-4x^2}\right)\left\{|x| \leqslant \frac{1}{2}\right\}$ where C is chosen so that $\int_{\mathbb{R}} \psi(x)dx = 1$. Then ρ given by the convolution $\rho(x) = (h * \psi)(x)$ satisfies assumption 1.

Assumption 2. *The Hessian $\partial^2_\theta L(\theta_0)$ exists and is strictly positive definite.*

An immediate implication of this assumption is the fact that in a sufficiently small neighborhood of θ_0, $c(\theta_0)\|\theta - \theta_0\|^2 \leqslant L(\theta) - L(\theta_0) \leqslant C(\theta_0)\|\theta - \theta_0\|^2$ for some $0 < c(\theta_0) \leqslant C(\theta_0) < \infty$.

Assumption 3A. *For any $\theta \in \Theta$, there exists a ball $B(\theta, r(\theta))$ such that for all $\theta', \theta'' \in B(\theta, r(\theta))$ and some measurable function M_θ, $|\ell(\theta', x) - \ell(\theta'', x)| \leqslant M_\theta(x)\|\theta' - \theta''\|$. Moreover, $\mathbb{E}M_\theta^{2+\tau}(X) < \infty$ for some $\tau \geqslant 0$. Finally, for every $\theta \in \Theta$, the map $\theta' \mapsto \ell(\theta', x)$ is differentiable at θ for P-almost all x (where the exceptional set of measure 0 can depend on θ), with derivative $\partial_\theta \ell(\theta, x)$.*

Assumption 3A is very similar to the standard set of sufficient conditions for the asymptotic normality of M-estimators, see for instance Theorem 5.23 in the book by Van der Vaart (2000).

Stronger requirement $\tau > 0$ is only needed to prove asymptotic normality of the estimator (3), while the rest of our results hold with $\tau = 0$. Next, define

$$\omega_N(\delta) := \mathbb{E} \sup_{\|\theta'-\theta_0\| \leqslant \delta} \sqrt{N} \left\| (P_N - P) \left(\partial_\theta \ell(\theta', \cdot) - \partial_\theta \ell(\theta_0, \cdot) \right) \right\|.$$

Assumption 3B. *Suppose that* $\lim_{\delta \to 0} \limsup_{N \to \infty} \omega_N(\delta) = 0$. *Moreover, there exists* $\tau > 0$ *such that* $\mathbb{E} \left|\ell(\theta, X)\right|^{2+\tau} < \infty$ *for all* $\theta \in \Theta$, *and the envelope function* $\mathcal{V}(x;\delta) := \sup_{\|\theta-\theta_0\| \leqslant \delta} \|\partial_\theta \ell(\theta, x)\|$ *of the class* $\{\partial_\theta \ell(\theta, \cdot) : \|\theta - \theta_0\| \leqslant \delta\}$ *satisfies* $\mathbb{E}\mathcal{V}^{2+\tau}(X;\delta) < \infty$ *for sufficiently small* δ.

Assumption 3B is only used in the proof of asymptotic normality of the estimator (3). When both assumption 3A and 3B are imposed, we will assume that they hold for the same constant $\tau > 0$. Assumption 3B requires that the empirical processes indexed by coordinates of the class of vector-valued functions $\{\partial_\theta \ell(\theta, \cdot) : \ \|\theta - \theta_0\| \leqslant \delta\}$ are asymptotically continuous at θ_0, plus an additional mild integrability condition on the envelope function. It is well known van der Vaart and Wellner (1996) that it holds for classes that satisfy uniform entropy bounds such as VC-subgraph classes (van der Vaart and Wellner, 1996) as well as classes that are Hölder continuous in parameter, meaning that $\|\partial_\theta \ell(\theta', x) - \partial_\theta \ell(\theta'', x)\| \leqslant \widetilde{M}_{\theta_0}(x) \|\theta' - \theta''\|^\alpha$ for some $\alpha > 0$ and all θ', θ'' in a neighborhood of θ_0, where $\mathbb{E}\left|\widetilde{M}_{\theta_0}^{2+\tau}(X)\right| < \infty$.

Assumption 4. *Given* $t, R > 0$ *and a positive integer* n, *define*

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

Then $\lim_{R \to \infty} \limsup_{n \to \infty} B(n, R, t) = 0$ *for some* $t > 0$.

Conditions similar to assumption 4 are commonly imposed to allow treatment of non-compact parameter spaces in the framework of M-estimation; see Van der Vaart (2000). When Θ is compact, assumption 4 holds automatically. To illustrate assumption 4, consider the framework of linear regression, where the data consist of i.i.d. copies of the random couple $(Z, Y) \in \mathbb{R}^d \times \mathbb{R}$ such that $Y = \langle \theta_*, Z \rangle + \varepsilon$ for some $\theta_* \in \mathbb{R}^d$ and a noise variable ε that is independent of Z and has variance σ^2. Moreover, assume that Z is centered and has positive definite covariance matrix Σ. It is easy to see that $\frac{1}{n} \sum_{j=1}^n \ell(\theta, Z_j, Y_j) = \frac{1}{n} \left(\|\vec{\varepsilon}\|^2 + \|\mathbb{Z}(\theta - \theta_*)\|^2 - 2 \langle \vec{\varepsilon}, \mathbb{Z}(\theta_* - \theta) \rangle \right)$, where $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and $\mathbb{Z} \in \mathbb{R}^{n \times d}$ has $Z_1, \ldots, Z_n$ as rows. Cauchy-Schwarz inequality combined with a simple relation $2|ab| \leqslant a^2/2 + 2b^2$ that holds for all $a, b \in \mathbb{R}$ yield that

$$\frac{1}{n} \sum_{j=1}^n \ell(\theta, Z_j, Y_j) \geqslant \frac{1}{2n} \|\mathbb{Z}(\theta - \theta_*)\|^2 - \frac{1}{n} \|\vec{\varepsilon}\|^2,$$

hence $\inf_{\|\theta - \theta_*\| \geqslant R} \frac{1}{n} \sum_{j=1}^n \ell(\theta, Z_j, Y_j) \geqslant \frac{R^2}{2} \inf_{\|u\|=1} \langle \Sigma_n u, u \rangle - \frac{1}{n} \|\vec{\varepsilon}\|^2$ where $\Sigma_n = \frac{1}{n} \sum_{j=1}^n Z_j Z_j^T$ is the sample covariance matrix. Since $\inf_{\|u\|=1} \langle \Sigma_n u, u \rangle \geqslant \lambda_{\min}(\Sigma) - \|\Sigma_n - \Sigma\| = \lambda_{\min}(\Sigma) - o_P(1)$ and $\frac{1}{n} \|\vec{\varepsilon}\|^2 = O_p(1)$, it is easy to conclude that assumption 4 holds.

We are ready to state the main results regarding consistency and asymptotic normality of $\widehat{\theta}_N$. Recall the adversarial contamination framework defined in section 1. It the statements below, we assume that the sequences $\{k_j\}_{j \geqslant 1}$ and $\{n_j\}_{j \geqslant 1}$ corresponding the the number of subgroups and their cardinality respectively are non-decreasing and converge to ∞, and that the total sample size is $N_j = k_j n_j$.

Theorem 1. *Let assumptions 1, 2, 3A and 4 be satisfied, where we take $\tau = 0$ in assumption 3A. Suppose that the contamination proportion $\kappa_j = \frac{\mathcal{O}_j}{N_j}$ satisfies $\kappa_j = o\left(\frac{1}{n_j^{1/2}\Delta_{n_j}}\right)$ and that $\limsup_{j\to\infty} \frac{\mathcal{O}_j}{k_j} \leqslant c$ for a sufficiently small constant c.*

1. *If moreover $\mathrm{Var}\left(\ell(\theta, X)\right) < \infty$ for all $\theta \in \Theta$, then the estimator $\widehat{\theta}_{N_j}$ defined in (3) is consistent: $\widehat{\theta}_{N_j} \to \theta_0$ in probability as $j \to \infty$.*
2. *Under no additional assumptions, $\widehat{\theta}^{(1)}_{N_j} \to \theta_0$ and $\widehat{\theta}^{(2)}_{N_j} \to \theta_0$ in probability as $j \to \infty$, where $\left(\widehat{\theta}^{(1)}_{N_j}, \widehat{\theta}^{(2)}_{N_j}\right)$ were defined in (4).* [1]

This result is essentially a corollary of the fact that under the stated assumptions, $\sup_{\theta\in\Theta'}\left|\widehat{L}(\theta) - L(\theta)\right| \to 0$ over compact subsets $\Theta' \subseteq \Theta$ (note however that we do not require Θ to be compact). Note that we impose an extra condition $\mathrm{Var}\left(\ell(\theta, X)\right) < \infty,\ \theta \in \Theta$ (in fact, it suffices to require this some $\theta' \in \Theta$) to obtain consistency of $\widehat{\theta}_{N_j}$; this assumption seems to be unavoidable and has been imposed in the prior work on the similar topic (Brownlees et al., 2015). The following statements constitute the main contribution of the paper.

Theorem 2. *Assume that the the sample is free of adversarial outliers ($\kappa = 0$). Let assumptions 1, 2, 3A and 4 be satisfied, where we take $\tau = 0$ in assumption 3A. Then*

$$\sqrt{N_j}\left(\widehat{\theta}^{(i)}_{N_j} - \theta_0\right) \xrightarrow{d} N\left(0, D^2(\theta_0)\right) \ \text{ as } j \to \infty,\ \ i = 1, 2,$$

where $D^2(\theta_0) = \left[\partial^2_\theta L(\theta_0)\right]^{-1} \Sigma \left[\partial^2_\theta L(\theta_0)\right]^{-1}$ and $\Sigma = \mathbb{E}\left[\partial_\theta \ell(\theta_0, X)\partial_\theta \ell(\theta_0, X)^T\right]$.

In essence, this result establishes that when the sample is free of adversarial corruption, no loss of asymptotic efficiency occurs if the standard M-estimator based on empirical risk minimization Van der Vaart (2000) is replaced by its robust counterpart $\widehat{\theta}^{(1)}_N$, as their asymptotic covariances are equal. For example, maximum likelihood estimation corresponds to the case when $\{p_\theta,\ \theta \in \Theta\}$ is a family of probability density functions with respect to some σ-finite measure μ and $\ell(\theta, \cdot) = -\log p_\theta(\cdot)$. If $-\partial^2_\theta \mathbb{E} \log p_{\theta_0}(X) = I(\theta_0) := \mathbb{E}\left[\partial_\theta \log p_{\theta_0}(X)\partial_\theta \log p_{\theta_0}(X)^T\right]$, then it follows that $\widehat{\theta}^{(i)}_{N_j},\ i = 1, 2$ are asymptotically equivalent to the maximum likelihood estimator. However, as the next result shows, the situation is different if we use the estimator $\widehat{\theta}_N$ instead.

Theorem 3. *Suppose that assumptions 1 – 4 hold where $\tau > 0$ in assumptions 3A and 3B. Moreover, suppose that $k_j = o(n_j^\tau)$ as $j \to \infty$. Then the estimator $\widehat{\theta}_{N_j}$ defined in (3) satisfies*

$$\sqrt{N_j}\left(\widehat{\theta}_{N_j} - \theta_0\right) \xrightarrow{d} N\left(0, V^2(\theta_0, \rho)\right) \ \text{ as } k, n \to \infty,$$

where $V^2(\theta_0, \rho) = \left[\partial^2_\theta L(\theta_0)\right]^{-1} A^2(\theta_0, \rho) \left[\partial^2_\theta L(\theta_0)\right]^{-1}$ and

$$A^2(\theta_0, \rho) = \left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^{-2} \left(\mathbb{E}\left[\left(\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2 Z_2 Z_2^T\right] \right.$$
$$\left. + \frac{1}{\Delta_\infty^2} \frac{\left(\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2 \mathbb{E}\left(\rho'\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}{\left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2} \Gamma(\theta_0)\Gamma^T(\theta_0)\right)$$

[1] Our bounds are strong enough to deduce convergence with probability 1, however, we omit the self-evident arguments as weaker results suffices to prove asymptotic normality.

$$+ \frac{2}{\Delta_\infty} \frac{\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\mathbb{E}\left[\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_1\right]}{\mathrm{Var}(Z_1)\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)}\Gamma(\theta_0)\Gamma^T(\theta_0)\Bigg).$$

Here, $\Delta_\infty = \lim_{j\to\infty}\Delta_{n_j}$, $\Gamma(\theta_0) = \mathbb{E}\left[\ell(\theta_0,X)\,\partial_\theta\ell(\theta_0,X)\right] \in \mathbb{R}^d$, *and* $(Z_2, Z_1) \in \mathbb{R}^d \times \mathbb{R}$ *is a centered multivariate normal vector with covariance matrix*

$$\Sigma(\theta_0) = \begin{pmatrix} \mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\,\partial_\theta\ell(\theta_0,X)^T\right] & \Gamma(\theta_0) \\ \Gamma(\theta_0)^T & \mathrm{Var}\left(\ell(\theta_0,X)\right)\end{pmatrix}.$$

Let us discuss this result. First of all, the requirement $k_j = o(n_j^\tau)$ is imposed to guarantee that the bias of the estimator $\widehat{L}(\theta)$ is of order $o(N_j^{-1/2})$; this condition "compensates" for the fact that estimator $\widehat{L}(\theta)$ is not linear with respect to $\ell(\theta,\cdot)$. The expression for the asymptotic variance in the previous result is somewhat hard to parse. Let us mention several examples where $A^2(\theta_0,\rho)$ takes a simple form

$$A^2(\theta_0,\rho) = \frac{\mathbb{E}\left(\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}{\left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}\mathbb{E}\left[Z_2Z_2^T\right],$$

while the second and third terms in the expression of $A^2(\theta_0,\rho)$ vanish. The most obvious situation occurs when the $\lim_{j\to\infty}\Delta_{n_j} = \infty$, whence $A^2(\theta_0,\rho) = \mathbb{E}\left[Z_2Z_2^T\right]$ and $\widehat{\theta}_N$ is asymptotically equivalent to the standard M-estimator $\widetilde{\theta}_N$. However, letting $\Delta_n \to \infty$ causes the estimator to be less robust, as indicated by Theorem 1, which makes this example less interesting. The second scenario corresponds to the case when $\Gamma(\theta_0) = 0$. One example that fits this scenario is linear regression where $X = (Z,Y) \in \mathbb{R}^d \times \mathbb{R}$, Z is a centered random vector and and $Y = \langle\theta_*, Z\rangle + \varepsilon$ for some $\theta_* \in \mathbb{R}^d$ and a noise variable ε that is independent of Z. Another common example is maximum likelihood estimation of the location parameter. Here, $\ell(\theta,x) = -\log f(x-\theta)$, $\theta \in \mathbb{R}^d$, where f is a probability density functions with respect to some shift-invariant measure μ, and X has density $f(x-\theta_0)$ for $\theta_0 \in \mathbb{R}^d$. Assuming that f is sufficiently regular, it is easy to check that $\Gamma(\theta_0)$ is equal to the gradient of the function $E(\theta) = \int_{\mathbb{R}^d}\log f(x-\theta)\cdot f(x-\theta)d\mu(x)$, which is simply the entropy corresponding to the density $f(x-\theta)$. If μ is shift-invariant, the entropy in the location family is constant, hence the conclusion follows.

3. Proofs.

The strategies of the proofs of Theorems 2 and 3 are different: the proof of Theorem 2 uses characterization of $\left(\widehat{\theta}_N^{(1)}, \widehat{\theta}_N^{(2)}\right)$ as the minimizer/maximizer of the risk, and follows a standard pattern of consequently establishing consistency, rate of convergence and finally the asymptotic normality. The arguments are quite general and can be extended beyond the classes that satisfy Lipschitz property imposed by assumption 3A. The proof on Theorem 3 on the other hand is based on the treating $\widehat{\theta}_N$ as a "Z-estimator" solving the estimating equation $\partial_\theta\widehat{L}(\widehat{\theta}_N) = 0$. Required reasoning is slightly more technical, and the Lipschitz property of the class plays an important role in controlling the product empirical process arising in the course of the proof.

3.1. Technical tools.

We recall some of the basic facts and existing results that our proofs often rely upon. 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 metric ρ if for all $t_1, t_2 \in \mathbb{T}$ and $s \in \mathbb{R}$,

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

Fact 1 (Dudley's entropy bound)**.** *Let $\{Y(t),\ t \in T\}$ be a centered stochastic process with sub-Gaussian increments. Then the following inequality holds:*

$$\mathbb{E}\sup_{t\in T}|Y(t)-Y(t_0)| \leqslant 12\int_0^{D(T)}\sqrt{\log N(T,\rho,\varepsilon)}d\varepsilon,$$

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

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

Lemma 1. *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 proof follows in a straightforward manner from the triangle inequality and is omitted. □

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

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

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

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

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

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

The proof of this bound is mimics the argument behind Theorem 3.1 in (Minsker, 2019b); we outline the necessary modifications in section 3.7. For the illustration purposes, assume that $\mathcal{O} = 0$, whence the result above implies that as long as $\mathbb{E}\sup_{\theta\in\Theta'}\frac{1}{\sqrt{N}}\sum_{j=1}^{N}|\ell(\theta,X_j)) - L(\theta)| = O(1)$ and $\sigma(\Theta') \lesssim \Delta_n = O(1)$,

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

Moreover, if $\mathcal{O} = \kappa N$, then, setting $k \asymp N\kappa^{\frac{2}{2+\tau}}$, we see that $\sup_{\theta\in\Theta'}\left|\widehat{L}(\theta)-L(\theta)\right| = O_p\left(N^{-1/2}+\kappa^{\frac{1+\tau}{2+\tau}}\right)$.

3.2. Proof of Theorem 1.

We will first establish the claim of Theorem 1. Observe that

$$\begin{aligned}L(\widehat{\theta}_N) &= L(\widehat{\theta}_N) - \widehat{L}(\widehat{\theta}_N) + \widehat{L}(\widehat{\theta}_N) \leqslant L(\widehat{\theta}_N) - \widehat{L}(\widehat{\theta}_N) + \widehat{L}(\theta_0) \\ &\leqslant L(\theta_0) + \left(L(\widehat{\theta}_N) - \widehat{L}(\widehat{\theta}_N)\right) + \left(\widehat{L}(\theta_0) - L(\theta_0)\right).\end{aligned}$$

Given $\varepsilon > 0$, assumption 2 implies that there exists $\delta > 0$ such that $\inf_{\|\theta-\theta_0\|\geqslant\varepsilon} L(\theta) > L(\theta_0)+\delta$ (in fact, this inequality holds under weaker condition than assumption 2). Therefore, for any $R > 0$

$$\mathbb{P}\Big(\|\widehat{\theta}_N-\theta_0\|\geqslant\varepsilon\Big) \leqslant \mathbb{P}\left(\sup_{\theta\in\Theta:\|\theta-\theta_0\|\leqslant R}|\widehat{L}(\theta)-L(\theta)| > \delta/2\right) + \mathbb{P}\Big(\|\widehat{\theta}_N-\theta_0\|\geqslant R\Big).$$

It follows from Lemma 2 that

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

as long as $\limsup_{\|,n\to\infty}\frac{\mathcal{O}(k,n)}{k} \leqslant c$ and $\Delta_n\kappa\sqrt{n} \to 0$ as $n,k\to\infty$, where $\kappa = \frac{\mathcal{O}(k,n)}{N}$ is the proportion of outliers. To see this, observe that $\mathbb{E}\sup_{\theta\in\Theta:\|\theta-\theta_0\|\leqslant R}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\ell(\theta,X_j)-L(\theta)\right| \leqslant$

$\mathbb{E}\sup_{\theta\in\Theta:\|\theta-\theta_0\|\leqslant R}\frac{1}{\sqrt{N}}\sum_{j=1}^N\left(\ell(\theta,X_j)-\ell(\theta_0,X_j)-(L(\theta)-L(\theta_0))\right)+\sqrt{\mathrm{Var}\left(\ell(\theta_0,X)\right)}$. Therefore, it suffices to show that

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

To this end, we use a well-known argument based on symmetrization inequality and Dudley's entropy integral bound (see Fact 1). Let $\varepsilon_1,\dots,\varepsilon_N$ be i.i.d. random signs, independent of the data $X_1,\dots,X_N$. Then the symmetrization inequality (van der Vaart and Wellner, 1996) yields that

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

Conditionally on $X_1,\dots,X_N$, 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 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 4 that there exist $\theta_1,\dots,\theta_{N(R)}$ such that $\bigcup_{j=1}^{N(R)}B(\theta_j,r(\theta_j))\supseteq B(\theta_0,R)$ and

$$|\ell(\theta',x)-\ell(\theta'',x)|\leqslant M_{\theta_j}(x)\|\theta'-\theta''\|$$

for all $\theta',\theta''\in B(\theta_j,r(\theta_j))$, where $M_{\theta_j}(x)$ is defined in assumption 3A. 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 $\left(\frac{6r(\theta_j)\|M_{\theta_j}\|_{L_2(P_N)}}{\tau}\right)^d$ balls of radius τ. Therefore,

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

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

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

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

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

and

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

It remains to establish that $\|\widehat{\theta}_N-\theta_0\|\leqslant R$ for R large enough with probability close to 1. Recall that

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

Assumption 4 states that $\lim_{R\to\infty}\limsup_{n\to\infty}B(n,R,t)=0$ for some $t>0$. In particular, one can choose R_0 and n_0 such that $B(n_0,R_0,t)<0.01$ for all $n\geqslant n_0$ and $R\geqslant R_0$. As

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

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

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

leading to a contradiction. Therefore,

$$\begin{aligned}\mathbb{P}&\left(\inf_{\|\theta-\theta_0\|\geqslant R}\widehat{L}(\theta)<L(\theta_0)+t/2\right)\\&\leqslant\mathbb{P}\left(\exists J\subseteq\{1,\ldots,k\},\ |J|\geqslant k/2:\ \inf_{\|\theta-\theta_0\|\geqslant R}\bar{L}_j(\theta)<L(\theta_0)+t/2+2\frac{\Delta_n}{\sqrt{n}},\ j\in J\right)\\&\leqslant\binom{k}{\lfloor k/2\rfloor}(B(n,R,t))^{k/2}\leqslant(2e)^{k/2}(B(n,R,t))^{k/2}\end{aligned}\tag{5}$$

whenever $2\frac{\Delta_n}{\sqrt{n}}\leqslant t/2$. Moreover, if $n\geqslant n_0$ and $R\geqslant R_0$, we deduce that $(2e)^{k/2}(B(n,R,t))^{k/2}<0.25^k\to 0$ as $k\to\infty$. As $\widehat{L}(\theta_0)\to L(\theta_0)$ P-almost surely, the preceding display implies that $\mathbb{P}\left(\|\widehat{\theta}_N-\theta_0\|<R\right)\to 1$ as $n,k,R\to\infty$.

The proof of of the second claim of the theorem proceeds in a similar way, with some modifications. For every $\theta'\in\Theta$, define $\widehat{\theta}(\theta'):=\operatorname{argmax}_{\theta\in\Theta}\widehat{L}(\theta',\theta)$. Observe that $\widehat{L}\left(\widehat{\theta}_N^{(1)},\widehat{\theta}_N^{(2)}\right)\leqslant\widehat{L}\left(\theta_0,\widehat{\theta}(\theta_0)\right)$, hence, whenever $\|\widehat{\theta}_N^{(j)}-\theta_0\|\leqslant R,\ j=1,2$,

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

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

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

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

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

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

Similarly, observe that by the definition of $\widehat{\theta}_N^{(2)}$, whenever $\|\widehat{\theta}_N^{(j)} - \theta_0\| \leqslant R,\ j = 1,2$ we have that

$$L(\widehat{\theta}_N^{(1)}) - L(\widehat{\theta}_N^{(2)}) = L(\widehat{\theta}_N^{(1)}) - L(\widehat{\theta}_N^{(2)}) \pm \widehat{L}(\widehat{\theta}_N^{(1)}, \widehat{\theta}_N^{(2)}) \geqslant \underbrace{\widehat{L}(\widehat{\theta}_N^{(1)}, \widehat{\theta}_N^{(1)})}_{=0} \\ - \sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1,\theta_2) - (L(\theta_1) - L(\theta_2))\right|.$$

In view of assumption 2, for any $\varepsilon > 0$,

$$\inf_{\theta_1} \sup_{\|\theta_2-\theta_0\|\geqslant\varepsilon} (L(\theta_1) - L(\theta_2)) < L(\theta_0) - L(\theta_0) - \delta = -\delta.$$

Hence,

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

Following the standard steps for bounding the expected supremum outlined in the fist part of the proof, it is easy to show that

$$\mathbb{E} \sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1,\theta_2) - (L(\theta_1) - L(\theta_2))\right| \to 0 \text{ as } k, n \to \infty,$$

so that $\mathbb{P}\Big(\sup_{\|\theta_j-\theta_0\|\leqslant R, j=1,2} \left|\widehat{L}(\theta_1,\theta_2) - (L(\theta_1) - L(\theta_2))\right| > \delta/2\Big) \to 0$. It remains to establish that $\mathbb{P}\Big(\left\|\widehat{\theta}_N^{(1)} - \theta_0\right\| > R \text{ or } \left\|\widehat{\theta}_N^{(2)} - \theta_0\right\| > R\Big) \to 0$. To this end, notice that by the definition of $\widehat{\theta}_N^{(1)}$,

$$0 \leqslant \widehat{L}(\widehat{\theta}_N^{(1)}, \widehat{\theta}_N^{(2)}) \leqslant \widehat{L}\left(\theta_0, \widehat{\theta}(\theta_0)\right) \leqslant \underbrace{L(\theta_0) - L\left(\widehat{\theta}(\theta_0)\right)}_{\leqslant 0} + \sup_{\|\theta-\theta_0\|\leqslant R} \left|\widehat{L}(\theta_0,\theta) - (L(\theta_0) - L(\theta))\right|$$

on the event $\left\{\|\widehat{\theta}(\theta_0)-\theta_0\| \leqslant R\right\}$. Repeating the steps presented in the first part of the proof for the class $\{\theta \mapsto \ell(\theta,\cdot)-\ell(\theta_0,\cdot),\ \theta\in\Theta\}$, it is easy to see that $\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta_0,\theta)-(L(\theta_0)-L(\theta))\right| \to 0$ and $\mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|>R\Big)\to 0$ for R large enough and as $n,k\to\infty$, implying that

$$\widehat{L}(\widehat{\theta}^{(1)}_N,\widehat{\theta}^{(2)}_N)\to 0 \text{ in probability.} \tag{6}$$

On the other hand, by the definition of $\widehat{\theta}^{(2)}_N$, it holds that $\widehat{L}(\widehat{\theta}^{(1)}_N,\widehat{\theta}^{(2)}_N)\geqslant \widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)$. Now, assume that $\|\widehat{\theta}^{(1)}_N-\theta_0\|>R$ while $\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)<L(\theta_0)+t/2-L(\theta_0)=t/2$. Arguing as in the first part of the proof, we see that there exists $J'\subset\{1,\dots,k\}$ such that $|J'|>k/2$ and $\bar{L}_j(\widehat{\theta}^{(1)}_N)-\bar{L}_j(\theta_0)< L(\theta_0)+t/2-L(\theta_0)+2\frac{\Delta_n}{\sqrt{n}}$ 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, $\mathbb{P}\left(\left|\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right|\geqslant t/4\right)\leqslant \frac{16}{nt^2}\mathrm{Var}\left(\ell(\theta_0,X)\right)$, therefore, for n and R large enough, $\mathbb{P}\Big(\inf_{\|\theta-\theta_0\|>R}\bar{L}_j(\theta)<L(\theta_0)+t/2+2\frac{\Delta_n}{\sqrt{n}}+\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\Big)<0.01$ for any j. Reasoning as in (5), we see that

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

We deduce that on the one hand,

$$\mathbb{P}\Big(\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)\geqslant t/2\bigcap\|\widehat{\theta}^{(1)}_N-\theta_0\|>R\Big)\to\mathbb{P}\Big(\|\widehat{\theta}^{(1)}_N-\theta_0\|>R\Big).$$

In view of (6), we see that on the other hand,

$$\mathbb{P}\Big(\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)\geqslant t/2\bigcap\|\widehat{\theta}^{(1)}_N-\theta_0\|>R\Big)\leqslant\mathbb{P}\Big(\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)\geqslant t/2\Big)\to 0,$$

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

Finally, assume that $\|\widehat{\theta}^{(2)}_N-\theta_0\|>R$ and that $\widehat{L}(\widehat{\theta}^{(1)}_N,\widehat{\theta}^{(2)}_N)>L(\widehat{\theta}^{(1)}_N)-L(\theta_0)-t/2$. Repeating the reasoning behind (5), we see that the latter implies that there exists $J'\subset\{1,\dots,k\}$ such that $|J'|>k/2$ and $\bar{L}_j(\widehat{\theta}^{(1)}_N)-\bar{L}_j(\widehat{\theta}^{(2)}_N)>L(\widehat{\theta}^{(1)}_N)-\Big(L(\theta_0)+t/2+2\frac{\Delta_n}{\sqrt{n}}\Big)$ for $j\in J'$, yielding that on the event $\left\{\|\widehat{\theta}^{(1)}_N-\theta_0\|\leqslant R\right\}$,

$$\begin{aligned}\inf_{\|\theta-\theta_0\|>R}\bar{L}_j(\theta)&<L(\theta_0)+t/2+2\frac{\Delta_n}{\sqrt{n}}+\Big(\bar{L}_j(\widehat{\theta}^{(1)}_N)-L(\widehat{\theta}^{(1)}_N)\Big)\\&\leqslant L(\theta_0)+t/2+2\frac{\Delta_n}{\sqrt{n}}+\sup_{\|\theta'-\theta_0\|\leqslant R}\left|\bar{L}_j(\theta')-L(\theta')\right|\end{aligned}$$

for $j\in J'$. We have shown before that $\mathbb{P}\Big(\|\widehat{\theta}^{(1)}_N-\theta_0\|>R\Big)\to 0$ 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 $n\to\infty$, for n and R large enough, the argument similar to (5) implies that

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

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

$$\mathbb{P}\Big(\widehat{L}\left(\widehat{\theta}_N^{(1)},\widehat{\theta}_N^{(2)}\right)\leqslant L(\widehat{\theta}_N^{(1)})-(L(\theta_0)+t/2)\Big)\leqslant\mathbb{P}\Big(\widehat{L}\left(\widehat{\theta}_N^{(1)},\theta_0\right)\leqslant L(\widehat{\theta}_N^{(1)})-(L(\theta_0)+t/2)\Big)$$

$$\mathbb{P}\left(L(\widehat{\theta}_N^{(1)})-L(\theta_0)-\sup_{\|\theta-\theta_0\|\leqslant R}\left|\widehat{L}(\theta,\theta_0)-(L(\theta)-L(\theta_0)\right|\leqslant L(\widehat{\theta}_N^{(1)})-(L(\theta_0)+t/2)\right)$$

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

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

3.3. Proof of Theorem 2.

The proof is divided into two steps. The first step consists in establishing the fact that the estimators $\widehat{\theta}_N^{(1)},\widehat{\theta}_N^{(1)}$ converge to θ_0 at $\sqrt{N}$-rate, while on the second step, we prove asymptotic normality by "zooming" to the resolution level $N^{-1/2}$; this proof pattern is quite standard in the empirical process theory (van der Vaart and Wellner, 1996).

We present a detailed argument establishing the convergence rate for $\widehat{\theta}_N^{(1)}$, and outline the modifications necessary to establish the result for $\widehat{\theta}_N^{(2)}$. Our goal is to show that

$$\lim_{M\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\Big(\sqrt{N}\|\widehat{\theta}_N^{(1)}-\theta_0\|\geqslant 2^M\Big)=0. \tag{7}$$

Define $S_{N,j}:=\big\{\theta:\ 2^{j-1}/\sqrt{N}<\|\theta-\theta_0\|\leqslant 2^j/\sqrt{N}\big\}$, $\bar{S}_{N,j}:=\big\{\theta:\ 0\leqslant\|\theta-\theta_0\|\leqslant 2^j/\sqrt{N}\big\}$, and observe that

$$\sqrt{N}\|\widehat{\theta}^{(1)}-\theta_0\|\geqslant 2^M\implies\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\widehat{\theta}(\theta))-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))\Big)\leqslant 0\text{ for some }j>M,$$

where $\widehat{\theta}(\theta'):=\operatorname{argmax}_{\theta\in\Theta}\widehat{L}(\theta',\theta)$. As $\widehat{L}(\theta,\widehat{\theta}(\theta))\geqslant\widehat{L}(\theta,\theta_0)$ for any θ, the inequality $\sqrt{N}\|\widehat{\theta}^{(1)}-\theta_0\|\geqslant 2^M$ implies that $\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0)\Big)\leqslant 0$ for some $j>M$, which in turn entails that

$$\inf_{\theta\in S_{N,j}}\Big(\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))+L(\theta_0,\widehat{\theta}(\theta_0))\Big)\leqslant L(\theta_0,\widehat{\theta}(\theta_0))-\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)$$

for some $j>M$. Since $L(\theta_0,\widehat{\theta}(\theta_0))-\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)\leqslant 0$ by the definition of θ_0, the previous display yields that

$$\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)-\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))+L(\theta_0,\widehat{\theta}(\theta_0))\right|$$
$$\geqslant\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)-L(\theta_0,\widehat{\theta}(\theta_0))\geqslant\inf_{\theta\in S_{N,j}}L(\theta,\theta_0),$$

which further implies that either

$$\sup_{\theta\in S_{N,j}}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right|\geqslant\inf_{\theta\in S_{N,j}}\frac{L(\theta,\theta_0)}{2},\text{ or }\left|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))-L(\theta_0,\widehat{\theta}(\theta_0))\right|\geqslant\inf_{\theta\in S_{N,j}}\frac{L(\theta,\theta_0)}{2}.$$

Let $0 < \eta_1 \leqslant r(\theta_0)$ be small enough so that $L(\theta) - L(\theta_0) \geqslant c\|\theta-\theta_0\|^2$ for θ such that $\|\theta-\theta_0\| \leqslant \eta_1$ (existence of η_1 follows from the assumption 2), and observe that $\mathbb{P}\left(\|\widehat{\theta}_N^{(1)} - \theta_0\| \geqslant \eta_1\right) \to 0$ as $n, k \to \infty$ due to consistency of the estimator. We then have

$$\mathbb{P}\left(\sqrt{N}\|\widehat{\theta}_N^{(1)} - \theta_0\| \geqslant 2^M\right) \leqslant \mathbb{P}\left(\sqrt{N}\left|\widehat{L}(\theta_0, \widehat{\theta}(\theta_0)) - L(\theta_0, \widehat{\theta}(\theta_0))\right| \geqslant c\frac{2^{2M}}{\sqrt{N}}\right)$$
$$+ \mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant \eta_1} \sup_{\theta\in S_{N,j}} \sqrt{N}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \geqslant c\frac{2^{2j-2}}{\sqrt{N}}\right) + \mathbb{P}\left(\|\widehat{\theta}_N^{(1)} - \theta_0\| \geqslant \eta_1\right). \quad (8)$$

To estimate $\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1} \sup_{\theta\in S_{N,j}} \sqrt{N}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \geqslant c\frac{2^{2j-2}}{\sqrt{N}}\right)$, we invoke Proposition 2 applied to the class $\left\{\ell(\theta,\cdot) - \ell(\theta_0,\cdot),\ \theta\in\bar{S}_{N,j}\right\}$. Since $|\ell(\theta,x) - \ell(\theta_0,x)| \leqslant M_{\theta_0}(x)\frac{2^j}{\sqrt{N}}$, it is easy to see that $\sigma^2(\delta) \leqslant \mathbb{E}M^2_{\theta_0}(X)\frac{2^{2j}}{N}$. Together with the union bound applied over $M < j \leqslant J_{\max} := \lfloor \log(\sqrt{N}\eta_1)\rfloor + 1$ with $s_j := j^2$, it implies that for all $\theta\in S_{N,j}$, $M+1 \leqslant j \leqslant J_{\max}$,

$$\sqrt{N}\left(\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right)$$
$$= \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right)}\frac{1}{\sqrt{k}}\sum_{i=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right)$$
$$+ \mathcal{R}_{n,k,j}(\theta), \quad (9)$$

where

$$\sup_{\theta\in\bar{S}_{N,j}} |\mathcal{R}_{n,k,j}(\theta)| \leqslant C(d,\theta_0)\left(\frac{2^{2j}}{N}\frac{j^4}{\sqrt{k}} + \frac{2^{3j}j^2}{N^{3/2}} + \sqrt{k}\frac{2^{6j}}{N^3}\right)$$

uniformly over all $M \leqslant j \leqslant J_{\max}$ with probability at least $1 - 3\sum_{j:j\geqslant M+1} j^{-2} \geqslant 1 - \frac{C}{M}$. Let $\mathcal{E}$ denote the event of probability at least $1 - \frac{C}{M}$ on which the previous representation holds. Moreover, observe that, in view of Lemma 5, for η_1 small enough and N large enough,

$$\sup_{\|\theta-\theta_0\|\leqslant\eta_1}\left|\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right) - \rho''(0)\right| \leqslant \frac{\rho''(0)}{2} = \frac{1}{2}.$$

Taking this fact into account and noting that $\frac{2^j}{\sqrt{N}}\frac{j^4}{\sqrt{k}} + \frac{2^{2j}j^2}{N} + \sqrt{k}\frac{2^{5j}}{N^{3/2}} \leqslant c2^j$ for any $j \leqslant J_{\max}$ and any $c > 0$ given that n is large enough (in particular, it implies that the remainder term $\mathcal{R}_{n,k,j}(\theta)$ is smaller than $\frac{c}{2}\frac{2^{2j-2}}{\sqrt{N}}$ on event $\mathcal{E}$), we deduce that

$$\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1} \sup_{\theta\in S_{N,j}} \sqrt{N}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \geqslant c\frac{2^{2j-2}}{\sqrt{N}}\right) \leqslant \frac{C}{M}$$
$$+ \sum_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1} \mathbb{P}\left(\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right)\right| \geqslant c_1\frac{2^{2j}}{\sqrt{N}}\right).$$

Invoking Lemma 5 again, we see that

$$\sup_{\theta\in\bar{S}_{N,j}}\left|\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - \bar{L}_1(\theta_0) - L(\theta,\theta_0)\right)\right)\right| \leqslant C\frac{2^{2j}}{N}.$$

Let us denote $\rho'_{n,i}(\theta,\theta_0) = \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_i(\theta) - \bar{L}_i(\theta_0) - L(\theta,\theta_0)\right)\right)$, $i = 1,\ldots,k$ for brevity. As $\sqrt{k}\frac{2^{2j}}{N} \leqslant c'\frac{2^{2j}}{\sqrt{N}}$ for any $c' > 0$ and sufficiently large n,

$$\begin{aligned}\mathbb{P}\Bigg(\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\rho'_{n,i}(\theta,\theta_0)\right| \geqslant c_1\frac{2^{2j}}{\sqrt{N}}\Bigg) \\ \leqslant \mathbb{P}\Bigg(\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\rho'_{n,i}(\theta,\theta_0) - \mathbb{E}\rho'_{n,i}(\theta,\theta_0)\right)\right| \geqslant c_2\frac{2^{2j}}{\sqrt{N}}\Bigg) \\ \leqslant \frac{\sqrt{N}}{c_2 2^{2j}}\mathbb{E}\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\rho'_{n,i}(\theta,\theta_0) - \mathbb{E}\rho'_{n,i}(\theta,\theta_0)\right)\right|\end{aligned}$$

where we used Markov's inequality on the last step. To bound the expected supremum, we proceed in exactly the same fashion (using symmetrization, contraction and desymmetrization inequalities) as in the proof of Proposition 1, and deduce that

$$\begin{aligned}\mathbb{E}\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\rho'_{n,i}(\theta,\theta_0) - \mathbb{E}\rho'_{n,i}(\theta,\theta_0)\right)\right| \\ \leqslant \frac{C}{\Delta_n}\mathbb{E}\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(\theta,X_j) - \ell(\theta_0,X_j) - L(\theta,\theta_0)\right)\right|.\end{aligned}$$

The right side of the display above can be bounded by $\frac{C(d,\theta_0)}{\Delta_n}\frac{2^j}{\sqrt{N}}$ (using Lemma 8), implying that

$$\mathbb{P}\Bigg(\sup_{\theta\in S_{N,j}}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\left(\rho'_{n,i}(\theta,\theta_0) - \mathbb{E}\rho'_{n,i}(\theta,\theta_0)\right)\right| \geqslant c_2\frac{2^{2j}}{\sqrt{N}}\Bigg) \leqslant \frac{C(d,\theta_0)}{\Delta_n}\frac{1}{2^j},$$

whence

$$\begin{aligned}\mathbb{P}\Bigg(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_1}\sup_{\theta\in S_{N,j}}\sqrt{N}\left|\widehat{L}(\theta,\theta_0) - L(\theta,\theta_0)\right| \geqslant c\frac{2^{2j-2}}{\sqrt{N}}\Bigg) \\ \leqslant \frac{C}{M} + \frac{C(d,\theta_0)}{\Delta_n}\sum_{j\geqslant M}2^{-j} \leqslant \frac{C}{M} + \frac{C(d,\theta_0)}{\Delta_n}2^{-M+1} \to 0 \text{ as } M\to\infty\end{aligned}$$

whenever n,k are large enough. In view of (8), it only remains to show that

$$\mathbb{P}\Bigg(\sqrt{N}\left|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0)) - L(\theta_0,\widehat{\theta}(\theta_0))\right| \geqslant c\frac{2^{2M}}{\sqrt{N}}\Bigg) \to 0 \text{ as } n,k\to\infty. \tag{10}$$

To this end, it suffices to repeat the argument presented above, with several simplifications. First, we will start by proving that $\lim_{M\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\left(\sqrt{N}\|\widehat{\theta}(\theta_0) - \theta_0\| \geqslant 2^M\right) = 0$. We have already shown (in the course of the proof of Theorem 1) that $\widehat{\theta}(\theta_0)$ is a consistent estimator of θ_0, so that $\mathbb{P}\left(\|\widehat{\theta}(\theta_0) - \theta_0\| \geqslant \eta_2\right) \to 0$ for any $\eta_2 > 0$. If $\sqrt{N}\|\widehat{\theta}(\theta_0) - \theta_0\| \geqslant 2^M$, then $\widehat{\theta}(\theta_0) \in S_{N,j}$ for some $j > M$, implying that $\sup_{\theta\in S_{N,j}}\widehat{L}(\theta_0,\theta) \geqslant \widehat{L}(\theta_0,\theta_0) = 0$, which entails the inequality $\sup_{\theta\in S_{N,j}}\left(\widehat{L}(\theta_0,\theta) - L(\theta_0,\theta)\right) \geqslant -\sup_{\theta\in S_{N,j}}L(\theta_0,\theta) = \inf_{\theta\in S_{N,j}}L(\theta,\theta_0) \geqslant c\frac{2^{2j-2}}{N}$ whenever $2^j/\sqrt{N} \leqslant \eta_2$ and η_2 is small enough. Therefore,

$$\mathbb{P}\Big(\sqrt{N}\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant 2^M\Big)\leqslant \mathbb{P}\Big(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant \eta_2\Big) + \mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_2}\sup_{\theta\in S_{N,j}}\sqrt{N}\left|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\right|\geqslant c\frac{2^{2j-2}}{\sqrt{N}}\right).$$

The probability of the union is estimated as before using Proposition 2, implying that it converges to 0 as $M\to\infty$. To complete the proof of (10), observe that

$$\mathbb{P}\left(\sqrt{N}\left|\widehat{L}(\theta_0,\widehat{\theta}(\theta_0))-L(\theta_0,\widehat{\theta}(\theta_0))\right|>c\frac{2^{2M}}{\sqrt{N}}\right)\leqslant \mathbb{P}\left(\|\widehat{\theta}(\theta_0)-\theta_0\|\geqslant\frac{2^M}{\sqrt{N}}\right) + \mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\frac{2^M}{\sqrt{N}}}\sqrt{N}\left|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\right|\geqslant c\frac{2^{2M}}{\sqrt{N}}\right)$$

and that $\mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\frac{2^M}{\sqrt{N}}}\sqrt{N}\left|\widehat{L}(\theta_0,\theta)-L(\theta_0,\theta)\right|\geqslant c\frac{2^{2M}}{\sqrt{N}}\right)\leqslant\frac{C}{M}+\frac{C(d,\theta_0)}{\Delta_n}2^{-M}\to 0$ as $M\to\infty$, which follows from the representation (9) in the same fashion as before. This completes the proof of relation (7). To prove that

$$\lim_{M\to\infty}\limsup_{n,k\to\infty}\mathbb{P}\Big(\sqrt{N}\|\widehat{\theta}^{(2)}_N-\theta_0\|\geqslant 2^M\Big)=0,$$

we begin by observing that the inequality $\sqrt{N}\|\widehat{\theta}^{(2)}_N-\theta_0\|\geqslant 2^M$ implies that $\sup_{\theta\in S_{N,j}}\widehat{L}(\widehat{\theta}^{(1)}_N,\theta)\geqslant\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)$ for some $j>M$. If $\frac{2^j}{\sqrt{N}}\leqslant\eta_3$ for sufficiently small constant $\eta_3>0$, we see that it further entails the inequality

$$\begin{aligned}\sup_{\theta\in S_{N,j}}&\Big(\widehat{L}(\widehat{\theta}^{(1)}_N,\theta)-L(\widehat{\theta}^{(1)}_N,\theta)-\widehat{L}(\widehat{\theta}^{(1)}_N,\theta_0)+L(\widehat{\theta}^{(1)}_N,\theta_0)\Big)\\&\geqslant-\sup_{\theta\in S_{N,j}}L(\widehat{\theta}^{(1)}_N,\theta)+L(\widehat{\theta}^{(1)}_N,\theta_0)=\inf_{\theta\in S_{N,j}}L(\theta,\widehat{\theta}^{(1)}_N)+L(\widehat{\theta}^{(1)}_N,\theta_0)\\&=\inf_{\theta\in S_{N,j}}L(\theta,\theta_0)\geqslant c\frac{2^{2j-2}}{N}.\end{aligned}$$

We deduce from the display above that

$$\begin{aligned}\mathbb{P}\Big(\sqrt{N}\|\widehat{\theta}^{(2)}_N-\theta_0\|\geqslant 2^M\Big)&\leqslant\mathbb{P}\Big(\|\widehat{\theta}^{(2)}_N-\theta_0\|\geqslant\eta_3\Big)+\mathbb{P}\Big(\sqrt{N}\|\widehat{\theta}^{(1)}_N-\theta_0\|\geqslant 2^{M/2}\Big)\\&+\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2^j}{\sqrt{N}}\leqslant\eta_3}\sup_{\theta\in S_{N,j},\theta'\in\bar{S}_{N,M/2}}\sqrt{N}\left|\widehat{L}(\theta',\theta)-L(\theta',\theta)\right|\geqslant c_1\frac{2^{2j-2}}{\sqrt{N}}\right)\\&+\mathbb{P}\left(\sup_{\theta\in\bar{S}_{N,M/2}}\sqrt{N}\left|\widehat{L}(\theta,\theta_0)-L(\theta,\theta_0)\right|\geqslant c_1\frac{2^{2M}}{\sqrt{N}}\right).\end{aligned}$$

We have previously shown that the first and second term on the right side of the previous display converge to 0 as M, n and k tend to infinity, while the last term converges to 0 in view of argument presented before (see representation (9) and the bounds that follow). It remains to estimate $\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2j}{\sqrt{N}}\leqslant\eta_3}\sup_{\theta\in S_{N,j},\theta'\in\bar S_{N,M/2}}\sqrt{N}\left|\widehat{L}(\theta',\theta)-L(\theta',\theta)\right|\geqslant c_1\frac{2^{2j-2}}{\sqrt{N}}\right)$. To this end, we again invoke Proposition 2 applied to the class $\left\{\ell(\theta_1,\cdot)-\ell(\theta_2,\cdot),\ \theta_1\in\bar S_{N,M/2},\theta_2\in\bar S_{N,j}\right\}$; here, the "reference point" is (θ_0,θ_0). Since $|\ell(\theta,x)-\ell(\theta',x)|\leqslant M_{\theta_0}(x)\frac{2^j+2^{M/2}}{\sqrt{N}}$, it is easy to see that $\sigma^2(\delta)\leqslant\mathbb{E}M^2_{\theta_0}(X)\frac{2^{2j+1}+2^M}{N}\leqslant C(\theta_0)\frac{2^{2j}}{N}$, and to deduce that

$$\begin{aligned}&\sqrt{N}\left(\widehat{L}(\theta',\theta)-L(\theta',\theta)\right)\\&\quad=\frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar L_1(\theta')-\bar L_1(\theta)-L(\theta',\theta)\right)\right)}\frac{1}{\sqrt{k}}\sum_{i=1}^k\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar L_i(\theta')-\bar L_i(\theta)-L(\theta',\theta)\right)\right)\\&\qquad+\mathcal{R}_{n,k,j}(\theta',\theta),\end{aligned}$$

where

$$\sup_{\theta\in\bar S_{N,j},\theta'\in\bar S_{N,M/2}}\left|\mathcal{R}_{n,k,j}(\theta',\theta)\right|\leqslant C(d,\theta_0)\left(\frac{2^{2j}}{N}\frac{j^4}{\sqrt{k}}+\frac{2^{3j}j^2}{N^{3/2}}+\sqrt{k}\frac{2^{6j}}{N^3}\right)$$

uniformly over all $M\leqslant j\leqslant J_{\max}$ with probability at least $1-\frac{C}{M}$. The remaining steps again closely mimic the argument outlined in detail after display (9) and yield that

$$\mathbb{P}\left(\bigcup_{j:j\geqslant M+1,\ \frac{2j}{\sqrt{N}}\leqslant\eta_3}\sup_{\theta\in S_{N,j},\theta'\in\bar S_{N,M/2}}\sqrt{N}\left|\widehat{L}(\theta',\theta)-L(\theta',\theta)\right|\geqslant c_1\frac{2^{2j-2}}{\sqrt{N}}\right)\leqslant\frac{C(d,\theta_0)}{\Delta_n}2^{-M+1}\to 0$$

as $M\to\infty$, therefore implying the last claim in the first part of the proof.

Now we are ready to establish asymptotic normality of $\widehat\theta^{(1)}_N$ and $\widehat\theta^{(2)}_N$. To this end, consider the stochastic process $M_N(h,q)$ indexed by $h,q\in\mathbb{R}^d$ and defined via

$$M_N(h,q):=N\left(\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))-L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})\right).$$

Below, we will show that $M_N(h,q)$ converges weakly to the Gaussian process $W(h,q):=W^T(h-q)$, $h,q\in\mathbb{R}^d$, where $W\sim N(0,\Sigma_W)$ and $\Sigma_W=\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\partial_\theta\ell(\theta_0,X)^T\right]$. Let us deduce the conclusion assuming that weak convergence has already been established. We have that

$$N\cdot\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))=N\cdot L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})+M_N(h,q).$$

Note that, in view of assumption 2 and the fact that θ_0 minimizes $L(\theta_0)$,

$$N\cdot L(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N})\to\frac{1}{2}h^T\partial^2_\theta L(\theta_0)h-\frac{1}{2}q^T\partial^2_\theta L(\theta_0)q\text{ as }N\to\infty,$$

therefore $N\cdot\widehat{L}(\theta_0+h/\sqrt{N},\theta_0+q/\sqrt{N}))\xrightarrow{d}W^Th+\frac{1}{2}h^T\partial^2_\theta L(\theta_0)h-\left(W^Tq-\frac{1}{2}q^T\partial^2_\theta L(\theta_0)q\right)$. It is easy to see that

$$\begin{aligned}&\left(-\left[\partial^2_\theta L(\theta_0)\right]^{-1}W,-\left[\partial^2_\theta L(\theta_0)\right]^{-1}W\right)\\&\qquad=\operatorname*{argmin}_h\max_q W^Th+\frac{1}{2}h^T\partial^d_\theta L(\theta_0)h-\left(W^Tq-\frac{1}{2}q^T\partial^d_\theta L(\theta_0)q\right),\end{aligned}$$

where $-\left[\partial^2_\theta L(\theta_0)\right]^{-1} W \sim N\left(0, \left[\partial^2_\theta L(\theta_0)\right]^{-1} \Sigma_W \left[\partial^2_\theta L(\theta_0)\right]^{-1}\right)$. Therefore, since

$$\left(\sqrt{N}\left(\widehat{\theta}^{(1)}_N - \theta_0\right), \sqrt{N}\left(\widehat{\theta}^{(2)}_N - \theta_0\right)\right) = \operatorname*{argmin}_h \max_q \widehat{L}(\theta_0 + h/\sqrt{N}, \theta_0 + q/\sqrt{N})),$$

continuous mapping theorem yields the desired conclusion.
Next, we will establish the required weak convergence. To this end, we apply Proposition 2 to the class

$$\widetilde{\mathcal{L}}_N := \left\{\widetilde{\ell}_N(h,q,\cdot) := \ell(\theta_0 + h/\sqrt{N}, \cdot) - \ell(\theta_0 + q/\sqrt{N}, \cdot), \ \left\|\begin{pmatrix} h \\ q \end{pmatrix}\right\| \leqslant R\right\}, \tag{11}$$

and note that $\left\|\begin{pmatrix} \theta_0 + h/\sqrt{N} \\ \theta_0 + q/\sqrt{N} \end{pmatrix} - \begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix}\right\| \leqslant \frac{R}{\sqrt{N}}$. We will also introduce the following notation for brevity (that will be used only in this proof):

$$\begin{aligned} \bar{L}_j(h,q) &:= \frac{1}{n}\sum_{i \in G_j} \widetilde{\ell}_N(h,q,X_i), \\ \widetilde{L}(h,q) &:= \mathbb{E}\widetilde{\ell}_N(h,q,X). \end{aligned} \tag{12}$$

The quantities δ and $\sigma^2(\delta)$ defined in Proposition 2 admit the bounds $\delta \leqslant \frac{R}{\sqrt{N}}$ and, in view of assumption 3A,

$$\sigma^2(\delta) := \sup_{\|(h,q)^T\| \leqslant R} \operatorname{Var}\left(\widetilde{\ell}_N(h,q,X)\right) \leqslant 2\mathbb{E}M^2_{\theta_0}(X)\frac{R^2}{N}, \tag{13}$$

hence Proposition 2 yields that

$$M_N(h,q) \ = \ \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q) - \widetilde{L}(h,q)\right)\right)} \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right) + o_P(1)$$

uniformly over $\left\|(h,q)^T\right\| \leqslant R$. In view of assumption 1,

$$\mathbb{P}\left(\left|\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right| \leqslant 1\right) \leqslant \mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q) - \widetilde{L}(h,q)\right)\right) \leqslant 1.$$

As $\sup_{\|(h,q)^T\| \leqslant R} \mathbb{P}\left(\left|\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right| \geqslant 1\right) \leqslant \sup_{\|(h,q)^T\| \leqslant R} \frac{\operatorname{Var}\left(\widetilde{\ell}(h,q,X)\right)}{\Delta_n^2} \to 0$ as $n,k \to \infty$, we deduce that $\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(h,q) - \widetilde{L}(h,q)\right)\right) \to 1$ and

$$M_N(h,q) = \Delta_n \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^k \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right) + o_P(1).$$

It remains to establish convergence of the finite dimensional distributions as well as asymptotic equicontinuity. Convergence of finite dimensional distributions will be deduced from Lindeberg-Feller's central limit theorem. As $\rho'(x) = x$ for $|x| \leqslant 1$ by assumption 1,

$$\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)\right) = \frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)$$

on the event $\mathcal{C}_j := \left\{ \left\| \frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h,q) - \widetilde{L}(h,q) \right) \right\| \leqslant 1 \right\}$. Chebyshev's inequality and assumption 3A imply that

$$\mathbb{P}(\bar{\mathcal{C}}_j) \leqslant \mathrm{Var}\left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h,q) - \widetilde{L}(h,q) \right) \right) \leqslant \frac{\mathbb{E}\widetilde{\ell}(h,q,X)}{\Delta_n^2} \leqslant \frac{\mathbb{E}M_{\theta_0}^2(X)\|h-q\|^2}{\Delta_n^2 N},$$

therefore, $\mathbb{P}\left(\bigcup_{j=1}^k \bar{\mathcal{C}}_j \right) \leqslant \frac{\mathbb{E}M_{\theta_0}^2(X)\|h-q\|^2}{\Delta_n^2 n} \to 0$ as $n \to \infty$, and

$$\begin{aligned} M_N(h,q) &= \Delta_n \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^k \frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h,q) - \widetilde{L}(h,q) \right) + o_P(1) \\ &= \frac{1}{\sqrt{N}} \sum_{j=1}^N \sqrt{N} \left(\widetilde{\ell}_N(h,q,X_j) - \widetilde{L}(h,q) \right) + o_P(1) \end{aligned}$$

on the event $\bigcap_{j=1}^k \mathcal{C}_j$. Hence, the limits of the finite dimensional distributions of the processes $M_N(h,q)$ and $\widehat{M}_N(h,q) := \frac{1}{\sqrt{N}} \sum_{j=1}^N \sqrt{N} \left(\widetilde{\ell}_N(h,q,X_j) - \widetilde{L}(h,q) \right)$ coincide. It is easy to conclude from the Lindeberg-Feller's theorem that the finite dimensional distributions of the process $(h,q) \mapsto \widehat{M}_N(h,q)$ are Gaussian, with covariance function

$$\lim_{N\to\infty} \mathrm{cov}\left(\widehat{M}_N(h_1,q_1), \widehat{M}_N(h_2,q_2) \right) = (h_1-q_1)^T \, \mathbb{E}\left[\partial_\theta \ell(\theta_0,X) \left(\partial_\theta \ell(\theta_0,X) \right)^T \right] (h_2-q_2),$$

Indeed, the aforementioned relation follows from the dominated convergence theorem, where pointwise convergence and the "domination" hold due to assumption 3A. Lindeberg's condition is also easily verified, as $\left(\sqrt{N}\widetilde{\ell}_N(h,q,X) \right)^2 \leqslant M_{\theta_0}^2(X)\|h-q\|^2$, implying that the sequence $\left\{ \left(\sqrt{N_j}\widetilde{\ell}_{N_j}(h,q,X) \right)^2 \right\}_{j\geqslant 1}$ is uniformly integrable.

Finally, we will prove asymptotic equicontinuity of the process $M_N(h,q)$. To this end, it suffices to establish that

$$\lim_{\delta\to 0} \limsup_{n,k\to\infty} \mathbb{E} \sup_{\|(h_1,q_1)^T - (h_2,q_2)^T\| \leqslant \delta} |M_N(h_1,q_1) - M_N(h_2,q_2)| = 0,$$

which is equivalent, in view of Proposition 2, to the requirement that

$$\begin{aligned} \lim_{\delta\to 0} \limsup_{n,k\to\infty} \mathbb{E} \sup_{\|(h_1,q_1)^T - (h_2,q_2)^T\| \leqslant \delta} \Bigg| \Delta_n \frac{\sqrt{N}}{\sqrt{k}} \sum_{j=1}^k \Bigg(\rho'\left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h_1,q_1) - \widetilde{L}(h_1,q_1) \right) \right) \\ - \rho'\left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(h_2,q_2) - \widetilde{L}(h_2,q_2) \right) \right) \Bigg) \Bigg| = 0. \quad (14) \end{aligned}$$

To establish relation (14), we first observe that for any h, q,

$$\sqrt{Nk} \left| \mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_1(h,q) - \widetilde{L}(h,q) \right) \right) \right| = o(1)$$

as $k, n \to \infty$ by Lemma 5 and inequality (13). Therefore, we only need to show that

$$\limsup_{n,k\to\infty} \mathbb{E} \sup_{\|(h_1,q_1)^T - (h_2,q_2)^T\| \leqslant \delta} |M_N(h_1,q_1) - M_N(h_2,q_2) - (\mathbb{E}M_N(h_1,q_1) - \mathbb{E}M_N(h_2,q_2))| \xrightarrow{\delta\to 0} 0.$$

Next, we apply symmetrization inequality with Gaussian weights (van der Vaart and Wellner, 1996). To this end, let $g_1, \ldots, g_k$ be i.i.d. $N(0,1)$ random variables independent of the data $X_1, \ldots, X_N$. Then, setting $B(\delta) := \left\{(h_1,q_1),(h_2,q_2) : \ \|(h_1,q_1)^T - (h_2,q_2)^T\| \leqslant \delta\right\}$, we have that

$$\begin{aligned}\mathbb{E}\sup_{B(\delta)} &\left|M_N(h_1,q_1) - M_N(h_2,q_2) - (\mathbb{E}M_N(h_1,q_1) - \mathbb{E}M_N(h_2,q_2))\right| \leqslant \\ &C(\rho,\Delta_\infty)\mathbb{E}\sup_{B(\delta)}\left|\frac{\sqrt{N}}{\sqrt{k}}\sum_{j=1}^k g_j\left(\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h_1,q_1) - \widetilde{L}(h_1,q_1)\right)\right)\right.\right. \\ &\qquad\qquad \left.\left. - \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h_2,q_2) - \widetilde{L}(h_2,q_2)\right)\right)\right)\right|.\end{aligned}$$

Let us condition everything on $X_1, \ldots, X_N$; we will write $\mathbb{E}_g$ to denote the expectation with respect to $g_1, \ldots, g_k$ only. Consider the Gaussian process $Y_{n,k}(t)$ defined via

$$\mathbb{R}^k \ni t \mapsto \frac{1}{\sqrt{k}}\sum_{j=1}^k g_j\sqrt{N}\rho'(t_j),$$

where $t_j := t_j(h,q) = \frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(h,q) - \widetilde{L}(h,q)\right)$, $j = 1, \ldots, k$. In what follows, we will rely on the ideas behind the proof of Theorem 2.10.6 in van der Vaart and Wellner (1996). Let us we partition the set $\{(h,q) : \|h,q\| \leqslant R\}$ into the subsets S_j, $j = 1, \ldots, N(\delta)$ of diameter at most δ with respect to the Euclidean distance $\|\cdot\|$, and let $t^{(j)} := t^{(j)}(h^{(j)}, q^{(j)}) \in S_j$ $j = 1, \ldots, N(\delta)$ be arbitrary points; we also note that $N(\delta) \leqslant \left(\frac{6R}{\delta}\right)^{2d}$. Next, set $T^{(j)} := \{t(h,q) : \ (h,q) \in S_j\}$. Our goal will be to show that

$$\limsup_{n,k\to\infty}\mathbb{E}\max_{j=1,\ldots,N(\delta)}\sup_{t\in T^{(j)}}\left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \to 0 \text{ as } \delta \to 0,$$

whence the desired conclusion would follow from Theorem 1.5.6 in van der Vaart and Wellner (1996). By Lemma 2.10.16 in van der Vaart and Wellner (1996),

$$\begin{aligned}\mathbb{E}_g\max_{j=1,\ldots,N(\delta)}\sup_{t\in T^{(j)}}\left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \leqslant C\Bigg(&\max_{j=1,\ldots,N(\delta)}\mathbb{E}_g\sup_{t\in T^{(j)}}\left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \\ &+ \log^{1/2}N(\delta)\max_{1\leqslant j\leqslant N(\delta)}\sup_{t\in T^{(j)}}\mathrm{Var}_g^{1/2}\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right)\Bigg). \qquad (15)\end{aligned}$$

Observe that $\mathrm{Var}_g\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right) = \frac{N}{k}\sum_{i=1}^k\left(\rho'(t_i) - \rho'(t_i^{(j)})\right)^2$, therefore,

$$\begin{aligned}\mathbb{E}\max_{1\leqslant j\leqslant N(\delta)}\sup_{t\in T^{(j)}}\mathrm{Var}_g^{1/2}\left(Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right) &\leqslant \mathbb{E}^{1/2}\sup_{t^{(1)},t^{(2)}}\frac{N}{k}\sum_{i=1}^k\left(\rho'(t_i^{(1)}) - \rho'(t_i^{(2)})\right)^2 \\ &\leqslant \sqrt{N}L(\rho')\mathbb{E}^{1/2}\sup_{t^{(1)},t^{(2)}}\left(t_1^{(1)} - t_1^{(2)}\right)^2 \\ = L(\rho')\mathbb{E}^{1/2}\sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta}&\left(\frac{\sqrt{nN}}{\Delta_n}\left(\bar{L}_1(h_1,q_1) - \bar{L}_1(h_2,q_2) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)\right)\right)^2,\end{aligned}$$

where the supremum is taken over all $t^{(1)}(h_1,q_1),\ t^{(2)}(h_2,q_2)$ such that $\|(h_1,q_1)-(h_2,q_2)\| \leqslant \delta$. To estimate the last expected supremum, we invoke Lemma 8 with $f_{h,q}(X) := \ell(\theta_0 + h/\sqrt{N}, X) - \ell(\theta_0 + q/\sqrt{N}, X)$, noting that, in view of assumption 3B,

$$\begin{aligned}\sqrt{N}|f_{h_1,q_1}(X) - f_{h_2,q_2}(X)| &\leqslant M_{\theta_0}(X)\left(\|h_1-h_2\| + \|q_1-q_2\|\right) \\ &\leqslant 2M_{\theta_0}(X)\left\|(h_1,q_1)-(h_2,q_2)\right\|. \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned}\mathbb{E}^{1/2} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \left(\frac{\sqrt{nN}}{\Delta_n}\left(\bar{L}_1(h_1,q_1) - \bar{L}_1(h_2,q_2) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)\right)\right)^2 \\ \leqslant C\sqrt{d}\mathbb{E}^{1/2} M^2_{\theta_0}(X)\cdot\delta,\end{aligned}$$

yielding that the second term on the right side of (15) converges in probability to 0 as $\delta \to 0$. It remains to show that the first term $\max_{j=1,\ldots,N(\delta)} \mathbb{E}_g \sup_{t\in T^{(j)}} \left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right|$ converges to 0 in probability. As ρ' is Lipschitz continuous, the covariance function of $Y_{n,k}(t)$ satisfies

$$\mathbb{E}\left(Y_{n,k}(t^{(1)}) - Y_{n,k}(t^{(2)})\right)^2 \leqslant L^2(\rho')\frac{N}{k}\sum_{j=1}^{k}\left(t_j^{(1)} - t_j^{(2)}\right)^2,$$

where the right side corresponds to the variance of increments of the process $Z_{n,k}(t) = \frac{L(\rho')}{\sqrt{k}}\sum_{j=1}^k g_j\sqrt{N}t_j$. Therefore, Slepian's lemma (Ledoux and Talagrand, 1991) implies that for any j,

$$\begin{aligned}&\mathbb{E}_g \sup_{t\in T^{(j)}} \left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \\ &\qquad\leqslant \mathbb{E}_g \sup_{(h,q)\in S_j} \frac{1}{\sqrt{k}}\left|\frac{\sqrt{Nn}}{\Delta_n}\sum_{i=1}^{k} g_j\left(\bar{L}_i(h,q) - \bar{L}_i(h^{(j)},q^{(j)}) - (\widetilde{L}(h,q) - L(h^{(j)},q^{(j)}))\right)\right|.\end{aligned}$$

In turn, it yields the inequality

$$\begin{aligned}&\mathbb{E} \max_{j=1,\ldots,N(\delta)} \mathbb{E}_g \sup_{t\in T^{(j)}} \left|Y_{n,k}(t) - Y_{n,k}(t^{(j)})\right| \\ &\leqslant \mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{k}}\left|\frac{\sqrt{Nn}}{\Delta_n}\sum_{i=1}^{k} g_j\left(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2))\right)\right|.\end{aligned}$$

To complete the proof, we will apply the multiplier inequality (Lemma 2.9.1 in van der Vaart and Wellner, 1996) to deduce that the last display is bounded, up to a multiplicative constant, by

$$\max_{m=1,\ldots,k} \mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{m}}\left|\frac{\sqrt{Nn}}{\Delta_n}\sum_{i=1}^{m} \varepsilon_j\left(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2))\right)\right|$$

where $\varepsilon_1,\ldots,\varepsilon_k$ are i.i.d. Rademacher random variables. Next, desymmetrization inequality (Lemma 2.3.6 in van der Vaart and Wellner, 1996) implies that for any $m = 1,\ldots,k$,

$$\mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{m}}\left|\frac{\sqrt{Nn}}{\Delta_n}\sum_{i=1}^{m} \varepsilon_j\left(\bar{L}_i(h_1,q_1) - \bar{L}_i(h_2,q_2) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2))\right)\right|$$

$$\leqslant 2\mathbb{E} \sup_{\|(h_1,q_1)-(h_2,q_2)\|\leqslant\delta} \frac{1}{\sqrt{mn}} \left| \frac{\sqrt{N}}{\Delta_n} \sum_{i=1}^{mn} \left(\widetilde{\ell}_N(h_1,q_1,X_i) - \widetilde{\ell}_N(h_2,q_2,X_i) - (\widetilde{L}(h_1,q_1) - \widetilde{L}(h_2,q_2)) \right) \right|$$

where $\widetilde{\ell}_N(h,q,X)$ and $\widetilde{L}(h,q)$ were defined in (11) and (12) respectively. It remains to apply Lemma 8 in exactly the same way as before (see (16)) to deduce that the last display is bounded from above by $C\sqrt{d}\mathbb{E}^{1/2}M^2_{\theta_0}(X)\cdot\delta \to 0$ as $\delta \to 0$. This completes the proof of asymptotic equicontinuity, and weak convergence, of the sequence of processes $M_N(h,q)$.

3.4. Proof of Theorem 3.

Define vector-valued functions $G(\theta) := \partial_\theta L(\theta)$ and let $G_N(\theta) := \partial_\theta \widehat{L}(\theta)$. Therefore, $G(\theta_0) = 0$ and $G_N(\widehat{\theta}_N) = 0$ by the definition of θ_0 and $\widehat{\theta}_N$.

Remark 1. *We give formal justification of differentiability of $\widehat{L}(\theta)$ in section 3.5 below.*

As $\widehat{\theta}_N \to \theta_0$ almost surely by Theorem 1,

$$G(\widehat{\theta}_N) - G(\theta_0) = \partial^2_\theta L(\theta_0)\left(\widehat{\theta}_N - \theta_0\right) + o\left(\|\widehat{\theta}_N - \theta_0\|\right). \tag{17}$$

On the other hand,

$$\begin{aligned} G(\widehat{\theta}_N) - G(\theta_0) &= G_N(\widehat{\theta}_N) - G_N(\theta_0) + \left(G(\widehat{\theta}_N) - G_N(\widehat{\theta}_N)\right) - (G(\theta_0) - G_N(\theta_0)) \\ &= G_N(\widehat{\theta}_N) - G_N(\widehat{\theta}_0) + r_N, \end{aligned} \tag{18}$$

where $r_N = \left(G(\widehat{\theta}_N) - G_N(\widehat{\theta}_N)\right) - (G(\theta_0) - G_N(\theta_0))$. Note that for any $\delta > 0$,

$$\sqrt{N}\|r_N\| \leqslant \sqrt{N} \sup_{\|\theta-\theta_0\|\leqslant\delta} \|G_N(\theta) - G(\theta)\| + \sqrt{N}\|r_N\| I\{\|\widehat{\theta}_N - \theta_0\| > \delta\}.$$

We need show that both terms in the sum above converge to 0 in probability. Uniform consistency of $\widehat{\theta}_N$ proved in Theorem 1 implies that the second term converges to 0, while result for the first term will follow from asymptotic continuity of the process $\theta \mapsto \sqrt{N}(G_N(\theta) - G(\theta))$ at θ_0.

Lemma 3. *Under assumptions of the theorem and for any $\varepsilon > 0$,*

$$\lim_{\delta\to 0}\limsup_{k,n\to\infty} \mathbb{P}\left(\sup_{\|\theta-\theta_0\|\leqslant\delta} \left\| \sqrt{N}\left(G_N(\theta) - G(\theta) - (G_N(\theta_0) - G(\theta_0))\right) \right\| \geqslant \varepsilon \right) = 0.$$

The proof of Lemma 3 is long and is presented in section 3.6. Combining the result of this lemma with (17) and (18), we deduce that

$$\partial^2_\theta L(\theta_0)(\widehat{\theta}_N - \theta_0) + o\left(\|\widehat{\theta}_N - \theta_0\|\right) = -\left(G_N(\theta_0) - G(\theta_0)\right) + o_p(N^{-1/2}).$$

It will be established in Lemma 4 below that $\sqrt{N}\left(G_N(\theta_0) - G(\theta_0)\right)$ is asymptotically multivariate normal, therefore, $\|G_N(\theta_0) - G(\theta_0)\| = O_p(N^{-1/2})$. Moreover, $\partial^2_\theta L(\theta_0)$ is non-singular by assumption 2. It follows that $\|\widehat{\theta}_N - \theta_0\| = O_p(N^{-1/2})$, and we conclude that

$$\sqrt{N}(\widehat{\theta}_N - \theta_0) = -\left(\partial^2_\theta L(\theta_0)\right)^{-1} \sqrt{N}\left(G_N(\theta_0) - G(\theta_0)\right) + o_p(1). \tag{19}$$

The claim of the theorem follows from display (19) above and Lemma 4 that is stated below.

Lemma 4. *Under assumptions of Theorem 3,*

$$\sqrt{N}\left(G_N(\theta_0) - G(\theta_0)\right) \to N\left(0, A^2(\theta_0,\rho)\right),$$

where

$$A^2(\theta_0,\rho) = \left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^{-2}\left(\mathbb{E}\left[\left(\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2 Z_2 Z_2^T\right]\right.$$
$$+\frac{1}{\Delta_\infty^2}\frac{\left(\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2\mathbb{E}\left(\rho'\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}{\left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}\Gamma(\theta_0)\Gamma^T(\theta_0)$$
$$\left.+\frac{2}{\Delta_\infty}\frac{\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\mathbb{E}\left[\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_1\right]}{\mathrm{Var}(Z_1)\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)}\Gamma(\theta_0)\Gamma^T(\theta_0)\right),$$

where $\Gamma(\theta_0) = \mathbb{E}\left[\partial_\theta \ell(\theta_0,X)\left(\ell(\theta_0,X) - L(\theta_0)\right)\right]$ *and* $(Z_2,Z_1) \in \mathbb{R}^d \times \mathbb{R}$ *is a centered multivariate normal vector with covariance matrix* $\Sigma(\theta_0) = \begin{pmatrix}\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\,\partial_\theta\ell(\theta_0,X)^T\right] & \Gamma(\theta_0)\\ \Gamma(\theta_0)^T & \mathrm{Var}\left(\ell(\theta_0,X)\right)\end{pmatrix}$.

The proof of this lemma is given in the following section.

3.5. Proof of Lemma 4.

Recall that $\widehat{L}(\theta)$ satisfies the equation $\sum_{j=1}^k \rho'\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-\widehat{L}(\theta)}{\Delta_n}\right) = 0$. Setting

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

we see that $H'(z) = \frac{\sqrt{n}}{\Delta_n}\sum_{j=1}^k \rho''\left(\sqrt{n}\,\frac{\bar{L}_j(\theta)-L(\theta)-z}{\Delta_n}\right)$. We will show that $H'(\widehat{L}(\theta_0)) \neq 0$ with high probability, whence the implicit function theorem applies and yields that

$$G_N(\theta_0) = \partial_\theta \widehat{L}(\theta_0) = \sum_{j=1}^k \frac{\rho_j''(\theta_0)\cdot\partial_\theta \bar{L}_j(\theta_0)}{\sum_{j=1}^k \rho_j''(\theta_0)} = \frac{1}{k}\sum_{j=1}^k \frac{\rho_j''(\theta_0)\cdot\partial_\theta\bar{L}_j(\theta_0)}{\frac{1}{k}\sum_{j=1}^k\rho_j''(\theta_0)}, \tag{20}$$

where we set $\rho_j''(\theta) := \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - \widehat{L}(\theta)\right)\right)$ for brevity. The following lemma justifies implicit differentiation.

Proposition 1. *Let* $\theta \in \Theta$, *and set* $\delta_0 := r(\theta)$, *where* $r(\theta)$ *is defined in assumption 3A. Then for all* $0 < \delta \leqslant \delta_0$,

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

therefore

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

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 3.8.

Remark 2. *We will also frequently use the corollary of the previous lemma stating that*

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

Indeed,

$$\frac{1}{k}\sum_{j=1}^{k}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-\widehat{L}(\theta')\right)\right)=\frac{1}{k}\sum_{j=1}^{k}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-L(\theta')\right)\right)$$
$$+\frac{1}{k}\sum_{j=1}^{k}\left[\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-\widehat{L}(\theta')\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-L(\theta')\right)\right)\right].$$

As ρ'' *is Lipschitz continuous with constant equal to 1,*

$$\left|\frac{1}{k}\sum_{j=1}^{k}\left[\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-\widehat{L}(\theta')\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta')-L(\theta')\right)\right)\right]\right|\leqslant\frac{\sqrt{n}}{\Delta_n}\left|\widehat{L}(\theta')-L(\theta')\right|.$$

Therefore, result follows from Proposition 1.

It follows from Lemma 8 that

$$\frac{1}{\sqrt{k}}\frac{1}{\sqrt{N}}\mathbb{E}\sup_{\|\theta'-\theta\|\leqslant\delta}\left|\sum_{j=1}^{N}\left(\ell(\theta',X_j)-L(\theta')\right)\right|\to 0\text{ as }k\to\infty.$$

Moreover, the bound of Lemma 5 implies that

$$\left|\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta')-L(\theta')\right)\right)-\mathbb{E}\rho''\left(\frac{Z(\theta')}{\Delta_n}\right)\right|\leqslant\frac{C(\rho,\Delta_n)}{n^{\tau/2}}\mathbb{E}\left|\ell(\theta',X)-L(\theta')\right|^{2+\tau},$$

where $Z(\theta')\sim N\left(0,\mathrm{Var}(\ell(\theta',X))\right)$. As ρ'' is positive near the origin, we see that for n,k large enough, $\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta')-L(\theta')\right)\right)>0$ with high probability uniformly over θ' in a neighborhood of θ. Preceding argument also implies that for any $\theta\in\Theta$,

$$\frac{1}{k}\sum_{j=1}^{k}\rho''_j(\theta)\to\mathbb{E}\rho''\left(\frac{Z(\theta)}{\Delta_\infty}\right)\tag{21}$$

in probability as $k, n \to \infty$. Invoking Slutsky's lemma, we see that it remains to find the weak limit of the sequence

$$\frac{\sqrt{N}}{k}\sum_{j=1}^{k}\rho_j''(\theta_0)\cdot\left(\partial_\theta \bar{L}_j(\theta_0)-\partial_\theta L(\theta_0)\right)=\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho_j''(\theta_0)\cdot\sqrt{n}\,\partial_\theta\bar{L}_j(\theta_0), \tag{22}$$

where we used the fact that $\partial_\theta L(\theta_0)=0$. Next, we will show that right side of the display (22) is asymptotically equivalent to

$$\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_j(\theta_0)+\Omega(\rho,\theta_0)\sqrt{N}\left(\widehat{L}(\theta_0)-L(\theta_0)\right) \tag{23}$$

for some fixed vector $\Omega(\rho,\theta_0)\in\mathbb{R}^d$ to be found. Indeed, Taylor's expansion around $L(\theta_0)$ yields that

$$\begin{aligned}
&\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-\widehat{L}(\theta_0)\right)\right)\\
&\quad=\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)-\frac{\sqrt{n}}{\Delta_n}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)(\widehat{L}(\theta_0)-L(\theta_0))\\
&\qquad\qquad+\frac{n}{2\Delta_n^2}\rho^{(4)}\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-\widetilde{L}_j(\theta_0)\right)\right)(\widehat{L}(\theta_0)-L(\theta_0))^2
\end{aligned}$$

for some $\widetilde{L}_j(\theta_0)\in[L(\theta_0),\widehat{L}(\theta_0)]$. Plugging in this expansion in the right-hand side of (22), it is easy to see (using the bound of Lemma 2), that the sum of terms involving $\rho^{(4)}$ is asymptotically negligible: indeed,

$$\begin{aligned}
&\left\|\frac{n}{2\Delta_n^2\sqrt{k}}(\widehat{L}(\theta_0)-L(\theta_0))^2\sum_{j=1}^{k}\rho^{(4)}\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\,\partial_\theta\bar{L}_j(\theta_0)\right\|\\
&\qquad\leqslant\|\rho^{(4)}\|_\infty\frac{n\sqrt{k}}{2\Delta_n^2}(\widehat{L}(\theta_0)-L(\theta_0))^2\frac{1}{k}\sum_{j=1}^{k}\sqrt{n}\left\|\partial_\theta\bar{L}_j(\theta_0)\right\|.
\end{aligned}\tag{24}$$

By Lemma 2, $(\widehat{L}(\theta_0)-L(\theta_0))^2=O_P(1/N)$, while $\frac{1}{k}\sum_{j=1}^{k}\sqrt{n}\left\|\partial_\theta\bar{L}_j(\theta_0)\right\|=O_P(1)$ since

$$\mathbb{E}\left\|\sqrt{n}\,\partial_\theta\bar{L}_1(\theta_0)\right\|\leqslant\mathbb{E}^{1/2}\left\|\sqrt{n}\,\partial_\theta\bar{L}_1(\theta_0)\right\|^2=\left(\operatorname{tr}\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\partial_\theta\ell(\theta_0,X)^T\right]\right)^{1/2}.$$

Therefore, the right-hand side in (24) is of order $O_P(k^{-1/2})=o_P(1)$ as $k\to\infty$. Next, observe that

$$\begin{aligned}
&\frac{\sqrt{n}\left(\widehat{L}(\theta_0)-L(\theta_0)\right)}{\sqrt{k}\Delta_n}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_j(\theta_0)\\
&\quad=\sqrt{N}\left(\widehat{L}(\theta_0)-L(\theta_0)\right)\left[\frac{\sqrt{n}}{\sqrt{k}\sqrt{N}\Delta_n}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_j(\theta_0)\right]\\
&\qquad=\sqrt{N}\left(\widehat{L}(\theta_0)-L(\theta_0)\right)\left[\frac{1}{k\Delta_n}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_j(\theta_0)\right].
\end{aligned}$$

We will show that $\frac{1}{k\Delta_n}\sum_{j=1}^k \rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta \bar{L}_j(\theta_0)\to\Omega(\rho,\theta_0)$ in probability as $k,n\to\infty$ and that $\sqrt{N}\left(\widehat{L}(\theta_0)-L(\theta_0)\right)$ is asymptotically normal, whence weak convergence of the product will follow from Slutsky's lemma. To this end, note that

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{k\Delta_n}\sum_{j=1}^k \rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta \bar{L}_j(\theta_0)\right)\\ &\quad\leqslant \frac{1}{k\Delta_n}\mathbb{E}\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta \bar{L}_1(\theta_0)\right)^2\leqslant\frac{\|\rho'''\|_\infty}{k\Delta_n}\mathbb{E}\left(\partial_\theta\ell(\theta_0,X)\right)^2\xrightarrow{k\to\infty}0.\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{1}{k\Delta_n}\sum_{j=1}^k \rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\partial_\theta \bar{L}_j(\theta_0)\\ &\qquad\qquad-\frac{1}{\Delta_n}\mathbb{E}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_1(\theta_0)\to 0\end{aligned}$$

in probability. It remains to find the limit of the vector of expectations $\mathbb{E}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_1(\theta_0)$ as $n\to\infty$. Note that for any jointly normal, centered random variables Z_1,Z_2, $\mathbb{E}[Z_2|Z_1]=\frac{\gamma}{\mathrm{Var}(Z_1)}Z_1$, where $\gamma=\mathrm{cov}(Z_1,Z_2)$. Therefore, for a differentiable function f,

$$\mathbb{E}f(Z_1)Z_2=\mathbb{E}\mathbb{E}\left[f(Z_1)Z_2|Z_1\right]=\frac{\gamma}{\mathrm{Var}(Z_1)}\mathbb{E}f(Z_1)Z_1=\gamma\,\mathbb{E}f'(Z_1),\tag{25}$$

where the last equality follows from Stein's identity. It is easy to see that the random vector $\sqrt{n}\begin{pmatrix}\partial_\theta\bar{L}_1(\theta_0)\\ \bar{L}_1(\theta_0)-L(\theta_0)\in\mathbb{R}^{d+1}\end{pmatrix}$ is asymptotically multivariate normal with covariance matrix

$$\Sigma(\theta_0)=\begin{pmatrix}\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\,\partial_\theta\ell(\theta_0,X)^T\right] & \Gamma(\theta_0)\\ \Gamma(\theta_0)^T & \mathrm{Var}\left(\ell(\theta_0,X)\right)\end{pmatrix},\tag{26}$$

moreover, the sequence of random variables $\left\{\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_1(\theta_0)\right\}_{n\geqslant 1}$ is uniformly integrable due to assumption 3B. Hence, it suffices to find the "cross-covariance" vector $\Gamma(\theta_0)$; the latter is equal to

$$\Gamma(\theta_0)=\mathbb{E}\left[\partial_\theta\ell(\theta_0,X)\left(\ell(\theta_0,X)-L(\theta_0)\right)\right].$$

Therefore, (25) applies coordinatewise with $f(x)=\rho'''\left(\frac{x}{\Delta_\infty}\right)$ and yields that

$$\Omega(\rho,\theta_0)=\frac{1}{\Delta_\infty^2}\Gamma(\theta_0)\mathbb{E}\rho^{(4)}\left(\frac{Z(\theta_0)}{\Delta_\infty}\right),\tag{27}$$

where $Z(\theta_0)\sim N\left(0,\mathrm{Var}(\ell(\theta_0,X))\right)$. We have thus established relation (23). Next, if we prove joint asymptotic normality of

$$\frac{1}{\sqrt{k}}\sum_{j=1}^k\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_j(\theta_0)$$

and $\sqrt{N}\left(\widehat{L}(\theta_0) - L(\theta_0)\right)$, then asymptotic normality of $\sqrt{N}\left(G_N(\theta_0) - G(\theta_0)\right)$ will follow from (23). The the key observation follows from Proposition 2 (stated and proved in section 4) which yields a representation of the form

$$\sqrt{N}\left(\widehat{L}(\theta) - L(\theta)\right) = \frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right)\right)} \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) + o_P(1) \quad (28)$$

that holds uniformly in a neighborhood of θ (to see that the remainder term is $o_P(1)$, it suffices to use the assumption stating that $k = o(n^{\tau})$). Furthermore, it follows from Lemma 5 that

$$\left|\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right)\right) - \mathbb{E}\rho''\left(\frac{Z(\theta)}{\Delta_\infty}\right)\right| \to 0$$

where $Z(\theta) \sim N\left(0, \mathrm{Var}(\ell(\theta, X))\right)$. Therefore, the limiting distribution of $\sqrt{N}\left(\widehat{L}(\theta) - L(\theta)\right)$ coincides with the limiting distribution of the sequence

$$\frac{\Delta_\infty}{\mathbb{E}\rho''\left(\frac{Z(\theta)}{\Delta_\infty}\right)} \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right).$$

Remark 3. *Representation (28) and Lemma 5 readily imply asymptotic normality of the sequence $\left\{\sqrt{N}\left(\widehat{L}(\theta) - L(\theta)\right)\right\}_{n,k\geqslant 1}$: indeed, we have that*

$$\left|\sqrt{k}\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)\right| \leqslant C(\rho, \Delta_n)\sqrt{\frac{k}{n^{\tau}}} \xrightarrow{k,n\to\infty} 0$$

by Lemma 5 and assumptions of Theorem 3, therefore asymptotic normality follows from the usual Lindeberg-Feller central limit theorem.

Combined with (22) and (23), Proposition 2 allows us to establish the asymptotic normality of $\sqrt{N}\,\partial_\theta \widehat{\ell}(\theta_0)$. Indeed, we only need to consider the asymptotic behavior of the sequence

$$\frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta \bar{L}_j(\theta_0) + \frac{\Delta_\infty\,\Omega(\rho,\theta_0)}{\mathbb{E}\rho''\left(\frac{Z(\theta_0)}{\Delta_\infty}\right)}\frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right). \quad (29)$$

This expression is a sum (over k) of independent random vectors, moreover, it is asymptotically centered. Indeed, the first term in (29) is asymptotically unbiased due to Lemma 6 and the fact that $k = o(n^{\tau})$ by assumption, while asymptotic unbiasedness of the second term follows from Lemma 5 together with an observation that, due to ρ' being odd, $\mathbb{E}\rho'(Z) = 0$ for normally distributed Z. Finally, asymptotic normality of the sum in (29) follows from Lindeberg-Feller's central limit theorem. To verify Lindeberg's condition, set $\xi_{n,j} := \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta \bar{L}_j(\theta_0)$ and $\eta_{n,j} := \frac{\Delta_n\,\Omega(\rho,\theta_0)}{\mathbb{E}\rho''\left(\frac{Z(\theta_0)}{\Delta_\infty}\right)}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)$, $j = 1, \ldots, k$. Then, for any unit vector $v \in \mathbb{R}^d$, $t \in \mathbb{R}$ and k large enough,

$$\begin{aligned}&\mathbb{E}\left(\langle \xi_{n,1} - \mathbb{E}\xi_{n,1}, v\rangle + t(\eta_{n,1} - \mathbb{E}\eta_{n,1})\right)^2 I\left\{\left|\langle \xi_{n,1} - \mathbb{E}\xi_{n,1}, v\rangle + t(\eta_{n,1} - \mathbb{E}\eta_{n,1})\right| \geqslant C(v,t)\sqrt{k}\right\} \\ &\leqslant 2\mathbb{E}\|\xi_{n,1} - \mathbb{E}\xi_{n,1}\|^2 I\left\{\|\xi_{n,1} - \mathbb{E}\xi_{n,1}\| \geqslant C'\sqrt{k}\right\} + 8M_\eta^2\mathbb{P}\Big(\|\xi_{n,1} - \mathbb{E}\xi_{n,1}\| \geqslant C'\sqrt{k}\Big)\end{aligned}$$

$$\leqslant 2\left(\mathbb{E}\|\xi_{n,1}-\mathbb{E}\xi_{n,1}\|^{2(1+\tau/2)}\right)^{2/(2+\tau)}\left(\mathbb{P}\Big(\|\xi_{n,1}-\mathbb{E}\xi_{n,1}\|\geqslant C'\sqrt{k}\Big)\right)^{\tau/(2+\tau)}$$
$$+\,8M_\eta^2\mathbb{P}\Big(\|\xi_{n,1}-\mathbb{E}\xi_{n,1}\|\geqslant C'\sqrt{k}\Big)\to 0 \text{ as } k\to\infty,$$

where we use the fact that $\eta_{n,1}$ is bounded almost surely by a constant M_η and that $\sup_n \mathbb{E}\|\xi_{n,1}\|^{2+\tau} < \infty$ in view of assumption 3A. It remains to find the limit of the vector of covariances [2]

$$\operatorname{cov}\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\cdot\partial_\theta\bar{L}_1(\theta_0),\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right),$$

which is equal to the limit, as $n\to\infty$, of

$$\mathbb{E}\left[\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\,\partial_\theta\bar{L}_1(\theta_0)\cdot\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right].$$

It is easy to see that for any $i\in\{1,\dots,d\}$, the sequence of random variables

$$\left\{\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_0)-L(\theta_0)\right)\right)\sqrt{n}\langle\partial_\theta\bar{L}_1(\theta_0),e_i\rangle\cdot\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right\}_{n\geqslant 1}$$

is uniformly integrable due to assumption 3A and the fact that both $\|\rho'\|_\infty$ and $\|\rho''\|_\infty$ are finite, therefore the limit is equal to $\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_2$, where $(Z_2,Z_1)\in\mathbb{R}^d\times\mathbb{R}$ is a centered multivariate normal vector with covariance $\Sigma(\theta_0)$ defined in (30). Using (25), we deduce that

$$\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_2=\Gamma(\theta_0)\frac{1}{\operatorname{Var}\left(\ell(\theta_0,X)\right)}\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_1.$$

Therefore, it follows from simple but tedious algebra that the limiting variance of the sequence in (29) is equal to

$$\mathbb{E}\left[\left(\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2 Z_2Z_2^T\right]+\frac{1}{\Delta_\infty^2}\frac{\left(\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2\mathbb{E}\left(\rho'\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}{\left(\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\right)^2}\Gamma(\theta_0)\Gamma^T(\theta_0)$$
$$+\,\frac{2}{\Delta_\infty}\frac{\mathbb{E}\rho^{(4)}\left(\frac{Z_1}{\Delta_\infty}\right)\mathbb{E}\left[\rho''\left(\frac{Z_1}{\Delta_\infty}\right)\rho'\left(\frac{Z_1}{\Delta_\infty}\right)Z_1\right]}{\operatorname{Var}(Z_1)\mathbb{E}\rho''\left(\frac{Z_1}{\Delta_\infty}\right)}\Gamma(\theta_0)\Gamma^T(\theta_0),$$

where (Z_1,Z_2) are defined same as before. The final result follows from (20), (21) and the display above.

3.6. Proof of Lemma 3.

Our goal is to show that for any $\varepsilon>0$,

$$\limsup_{n\to\infty}\mathbb{P}\left(\sup_{\|\theta-\theta_0\|_2\leqslant\delta}\left|\sqrt{N}\left(\partial_\theta\widehat{L}(\theta)-\partial_\theta L(\theta)-\left(\partial_\theta\widehat{L}(\theta_0)-\partial_\theta L(\theta_0)\right)\right)\right|\geqslant\varepsilon\right)$$

[2] Given a random variable ξ and a random vector $X\in\mathbb{R}^d$, we write $\operatorname{cov}(\xi,X)$ for the vector $(\operatorname{cov}(\xi,X_1),\dots,\operatorname{cov}(\xi,X_d))^T$.

converges to 0 as $\delta \to 0$. Recall representation (20) which states that

$$\partial_\theta \widehat{L}(\theta) - \partial_\theta L(\theta) = \frac{1}{k}\sum_{j=1}^{k} \frac{\rho_j''(\theta)\cdot\left(\partial_\theta \bar{L}_j(\theta) - \partial_\theta L(\theta)\right)}{\frac{1}{k}\sum_{j=1}^{k}\rho_j''(\theta)},$$

where $\rho_j''(\theta) = \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - \widehat{L}(\theta)\right)\right)$. In view of Lemma 1, it suffices to treat the numerator and the denominator in the representation above separately. We will first show that

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

and $\liminf_{n,k\to\infty} \mathbb{P}(|B_{n,k}(\theta_0)| \geqslant c) = 1$ for some $c > 0$, where

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

These claims immediately follow from Proposition 1 (see formula (21)), and the fact that the function $\theta \mapsto \mathbb{E}\rho\left(\frac{Z(\theta)}{\Delta_\infty}\right)$ is continuous under our assumptions, moreover, $\mathbb{E}\rho\left(\frac{Z(\theta_0)}{\Delta_\infty}\right) > 0$. The main part of the proof consists in establishing the asymptotic continuity of the process

$$A_{n,k}(\theta) = \sqrt{N}\frac{1}{k}\sum_{j=1}^{k} \rho_j''(\theta)\left(\partial_\theta \bar{L}_j(\theta) - \partial_\theta L(\theta)\right).$$

As we have shown before (see display (23)),

$$A_{n,k}(\theta) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)\sqrt{n}\left(\partial_\theta \bar{L}_j(\theta) - \partial_\theta L(\theta)\right) \\ + \Omega(\rho,\theta)\sqrt{N}\left(\widehat{L}(\theta) - L(\theta)\right) + o_p(1)$$

uniformly over the neighborhood of θ_0, where $\Omega(\rho,\theta)$ was defined in (27). Therefore, it suffices to control each part in the sum above separately.

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

It follows from Lemma 6 that whenever $k/n^\tau \to 0$ as $k, n \to \infty$, the process $K_{n,k}(\theta)$ is asymptotically unbiased, hence it suffices to prove the asymptotic continuity of the centered process $K_{n,k}(\theta) - \mathbb{E}K_{n,k}(\theta)$. Moreover, we can establish the result for each coordinate of the vector-valued process $K_{n,k}(\theta)$ separately, and we will set $K_{n,k}^{(i)}(\theta) = \langle K_{n,k}(\theta), e_i\rangle$, $\partial_\theta^{(i)}\bar{L}_j(\theta) = \langle \partial_\theta \bar{L}_j(\theta), e_i\rangle$, and $\partial_\theta^{(i)} L(\theta) = \langle \partial_\theta L(\theta), e_i\rangle$. Without the loss of generality, we only treat the case of $i = 1$ below. To this end, we will apply symmetrization inequality with Gaussian weights (Ledoux and Talagrand, 1991) to deduce that

$$\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|K_{n,k}^{(1)}(\theta) - K_{n,k}^{(1)}(\theta) - \left(\mathbb{E}K_{n,k}^{(1)}(\theta) - \mathbb{E}K_{n,k}^{(1)}(\theta_0)\right)\right| \leqslant C\mathbb{E}\sup_{\|\theta-\theta_0\|_2\leqslant\delta}|W_{n,k}(\theta) - W_{n,k}(\theta_0)|,$$

where

$$W_{n,k}(\theta) = \frac{1}{\sqrt{k}}\sum_{j=1}^{k} g_j\,\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)\sqrt{n}\left(\partial_\theta^{(1)}\bar{L}_j(\theta) - \partial_\theta^{(1)} L(\theta)\right)$$

and $g_1, \ldots, g_k$ are i.i.d. $N(0,1)$ random variables independent of $X_1, \ldots, X_N$. Let $\mathbb{E}_g$ denote the expectation with respect to $g_1, \ldots, g_k$ only, conditionally on $X_1, \ldots, X_N$. Observe that

$$\begin{aligned}
&\mathbb{E}_g (W_{n,k}(\theta_1) - W_{n,k}(\theta_2))^2 \\
&\quad \leqslant \frac{2\|\rho''\|_\infty^2}{k} \sum_{j=1}^k \left(\sqrt{n} \left(\partial_\theta^{(1)} \bar{L}_j(\theta_1) - \partial_\theta^{(1)} L(\theta_1) - \left(\partial_\theta^{(1)} \bar{L}_j(\theta_2) - \partial_\theta^{(1)} L(\theta_2) \right) \right) \right)^2 \\
&\quad + \frac{2}{k} \sum_{j=1}^k F_{1,j}^2(\delta) \left(\rho'' \left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(\theta_1) - L(\theta_1) \right) \right) - \rho'' \left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(\theta_2) - L(\theta_2) \right) \right) \right)^2,
\end{aligned}$$

where we set $F_1(\delta; x_1, \ldots, x_n) := \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| \frac{1}{\sqrt{n}} \left(\sum_{j=1}^n (\partial_\theta^{(1)} \ell(\theta, x_j) - \partial_\theta^{(1)} L(\theta)) \right) \right|$ and

$$F_{1,j}(\delta) := F_1(\delta; X_i,\ i \in G_j) = \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| \sqrt{n} \left(\partial_\theta^{(1)} \bar{L}_j(\theta) - \partial_\theta^{(1)} L(\theta) \right) \right|.$$

Slepian's lemma (Ledoux and Talagrand, 1991) therefore implies that for an absolute constant $C > 0$,

$$\begin{aligned}
&\mathbb{E}_g \sup_{\|\theta-\theta_0\| \leqslant \delta} |W_{n,k}(\theta) - W_{n,k}(\theta_0)| \\
&\qquad \leqslant C \mathbb{E}_g \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| W'_{n,k}(\theta) - W'_{n,k}(\theta_0) \right| + C \mathbb{E}_g \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| W''_{n,k}(\theta) - W''_{n,k}(\theta_0) \right|,
\end{aligned}$$

where

$$\begin{aligned}
W'_{n,k}(\theta) &= \frac{1}{\sqrt{k}} \sum_{j=1}^k g_j \sqrt{n} \left(\partial_\theta^{(1)} \bar{L}_j(\theta) - \partial_\theta^{(1)}(\theta) \right), \\
W''_{n,k}(\theta) &= \frac{1}{\sqrt{k}} \sum_{j=1}^k g_j F_{1,j}(\delta) \rho'' \left(\frac{\sqrt{n}}{\Delta_n} \left(\bar{L}_j(\theta_1) - L(\theta_1) \right) \right).
\end{aligned}$$

The estimate for $\mathbb{E} \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| W'_{n,k}(\theta) - W'_{n,k}(\theta_0) \right|$ is standard: it follows from the multiplier inequality (van der Vaart and Wellner, 1996) that

$$\begin{aligned}
&\mathbb{E} \sup_{\|\theta-\theta_0\| \leqslant \delta} \left| W'_{n,k}(\theta) - W'_{n,k}(\theta_0) \right| \\
&\leqslant C \max_{m=1,\ldots,k} \mathbb{E} \sup_{\|\theta-\theta_0\| \leqslant \delta} \frac{1}{\sqrt{m}} \left| \sum_{j=1}^m \varepsilon_j \sqrt{n} \left(\partial_\theta^{(1)} \bar{L}_j(\theta) - \partial_\theta^{(1)} L(\theta) - \left(\partial_\theta^{(1)} \bar{L}_j(\theta_0) - \partial_\theta^{(1)} L(\theta_0) \right) \right) \right|,
\end{aligned}$$

where $\varepsilon_1, \ldots, \varepsilon_k$ are i.i.d. Rademacher random variables independent of $X_1, \ldots, X_N$. Moreover, desymmetrization inequality (Lemma 2.3.6 in van der Vaart and Wellner, 1996) yields that for all m,

$$\begin{aligned}
&\mathbb{E} \sup_{\|\theta-\theta_0\| \leqslant \delta} \frac{1}{\sqrt{m}} \left| \sum_{j=1}^m \varepsilon_j \sqrt{n} \left(\partial_\theta^{(1)} \bar{L}_j(\theta) - \partial_\theta^{(1)} L(\theta) - \left(\partial_\theta^{(1)} \bar{L}_j(\theta_0) - \partial_\theta^{(1)} L(\theta_0) \right) \right) \right| \\
&\quad \leqslant C \mathbb{E} \sup_{\|\theta-\theta_0\| \leqslant \delta} \frac{1}{\sqrt{mn}} \left| \sum_{j=1}^{mn} \left(\partial_\theta^{(1)} \ell(\theta, X_j) - \partial_\theta^{(1)} L(\theta) - \left(\partial_\theta^{(1)} \ell(\theta_0, X_j) - \partial_\theta^{(1)} L(\theta_0) \right) \right) \right| \xrightarrow{\delta \to 0} 0,
\end{aligned}$$

where the last relation follows from assumption 3B. The bounds for the process $W''_{n,k}(\theta)$ are more involved. Our goal will be to estimate the covering numbers associated with this process. Observe that for $\tau > 0$,

$$
\begin{aligned}
(d''_{n,k}(\theta_1,\theta_2))^2 &:= \mathbb{E}_g \left(W''_{n,k}(\theta_1) - W''_{n,k}(\theta_2)\right)^2 \\
&= \frac{1}{k}\sum_{j=1}^k F^2_{1,j}(\delta)\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_1) - L(\theta_1)\right)\right) - \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2) - L(\theta_2)\right)\right)\right)^2 \\
&\leqslant \left(\frac{1}{k}\sum_{j=1}^k F^{2+\tau}_{1,j}(\delta)\right)^{2/(2+\tau)} \\
&\times \left(\frac{1}{k}\sum_{j=1}^k \left|\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_1) - L(\theta_1)\right)\right) - \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2) - L(\theta_2)\right)\right)\right|^{2(2+\tau)/\tau}\right)^{\tau/(2+\tau)}, \qquad (30)
\end{aligned}
$$

where the last relation follows from Hölder's inequality. Define the classes

$$
\mathcal{F}(\delta) := \left\{(x_1,\ldots,x_n) \mapsto F_1(\delta; x_1,\ldots,x_n)\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\frac{1}{n}\sum_{j=1}^n \ell(\theta,x_j) - L(\theta)\right)\right),\ \theta\in\Theta,\ \|\theta-\theta_0\| \leqslant \delta\right\},
$$

$$
G(\delta) := \left\{(x_1,\ldots,x_n) \mapsto \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\frac{1}{n}\sum_{j=1}^n \ell(\theta,x_j) - L(\theta)\right)\right),\ \theta\in\Theta,\ \|\theta-\theta_0\| \leqslant \delta\right\}.
$$

We will also define the process $Y_{n,k}(\theta) := \frac{1}{\sqrt{k}}\sum_{j=1}^k g_j\,\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)$ that is Gaussian conditionally on $X_1,\ldots,X_N$, and the associated L_2 distance

$$
d^2_{n,k}(\theta_1,\theta_2) := \mathbb{E}_g\left(Y_{n,k}(\theta_1) - Y_{n,k}(\theta_2)\right)^2.
$$

Inequality (30) and the fact that $\|\rho''\|_\infty < \infty$ imply the bound

$$
N(\mathcal{F}(\delta),\varepsilon,d''_{n,k}) \leqslant N\left(\mathcal{G}(\delta), \left(\frac{c(\rho)\varepsilon}{\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})}}\right)^{(2+\tau)/\tau}, d_{n,k}\right),
$$

where $\left\|F_1^2(\delta)\right\|_{L_{1+\tau/2}(P_{n,k})} := \left(\frac{1}{k}\sum_{j=1}^k F^{2+\tau}_{1,j}(\delta)\right)^{2/(2+\tau)}$. To estimate the covering numbers of $G(\delta)$, observe that the Euclidean ball $\{\theta : \|\theta-\theta_0\| \leqslant \delta\}$ can be covered by at most $K(\varepsilon) = \left(\frac{6\delta}{\varepsilon}\right)^d$ balls of radius $\varepsilon < \delta$ with centers $\theta_1,\ldots,\theta_{K(\varepsilon)}$. Hence, $G(\delta)$ can be covered by $K(\varepsilon)$ sets of the form $\left\{(x_1,\ldots,x_n) \mapsto \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\frac{1}{n}\sum_{j=1}^n \ell(\theta,x_j) - L(\theta)\right)\right), \|\theta-\theta_j\| \leqslant \varepsilon\right\}$, $j = 1,\ldots,K(\varepsilon)$. It remains to find an upper bound for the covering number of such a set by the balls with respect to the metric $d_{n,k}(\cdot,\cdot)$. Sudakov minoration principle (Ledoux and Talagrand, 1991) yields the inequality

$$
\begin{aligned}
&\sup_{0<z\leqslant\varepsilon} z\sqrt{\log N\left(\mathcal{G}(\varepsilon), z, d_{n,k}\right)} \\
&\leqslant C\sup_{\theta'}\mathbb{E}_g \sup_{\|\theta-\theta'\|\leqslant\varepsilon}\frac{1}{\sqrt{k}}\left|\sum_{j=1}^k g_j\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) - \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta') - L(\theta')\right)\right)\right)\right|.
\end{aligned}
$$

Therefore,

$$\sqrt{\log N\left(\mathcal{G}(\varepsilon), \varepsilon, d_{n,k}\right)}$$
$$\leqslant \frac{C}{\varepsilon} \sup_{\theta':\|\theta'-\theta_0\|\leqslant\delta} \mathbb{E}_g \sup_{\|\theta-\theta'\|\leqslant\varepsilon} \frac{1}{\sqrt{k}} \left| \sum_{j=1}^{k} g_j \left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) - \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta') - L(\theta')\right)\right)\right)\right|.$$

We apply Slepian's lemma again (where we use the fact that the function $x \mapsto \rho''(x)$ is Lipschitz continuous) to deduce that the latter supremum is dominated by

$$Q_{n,k}(\varepsilon) := \frac{C(\rho)}{\Delta_n} \sup_{\theta':\|\theta'-\theta_0\|\leqslant\delta} \mathbb{E}_g \sup_{\|\theta-\theta'\|\leqslant\varepsilon} \frac{1}{\sqrt{k}} \left| \sum_{j=1}^{k} g_j \sqrt{n}\left(\bar{L}_j(\theta) - L(\theta) - (\bar{L}_j(\theta') - L(\theta'))\right)\right|,$$

whence $\log^{1/2} N(\mathcal{F}(\delta), \varepsilon, d''_{n,k}) \leqslant \sqrt{d \log_+\left(\frac{6\delta}{\varepsilon_\tau}\right)} + \frac{1}{\varepsilon_\tau} Q_{n,k}(\varepsilon_\tau)$, where $\varepsilon_\tau = \left(\frac{c(\rho)\varepsilon}{\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})}}\right)^{(2+\tau)/\tau}$ and $\log_+(x) = \log x \vee 0$. Let $D''_{n,k}(\delta)$ be the diameter of $\mathcal{F}(\delta)$ with respect to the metric $d''_{n,k}$. We will apply Dudley's entropy integral estimate (Fact 1) to deduce that

$$\mathbb{E}_g \sup_{\|\theta-\theta_0\|\leqslant\delta} \left|W''_{n,k}(\theta) - W''_{n,k}(\theta_0)\right| \leqslant \int_0^{D''_{n,k}(\delta)} H^{1/2}\left(\mathcal{F}(\delta), \varepsilon, d''_{n,k}\right) d\varepsilon$$
$$\leqslant C(\tau) \int_0^{D''_{n,k}(\delta)} \sqrt{d} \log_+^{1/2}\left(\frac{C\delta^{\tau/(2+\tau)} \|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})}}{\varepsilon}\right) d\varepsilon + C \int_0^{D''_{n,k}(\delta)} \frac{1}{\varepsilon_\tau} Q_{n,k}(\varepsilon_\tau) d\varepsilon$$
$$\leqslant C(\tau)\sqrt{d}\delta^{\tau/(2+\tau)} \|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})} \int_0^1 \sqrt{\log(1/z)}dz$$
$$+ C(\rho)\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})} \int_0^{D''_{n,k}(\delta)/C\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_k)}} \frac{Q_{n,k}(\varepsilon^{(2+\tau)/\tau})}{\varepsilon^{(2+\tau)/\tau}} d\varepsilon. \quad (31)$$

It follows from (30) that

$$D''_{n,k}(\delta)/\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})}$$
$$\leqslant C(\rho) \sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta} \left(\frac{1}{k}\sum_{j=1}^{k} \left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_1) - L(\theta_1)\right)\right) - \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2) - L(\theta_2)\right)\right)\right)^2\right)^{\tau/(4+2\tau)}$$
$$= C(\rho)\mathrm{diam}(G(\delta); d_{n,k})^{\tau/(2+\tau)}.$$

Clearly, as ρ'' is bounded, $D_{n,k}(\delta) := \mathrm{diam}(G(\delta); d_{n,k})$ is also bounded by a constant $C_1(\rho)$ that depends only on ρ. Therefore, application of Cauchy-Schwarz inequality yields that

$$\int_0^{D''_{n,k}(\delta)/C\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_k)}} \frac{Q_{n,k}(\varepsilon^{(2+\tau)/\tau})}{\varepsilon^{(2+\tau)/\tau}} d\varepsilon$$
$$\leqslant C_1 D''_{n,k}(\delta)/\|F_1^2(\delta)\|^{1/2}_{L_{1+\tau/2}(P_{n,k})} \left(\int_0^{C_1(\rho)} \left(\frac{Q_{n,k}(\varepsilon^{(2+\tau)/\tau})}{\varepsilon^{(2+\tau)/\tau}}\right)^2 d\varepsilon\right)^{1/2}.$$

Taking expectation with respect to $X_1, \ldots, X_N$ in (31) and applying the Cauchy-Schwartz inequality to the last term, we deduce that

$$\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|W''_{n,k}(\theta)-W''_{n,k}(\theta_0)\right| \leqslant C(\tau)\sqrt{d}\delta^{\tau/(2+\tau)}\mathbb{E}^{1/2}\|F_1^2(\delta)\|_{L_{1+\tau/2}(P_{n,k})}$$
$$+ C\mathbb{E}^{1/2}D^2_{n,k}(\delta)\,\mathbb{E}^{1/2}\left[\int_0^{C_1(\rho)}\left(\frac{Q_{n,k}(\varepsilon^{(2+\tau)/\tau})}{\varepsilon^{(2+\tau)/\tau}}\right)^2 d\varepsilon\right].$$

We will show that $\limsup_{n,k\to\infty}\mathbb{E}D^2_{n,k}(\delta)\to 0$ as $\delta\to 0$ and that $\mathbb{E}\left[\int_0^{C_1(\rho)}\left(\frac{Q_{n,k}(\varepsilon^{(2+\tau)/\tau})}{\varepsilon^{(2+\tau)/\tau}}\right)^2 d\varepsilon\right]$ is bounded, which suffices to complete the proof of the lemma. First, observe that relation (30) combined with Hölder's inequality implies that

$$\mathbb{E}D^2_{n,k}(\delta)\leqslant \mathbb{E}^{2/(2+\tau)}\left[\frac{1}{k}\sum_{j=1}^k F_{1,j}^{2+\tau}(\delta)\right]$$
$$\times\mathbb{E}^{\tau/2(2+\tau)}\sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta}\frac{1}{k}\sum_{j=1}^k\left|\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar L_j(\theta_1)-L(\theta_1)\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar L_j(\theta_2)-L(\theta_2)\right)\right)\right|^{2(2+\tau)/\tau}.$$

Note that

$$\mathbb{E}F_{1,1}^{2+\tau}(\delta)=\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}\ell(\theta,X_j)-\partial_\theta^{(1)}L(\theta)\right)\right|^{2+\tau}$$
$$\leqslant 2^{1+\tau}\left(\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}\ell(\theta_0,X_j)-\partial_\theta^{(1)}L(\theta_0)\right)\right|^{2+\tau}\right.$$
$$\left.+\,\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}(\ell(\theta,X_j)-\ell(\theta_0,X_j))-\partial_\theta^{(1)}(L(\theta)-L(\theta_0)\right)\right|^{2+\tau}\right). \quad (32)$$

Rosenthal's inequality (see e.g. Ibragimov and Sharakhmetov, 2001) implies that

$$\mathbb{E}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}\ell(\theta_0,X_j)-\partial_\theta^{(1)}L(\theta_0)\right)\right|^{2+\tau}$$
$$\leqslant C(\tau)\left(\frac{\mathbb{E}\left|\partial_\theta^{(1)}\left(\ell(\theta_0,X)-L(\theta_0)\right)\right|^{2+\tau}}{n^{\tau/2}}+\mathrm{Var}^{1+\tau/2}\left(\partial_\theta^{(1)}\ell(\theta_0,X)\right)\right),$$

where the right side of the display above is finite due to the second part of assumption 3B. To estimate the second term on the right side of (32), we use Theorem 2.4.15 in van der Vaart and Wellner (1996), which gives that

$$\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}(\ell(\theta,X_j)-\ell(\theta_0,X_j))-\partial_\theta^{(1)}(L(\theta)-L(\theta_0))\right)\right|^{2+\tau}$$
$$C\left(\mathbb{E}\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^n\left(\partial_\theta^{(1)}(\ell(\theta,X_j)-\ell(\theta_0,X_j))-\partial_\theta^{(1)}(L(\theta)-L(\theta_0))\right)\right|\right.$$
$$\left.+\,n^{-\tau/(4+2\tau)}\mathbb{E}\left|\mathcal{V}(X;\delta)\right|^{2+\tau}\right),$$

where $\mathcal{V}(x;\delta)$ is the envelope function defined in assumption 3B. In view of the first part of the same assumption, both terms on the right side of the display above converge to 0 as $n \to \infty$ and $\delta \to 0$. Next, as $\|\rho''\|_\infty < \infty$,

$$\begin{aligned}
&\mathbb{E} \sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta} \frac{1}{k}\sum_{j=1}^k \left|\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_1)-L(\theta_1)\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2)-L(\theta_2)\right)\right)\right|^{2(1+\tau)/\tau} \\
&\quad\leqslant C(\rho)\mathbb{E} \sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta} \frac{1}{k}\sum_{j=1}^k \left|\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_1)-L(\theta_1)\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2)-L(\theta_2)\right)\right)\right| \\
&\quad\leqslant C(\rho)\mathbb{E} \sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta} \left|\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta_1)-L(\theta_1)\right)\right)-\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_2)-L(\theta_2)\right)\right)\right| \\
&\quad\leqslant \frac{C(\rho)}{\Delta_n}\mathbb{E} \sup_{\theta_1,\theta_2:\|\theta_j-\theta_0\|\leqslant\delta} \frac{1}{\sqrt{n}}\left|\sum_{j=1}^n \left(\ell(\theta_1,X_j)-\ell(\theta_2,X_j)-(L(\theta_1)-L(\theta_2))\right)\right|,
\end{aligned}$$

where we used Lipschitz continuity of ρ'' on the last step. Lemma 8 implies that the last display is bounded by $\frac{C'(\rho)}{\Delta_n}\sqrt{d}\,\delta\,\mathbb{E}^{1/2}M^2_{\theta_0}(X)$, allowing us to conclude that $\limsup_{n,k\to\infty}\mathbb{E}D^2_{n,k}(\delta) \to 0$ as $\delta \to 0$. It remains to estimate $\left[\int_0^{C_1(\rho)}\mathbb{E}\left(\frac{Q_{n,k}(\varepsilon^{(1+\tau)/\tau})}{\varepsilon^{(1+\tau)/\tau}}\right)^2 d\varepsilon\right]$. Set $\varepsilon' = \varepsilon^{(1+\tau)/\tau}$ so that

$$\mathbb{E}Q^2_{n,k}(\varepsilon') \leqslant \frac{C(\rho)}{\Delta_n^2}\mathbb{E}\sup_{\substack{\tilde{\theta},\theta:\|\tilde{\theta}-\theta_0\|\leqslant\delta \\ \|\theta-\tilde{\theta}\|\leqslant\varepsilon'}}\left(\frac{1}{\sqrt{k}}\left|\sum_{j=1}^k g_j\sqrt{n}\left(\bar{L}_j(\theta)-L(\theta)-(\bar{L}_j(\tilde{\theta})-L(\tilde{\theta}))\right)\right|\right)^2.$$

Application of Lemma 7 allows us to replace Gaussian random variables g_j's by a sequence of Rademacher random variables $\varepsilon_1,\ldots,\varepsilon_k$ independent from $X_1,\ldots,X_N$, implying that

$$\mathbb{E}Q^2_{n,k}(\varepsilon') \leqslant \frac{C(\rho)}{\Delta_n^2}\max_{m=1,\ldots,k}\mathbb{E}\sup_{\substack{\tilde{\theta},\theta:\|\tilde{\theta}-\theta_0\|\leqslant\delta \\ \|\theta-\tilde{\theta}\|\leqslant\varepsilon'}}\left(\frac{1}{\sqrt{m}}\left|\sum_{j=1}^m \varepsilon_j\sqrt{n}\left(\bar{L}_j(\theta)-L(\theta)-(\bar{L}_j(\tilde{\theta})-L(\tilde{\theta}))\right)\right|\right)^2.$$

Furthermore, desymmetrization inequality followed by the application of Lemma 8 yields that for any $m \in \{1,\ldots,k\}$,

$$\begin{aligned}
&\mathbb{E}\sup_{\substack{\tilde{\theta},\theta:\|\tilde{\theta}-\theta_0\|\leqslant\delta \\ \|\theta-\tilde{\theta}\|\leqslant\varepsilon'}}\left(\frac{1}{\sqrt{m}}\left|\sum_{j=1}^m \varepsilon_j\sqrt{n}\left(\bar{L}_j(\theta)-L(\theta)-(\bar{L}_j(\tilde{\theta})-L(\tilde{\theta}))\right)\right|\right)^2 \\
&\qquad\leqslant 4\mathbb{E}\sup_{\substack{\tilde{\theta},\theta:\|\tilde{\theta}-\theta_0\|\leqslant\delta \\ \|\theta-\tilde{\theta}\|\leqslant\varepsilon'}}\left(\frac{1}{\sqrt{mn}}\sum_{j=1}^{mn}\left(\ell(\theta,X_j)-\ell(\tilde{\theta},X_j)-(L(\theta)-L(\tilde{\theta}))\right)\right)^2 \\
&\qquad\leqslant Cd\varepsilon^2\mathbb{E}M^2_{\theta_0}(X),
\end{aligned}$$

Therefore, $\mathbb{E}\left[\int_0^{C_1(\rho)}\mathbb{E}\left(\frac{Q_{n,k}(\varepsilon')}{\varepsilon'}\right)^2 d\varepsilon\right] \leqslant C_2(\rho)d\mathbb{E}M^2_{\theta_0}(X) < \infty$, therefore completing the proof of asymptotic continuity of the process $K_{n,k}(\theta)$ at $\theta = \theta_0$.

2. Controlling $L_{n,k}(\theta) = \Omega(\rho,\theta)\sqrt{N}\left(\widehat{L}(\theta)-L(\theta)\right)$.

Asymptotic continuity of the process $\theta \mapsto \sqrt{N}\left(\widehat{L}(\theta) - L(\theta)\right)$ follows from Proposition 2. Indeed, it implies that for δ small enough,

$$\begin{aligned}
&\sup_{\|\theta-\theta_0\|\leqslant\delta} \left|\sqrt{N}\left(\widehat{L}(\theta) - \widehat{L}(\theta_0) - (L(\theta) - L(\theta_0)\right)\right| \\
&\qquad = \frac{\Delta_\infty}{\sqrt{k}} \sup_{\|\theta-\theta_0\|\leqslant\delta} \left| \frac{1}{\mathbb{E}\rho''\left(\frac{Z(\theta)}{\Delta_\infty}\right)} \sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) \right. \\
&\qquad\qquad \left. - \frac{1}{\mathbb{E}\rho''\left(\frac{Z(\theta_0)}{\Delta_\infty}\right)} \sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right)\right| + o_p(1).
\end{aligned}$$

As the function $\theta \mapsto \mathbb{E}\rho''\left(\frac{Z(\theta)}{\Delta_\infty}\right)$ is continuous under our assumptions, it suffices to show that

$$\sup_{\|\theta-\theta_0\|\leqslant\delta} \frac{1}{\sqrt{k}} \left| \sum_{j=1}^{k} \left[\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) - \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta_0) - L(\theta_0)\right)\right)\right]\right| \tag{33}$$

converges to 0 in probability as $\delta \to 0$. The process $\theta \mapsto \frac{1}{\sqrt{k}}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)$ is asymptotically unbiased in view of Lemma 5 and a fact that $\mathbb{E}\rho'(Z) = 0$ for normally distributed Z, therefore, we may consider centered processes in (33). Setting $\rho'_j(\theta) := \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right)$ and applying symmetrization and contraction inequalities, we see that

$$\begin{aligned}
&\mathbb{E} \sup_{\|\theta-\theta_0\|\leqslant\delta} \frac{1}{\sqrt{k}} \left| \sum_{j=1}^{k} \left(\rho'_j(\theta) - \rho'_j(\theta_0) - \mathbb{E}(\rho'_j(\theta) - \rho'_j(\theta_0))\right)\right| \\
&\qquad \leqslant 2\mathbb{E} \sup_{\|\theta-\theta_0\|\leqslant\delta} \frac{1}{\sqrt{k}} \left| \sum_{j=1}^{k} \varepsilon_j \left(\rho'_j(\theta) - \rho'_j(\theta_0)\right)\right| \\
&\qquad \leqslant \frac{4\|\rho'\|_\infty}{\Delta_n} \mathbb{E} \sup_{\|\theta-\theta_0\|\leqslant\delta} \frac{1}{\sqrt{N}} \left| \sum_{j=1}^{N} \left(\ell(\theta, X_j) - \ell(\theta_0, X_j) - (L(\theta) - L(\theta_0)\right)\right| \\
&\qquad \leqslant C(\rho)\sqrt{d}\sqrt{\delta}\mathbb{E}^{1/2} M^2_{\theta_0}(X),
\end{aligned}$$

where the last inequality follows from Lemma 8. This completes the proof of the main claim.

3.7. Proof of Lemma 2.

As it was mentioned after the statement of the lemma, the proof repeats the argument behind Theorem 3.1 in (Minsker, 2019b). The only difference is in the way we estimate

$$T_{n,k}(\Theta') := \sup_{\theta\in\Theta'} \left(\frac{1}{k}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) - \mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right)\right)\right).$$

Instead of applying Talagrand's concentration inequality, we will use Markov's inequality, which yields that

$$T_{n,k}(\Theta') \leqslant s\,\mathbb{E}\sup_{\theta\in\Theta'} \left| \frac{1}{k}\sum_{j=1}^{k} \rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta) - L(\theta)\right)\right) - \mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right)\right)\right|$$

with probability at least $1 - \frac{1}{s}$. It remains to check that (see (Minsker, 2019b) for the details)

$$\mathbb{E}\sup_{\theta\in\Theta'}\left|\frac{1}{k}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)-\mathbb{E}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta)-L(\theta)\right)\right)\right| \\ \leqslant C(\rho)\frac{1}{\Delta_n\sqrt{k}}\mathbb{E}\sup_{\theta\in\Theta'}\left|\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left(\ell(\theta,X_j)\right)-\mathbb{E}\ell(\theta,X)\right|.$$

3.8. Proof of Proposition 1.

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

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

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

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

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.

4. Auxiliary results.

The following proposition is one of the key technical results that the proofs of Theorems 2 and 3 rely on.

Proposition 2. *Let $\mathcal{L}=\{\ell(\theta,\cdot),\ \theta\in\Theta\}$ be a class of functions, and, given $\theta_0\in\Theta$, set $\sigma^2(\delta):=\sup_{\|\theta-\theta_0\|\leqslant\delta}\mathrm{Var}\left(\ell(\theta,X)\right)$. Moreover, let assumption 3A hold. Then for every $\delta\leqslant r(\theta_0)$ [3], the following representation holds uniformly over $\|\theta-\theta_0\|\leqslant\delta$:*

$$\sqrt{N}\left(\widehat{L}(\theta)-L(\theta)\right)=\frac{\Delta_n}{\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta)-L(\theta)\right)\right)}\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\rho'\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)+\mathcal{R}_{n,k,j}(\theta),$$

[3] $r(\theta_0)$ was defined in assumption 3A

where

$$\sup_{\|\theta-\theta_0\|\leqslant\delta} |\mathcal{R}_{n,k,j}(\theta)| \leqslant C(d,\theta_0)\left((\delta\vee\sigma(\delta))^2\frac{s^2}{\sqrt{k}} + s(\sigma^2(\delta)\wedge n^{-\tau/2})\left(\delta\vee\sigma(\delta)\right) + \sqrt{k}\left(\sigma^2(\delta)\wedge n^{-\tau/2}\right)^3\right)$$

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

Proof. Define

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

so that $\widehat{G}_k(\widehat{L}(\theta)-L(\theta);\theta)=0$, and let $G_k(z;\theta):=\mathbb{E}\widehat{G}_k(z;\theta)$. Next, consider the stochastic process

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

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

$$\sqrt{N}\frac{R_k(\theta')}{\partial_z G_k(z;\theta')|_{z=0}} = O_P\left(\frac{(\delta\vee\sigma(\delta))^2}{\sqrt{k}} + (\sigma^2(\delta)\wedge n^{-\tau/2})\left(\delta\vee\sigma(\delta)\right) + \sqrt{k}\left(\sigma^2(\delta)\wedge n^{-\tau/2}\right)^3\right) \tag{34}$$

uniformly over θ' in the neighborhood of θ. Taking this claim for granted for now, we see that $\sqrt{N}\left(\widehat{L}(\theta)-L(\theta)\right) = -\sqrt{N}\frac{\widehat{G}_k(0)}{\partial_z G_k(z;\theta)|_{z=0}} + \sqrt{N}\frac{R_k(\theta)}{\partial_z G_k(z;\theta)|_{z=0}}$, and in particular it follows from the claim above that the weak limits of $\sqrt{N}(\widehat{L}(\theta)-L(\theta))$ and

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

coincide. It remains to establish the relation (34). To this end, define $\widehat{e}_N(\theta):=\widehat{L}(\theta)-L(\theta)$ so that $G_k(\widehat{e}_N(\theta);\theta)=0$, and let $G_k(z;\theta):=\mathbb{E}\widehat{G}_k(z)$. Next, consider the stochastic process

$$R_k(\theta) = \widehat{G}_k(0;\theta) + \partial_z G_k(z;\theta)\big|_{z=0}\cdot\widehat{e}_N(\theta).$$

The following identity is immediate:

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

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

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

Therefore,

$$\widehat{G}_k\left(\widehat{e}_N(\theta);\theta\right) - \widehat{G}_k(0;\theta) = -\frac{\sqrt{n}}{k\Delta_n}\sum_{j=1}^k \rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\cdot\widehat{e}_N(\theta)$$

$$+\frac{n}{k\Delta_n^2}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2$$

$$+\frac{n}{k\Delta_n^2}\sum_{j=1}^{k}\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)-\tau_j\widehat{e}_N(\theta)\right)\right)-\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2$$

and

$$\begin{aligned}R_k(\theta)=\frac{\sqrt{n}}{\Delta_n}\frac{1}{k}\sum_{j=1}^{k}&\left(\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)-\mathbb{E}\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right)\cdot\widehat{e}_N(\theta)\\&-\frac{n}{\Delta_n^2}\frac{1}{k}\sum_{j=1}^{k}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2\\-\frac{n}{\Delta_n^2}\frac{1}{k}\sum_{j=1}^{k}&\left(\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)-\tau_j\widehat{e}_N(\theta)\right)\right)-\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right)\cdot\left(\widehat{e}_N(\theta)\right)^2\\&=R'(\theta)+R''(\theta)+R'''(\theta).\end{aligned}\tag{35}$$

It follows from Lemma 2 (with $\mathcal{O}=0$) and Lemma 8 that

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|\widehat{e}_N(\theta)\right|\leqslant C(d,\theta_0)\left(\frac{\delta\vee\sigma(\delta)}{\sqrt{N}}s+\frac{\sigma^2(\delta)\wedge n^{-\tau/2}}{\sqrt{n}}\right)$$

with probability at least $1-s^{-1}$ whenever $s\lesssim\sqrt{k}$. Moreover, Proposition 1 combined with Lemma 8 yields that

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

with probability at least $1-s^{-1}$. Therefore, the first term $R'(\theta)$in (35) satisfies

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|R'(\theta)\right|\leqslant C\left((\delta\vee\sigma(\delta))^2\frac{s^2}{k}+s(\sigma^2(\delta)\wedge n^{-\tau/2})\frac{\delta\vee\sigma(\delta)}{\sqrt{k}}\right)$$

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

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

with probability at least $1-s^{-1}$, again by Proposition 1, and $\left|\mathbb{E}\rho'''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_j(\theta)-L(\theta)\right)\right)\right|\leqslant C\left(\sigma^2(\delta)\wedge n^{-\tau/2}\right)$ by Lemma 5. Therefore, the term $R''(\theta)$ admits an upper bound

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|R''(\theta)\right|\leqslant C\left(\frac{(\delta\vee\sigma(\delta))^3}{k^{3/2}}s^3+\left(\sigma^2(\delta)\wedge n^{-\tau/2}\right)^3\right)$$

which holds with probability at least $1-s^{-1}$. Finally, as ρ''' is Lipschitz continuous by assumption, the third term $R'''(\theta)$ can be estimated via

$$\sup_{\|\theta-\theta_0\|\leqslant\delta}\left|R'''(\theta)\right|\leqslant C\frac{n}{\Delta_n^2}\left|\widehat{e}_N(\theta)\right|^3\leqslant\frac{C}{\sqrt{n}}\left(\frac{(\delta\vee\sigma(\delta))^3}{k^{3/2}}s^3+\left(\sigma^2(\delta)\wedge n^{-\tau/2}\right)^3\right)$$

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 (taking into account the fact that $s \leqslant c(d,\theta_0)\sqrt{k}$), it is easy to conclude that

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

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

Lemma 5. *Let $F:\mathbb{R}\mapsto\mathbb{R}$ be a function such that F'' is bounded and Lipschitz continuous. Moreover, suppose that $\xi_1,\ldots,\xi_n$ are independent centered random variables such that $\mathbb{E}|\xi_j|^2<\infty$ for all j, and that $Z_j,\ j=1,\ldots,n$ are independent with normal distribution $N(0,\mathrm{Var}(\xi_j))$. Then*

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n Z_j\right)\right| \leqslant C(F)\sum_{j=1}^n \mathbb{E}\left[\xi_j^2\cdot\min(|\xi_j,1)\right].$$

In particular, if $\mathbb{E}|\xi_j|^{2+\tau}<\infty$ for some $\tau\in(0,1]$ and all j, then

$$\left|\mathbb{E}F\left(\sum_{j=1}^n \xi_j\right) - \mathbb{E}F\left(\sum_{j=1}^n Z_j\right)\right| \leqslant C(F)\sum_{j=1}^n \mathbb{E}|\xi_j|^{2+\tau}.$$

Proof. We will apply the standard Lindeberg's replacement method (see e.g. 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

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

Similarly,

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

Lipschitz continuity and boundedness of F'' imply that $|F''(x)-F''(y)|\leqslant C(F)\min(1,|x-y|)$ with $C(F)=\max(\|F\|_\infty, L(F''))$. Therefore,

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

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 $\{|\xi_j| > 1\}$. □

Lemma 6. *Assume that* $\mathbb{E}\left\|\partial_\theta \ell(\theta, X)\right\|^{2+\tau} < \infty$ *for some* $\tau \in (0, 1]$. *Then*

$$\left\| \mathbb{E}\left[\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right) \right) \sqrt{n}\left(\partial_\theta \bar{L}_1(\theta) - \partial_\theta L(\theta)\right) \right] \right\|_\infty$$
$$\leqslant C(\rho, \Delta_n) \frac{\mathbb{E}\left|\ell(\theta, X) - L(\theta)\right|^{2+\tau} \left(\mathrm{Var}^{1/2}\left(\ell(\theta, X)\right) \vee \left\| V_\theta^{1/2}(\theta) \right\| \right)}{n^{\tau/2}},$$

where $V(\theta)$ *is the covariance matrix of the random vector* $\partial_\theta \ell(\theta, X)$.

Proof. It suffices to establish the upper bound for each coordinate of the vector

$$\rho''\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar{L}_1(\theta) - L(\theta)\right) \right) \sqrt{n}\left(\partial_\theta \bar{L}_1(\theta) - \partial_\theta L(\theta)\right).$$

Without loss of generality, it suffices to consider the first coordinate. To this end, let $\xi_j = \frac{1}{\sqrt{n}}(\ell(\theta, X_j) - L(\theta))$ and $\eta_j = \frac{1}{\sqrt{n}}\langle \partial_\theta(\ell(\theta, X_j) - L(\theta)), e_1\rangle$, $j = 1, \ldots, n$. Moreover, let (W_j, Z_j), $j = 1, \ldots, n$ be a sequence of i.i.d. centered bivariate normal vectors such that $\mathrm{cov}(W_j, Z_j) = \mathrm{cov}(\xi_j, \eta_j)$. Next, observe that $\mathbb{E}\rho''\left(\frac{1}{\Delta_n}\sum_{j=1}^n W_j\right)\left(\sum_{j=1}^n Z_j\right) = 0$. Indeed, the random vector $(S_{n,Z}, S_{n,W}) = \left(\sum_{j=1}^n W_j, \sum_{j=1}^n Z_j\right)$ is jointly Gaussian, hence

$$\mathbb{E}\rho''\left(\frac{1}{\Delta_n} S_{n,W}\right) S_{n,Z} = \mathbb{E}\left[\rho''\left(\frac{1}{\Delta_n} S_{n,W}\right) S_{n,Z} | S_{n,W}\right]$$
$$= \mathbb{E}\rho''\left(\frac{1}{\Delta_n} S_{n,W}\right) \mathbb{E}\left[S_{n,Z} | S_{n,W}\right] = \alpha_{Z,W} \mathbb{E}\rho''\left(\frac{1}{\Delta_n} S_{n,W}\right) S_{n,W} = 0,$$

where $\alpha_{Z,W} = \frac{\mathbb{E}S_{n,W}S_{n,Z}}{\mathbb{E}S_{n,W}^2}$ and the last equality follows since the function $x \mapsto \rho''\left(\frac{x}{\Delta_n}\right)x$ is odd. We can write $\mathbb{E}\rho''\left(\frac{\sum_{j=1}^n \xi_j}{\Delta_n}\right)\sum_{j=1}^n \eta_j$ as

$$\mathbb{E}\rho''\left(\frac{\sum_{j=1}^n \xi_j}{\Delta_n}\right)\sum_{j=1}^n \eta_j = \sum_{j=1}^n \eta_j \left(\rho''\left(\frac{\sum_{i=1}^n \xi_i}{\Delta_n}\right) - \rho''\left(\frac{\sum_{i\neq j} \xi_i}{\Delta_n}\right) \right),$$

where we used the fact that $\mathbb{E}\eta_j \rho''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) = 0$ for all j. Taylor expansion of the the function $\rho''\left(\frac{\sum_{i=1}^n \xi_i}{\Delta_n}\right)$ around the point $\sum_{i\neq j} \xi_i$ gives that

$$\rho''\left(\frac{\sum_{i=1}^n \xi_i}{\Delta_n}\right) - \rho''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) = \rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right)\frac{\xi_j}{\Delta_n} + \left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i + \zeta\cdot\xi_j}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right)\right)\frac{\xi_j}{\Delta_n}$$

for some $\zeta \in [0,1]$. Therefore,

$$\begin{aligned}
&\mathbb{E}\rho''\left(\frac{\sum_{j=1}^n \xi_j}{\Delta_n}\right)\left(\sum_{j=1}^n \eta_j\right) \\
&= \mathbb{E}\rho''\left(\frac{\sum_{j=1}^n \xi_j}{\Delta_n}\right)\left(\sum_{j=1}^n \eta_j\right) - \mathbb{E}\rho''\left(\frac{1}{\Delta_n}\sum_{j=1}^n W_j\right)\left(\sum_{j=1}^n Z_j\right) \\
&= \sum_{j=1}^n \frac{1}{\Delta_n}\mathbb{E}\left(\xi_j\eta_j\rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) - W_jZ_j\rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right) \\
&+ \sum_{j=1}^n \frac{1}{\Delta_n}\mathbb{E}\eta_j\xi_j\left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i + \zeta_j\cdot\xi_j}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right)\right) \\
&- \sum_{j=1}^n \frac{1}{\Delta_n}\mathbb{E}W_jZ_j\left(\rho'''\left(\frac{\sum_{i\neq j}W_i + \zeta_j'\cdot W_j}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right).
\end{aligned}$$

Using the fact that $\mathbb{E}\xi_j\eta_j = \mathbb{E}W_jZ_j$, we deduce that

$$\sum_{j=1}^n \frac{1}{\Delta_n}\mathbb{E}\left(\xi_j\eta_j\rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) - W_jZ_j\rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right) = n\frac{\mathbb{E}W_jZ_j}{\Delta_n}\mathbb{E}\left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right).$$

It remains to estimate $\mathbb{E}\left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right)$. Applying the standard Lindeberg's replacement method (see Lemma 5 above), it is easy to deduce that whenever $\rho^{(5)}(\cdot)$ is bounded and Lipschitz continuous,

$$\mathbb{E}\left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}W_i}{\Delta_n}\right)\right) \leqslant C(\rho)\frac{\mathbb{E}|\xi_1|^{2+\tau}}{\Delta_n^2 n^{\tau/2}}.$$

Moreover, as $\|\rho'''\|_\infty < \infty$ and ρ''' is Lipschitz continuous by assumption, $|\rho'''(x) - \rho'''(y)| \leqslant C(\rho,\gamma)\|x-y\|^\gamma$ for any $\gamma \in (0,1]$. Taking $\gamma = \tau/2$, we see that

$$\begin{aligned}
&\left|\sum_{j=1}^n \frac{1}{\Delta_n}\mathbb{E}\eta_j\xi_j\left(\rho'''\left(\frac{\sum_{i\neq j}\xi_i + \tau_j\xi_j}{\Delta_n}\right) - \rho'''\left(\frac{\sum_{i\neq j}\xi_i}{\Delta_n}\right)\right)\right| \\
&\leqslant \frac{C(\rho,\tau)}{\Delta_n}\sum_{j=1}^n \mathbb{E}\left|\eta_j\xi_j^{1+\tau/2}\right| \leqslant \frac{C(\rho,\tau)}{\Delta_n} n\,\mathbb{E}^{1/2}\eta^2\mathbb{E}^{1/2}|\xi_j|^{2+\tau} \leqslant \frac{C}{n^{\tau/2}}.
\end{aligned}$$

□

Lemma 7. *Let $Z_1(f), \dots, Z_k(f),\ f \in \mathcal{F}$ be i.i.d. stochastic processes such that $\mathbb{E}\sup_{f\in\mathcal{F}} |Z_1(f)| \leqslant \infty$. Assume that $g_1, \dots, g_k$ are i.i.d. $N(0,1)$ random variables independent of $Z_1, \dots, Z_k$. Similarly, let $\varepsilon_1, \dots, \varepsilon_k$ be i.i.d. Rademacher random variables independent of $Z_1, \dots, Z_k$. Then the following inequality holds:*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k} g_j Z_j(f)\right)^2 \leqslant C \max_{m=1,\dots,k} \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} \varepsilon_j Z_j(f)\right)^2$$

for some absolute constant $C > 0$.

Let us first deduce the final bound from the Lemma. Desymmetrization inequality (Lemma 2.3.6 in van der Vaart and Wellner, 1996) yields that

$$\begin{aligned}\mathbb{E}\sup_{\|\theta-\tilde\theta\|\leqslant\varepsilon'}&\left(\frac{1}{\sqrt{m}}\left|\sum_{j=1}^{m}\varepsilon_j\left(\frac{\sqrt{n}}{\Delta_n}\left(\bar\ell_j(\theta)-\ell(\theta)\right)-\frac{\sqrt{n}}{\Delta_n}\left(\bar\ell_j(\tilde\theta)-\ell(\tilde\theta)\right)\right)\right|\right)^2 \\ &\leqslant \frac{1}{\Delta_n^2}\mathbb{E}\sup_{\|\theta-\tilde\theta\|\leqslant\varepsilon'}\left(\frac{1}{\sqrt{mn}}\left|\sum_{j=1}^{mn}\left(\ell(\theta,X_j)-\ell(\tilde\theta,X_j)-\ell(\theta)+\ell(\tilde\theta)\right)\right|\right)^2.\end{aligned}$$

To estimate the last expression, it suffices to bound

$$\mathbb{E}\sup_{\|\theta-\tilde\theta\|\leqslant\varepsilon'}\frac{1}{\sqrt{mn}}\left|\sum_{j=1}^{mn}\left(\ell(\theta,X_j)-\ell(\tilde\theta,X_j)-\ell(\theta)+\ell(\tilde\theta)\right)\right|$$

and observe that the envelope $x \mapsto \sup_{\|\theta-\tilde\theta\|\leqslant\varepsilon}|\ell(\theta,x)-\ell(\tilde\theta,x)|$ is square integrable.

Proof of Lemma 7. Let $\tilde g_1 \geqslant \ldots \geqslant \tilde g_k \geqslant \tilde g_{k+1} := 0$ be the reversed order statistics corresponding to $|g_1|, \dots, |g_k|$, and note that $\tilde g_j = \sum_{i=j}^{k}(\tilde g_i - \tilde g_{i+1})$. Therefore,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\cdot}\sum_{j=1}^{k} g_j Z_j(f)\right)^2 &= \mathbb{E}_{|g|}\mathbb{E}_{\varepsilon,X}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\varepsilon_j|g_j|Z_j(f)\right)^2 \\ = \mathbb{E}_{|g|}\mathbb{E}_{\varepsilon,X}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\varepsilon_j\tilde g_j Z_j(f)\right)^2 &= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{j=1}^{k}\sum_{i=j}^{k}(\tilde g_i-\tilde g_{i+1})\varepsilon_j Z_j(f)\right)^2 \\ &= \mathbb{E}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\sqrt{i}(\tilde g_i-\tilde g_{i+1})\sum_{j=1}^{i}\frac{1}{\sqrt{i}}\varepsilon_j Z_j(f)\right)^2 \\ &\leqslant \mathbb{E}_{|g|}\mathbb{E}_{\varepsilon,X}\sup_{f\in\mathcal{F}}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\sqrt{i}(\tilde g_i-\tilde g_{i+1})\left|\frac{1}{\sqrt{i}}\sum_{j=1}^{i}\varepsilon_j Z_j(f)\right|\right)^2 \\ = \mathbb{E}_{|g|}\frac{1}{k}\sum_{i,l=1}^{k}&\left[\sqrt{i}\sqrt{l}(\tilde g_i-\tilde g_{i+1})(\tilde g_l-\tilde g_{l+1})\mathbb{E}_{\varepsilon,X}\left(\left|\frac{1}{\sqrt{i}}\sum_{j=1}^{i}\varepsilon_j Z_j(f)\right|\left|\frac{1}{\sqrt{l}}\sum_{v=1}^{l}\varepsilon_v Z_v(f)\right|\right)\right] \\ &\leqslant \max_{i=1,\dots,k}\mathbb{E}_{\varepsilon,X}\left|\frac{1}{\sqrt{i}}\sum_{j=1}^{i}\varepsilon_j Z_j(f)\right|^2\mathbb{E}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\sqrt{i}(\tilde g_i-\tilde g_{i+1})\right)^2.\end{aligned}$$

Finally, note that

$$\mathbb{E}\left(\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\sqrt{i}(\tilde{g}_i-\tilde{g}_{i+1})\right)^2 \leqslant \mathbb{E}\left(\frac{1}{\sqrt{k}}\int_0^\infty \sqrt{\#\{i\leqslant k:\ |g_i|\geqslant t\}}dt\right)^2$$
$$= \mathbb{E}\left(\frac{1}{\sqrt{k}}\int_0^\infty \frac{1}{\sqrt{1+t^2}}\sqrt{1+t^2}\sqrt{\#\{i\leqslant k:\ |g_i|\geqslant t\}}dt\right)^2$$
$$\leqslant \int_0^\infty \frac{dt}{1+t^2}\int_0^\infty (1+t^2)\mathbb{P}(|g|\geqslant t)dt \leqslant \pi\int_0^\infty (1+t^2)e^{-t^2/2}dt < \infty.$$

□

Lemma 8. *Let $\mathcal{F}=\left\{f_\theta,\ \theta\in\Theta'\subseteq\mathbb{R}^d\right\}$ be a class of functions that is Lipschitz in parameter, meaning that $|f_{\theta_1}(x)-f_{\theta_2}(x)|\leqslant M(x)\|\theta_1-\theta_2\|$. Moreover, assume that $\mathbb{E}M^p(X)<\infty$ for some $p\geqslant 1$. Then*

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

and

$$\mathbb{E}\sup_{\theta\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n\left(f_\theta(X_j)-Pf_{\theta_1}\right)\right|\right)^p$$
$$\leqslant C(p)\left(d^{p/2}\mathrm{diam}^p(\Theta',\|\cdot\|)\mathbb{E}\|M\|^p_{L_2(\Pi_n)}+\mathbb{E}^{1\wedge\frac{p}{2}}\left|f_{\theta_0}(X)-Pf_{\theta_0}\right|^{2\vee p}\right)$$

for any $\theta_0\in\Theta'$.

Proof. Symmetrization inequality yields that

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

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

$$\mathbb{E}_\varepsilon\sup_{\theta_1,\theta_2\in\Theta'}\left(\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n\varepsilon_j\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)\right|\right)^p$$
$$\leqslant C(p)\left(\mathbb{E}_\varepsilon\sup_{\theta_1,\theta_2\in\Theta'}\frac{1}{\sqrt{n}}\left|\sum_{j=1}^n\varepsilon_j\left(f_{\theta_1}(X_j)-f_{\theta_2}(X_j)\right)\right|\right)^p$$

$$\leqslant C(p)\left(\int_0^{D_n(\Theta')} H^{1/2}(z, T_n, d_n)dz\right)^p,$$

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. 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

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

Therefore,

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

and

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

□

Proof of the second bound follows from the triangle inequality

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

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

References

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

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

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

Brownlees, C., Joly, E., Lugosi, G., 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, pages 1148–1185. Institut Henri Poincaré.

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

Chinot, G., Guillaume, L., and Matthieu, L. (2018). Statistical learning with Lipschitz and convex loss functions. *arXiv preprint arXiv:1810.01090*.

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

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

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

Lecué, G. and Lerasle, M. (2017). Robust machine learning by median-of-means: theory and practice. *arXiv preprint arXiv:1711.10306*.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via MOM minimization. *arXiv preprint arXiv:1808.03106*.

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

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

Lugosi, G. and Mendelson, S. (2016). Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*.

Lugosi, G. and Mendelson, S. (2017). Regularization, sparse recovery, and median-of-means tournaments. *arXiv preprint arXiv:1701.04112*.

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

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 Mathieu, T. (2019). Excess risk bounds in robust empirical risk minimization. *arXiv preprint arXiv:1910.07485*.

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

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

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

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

Yin, D., Chen, Y., Ramchandran, K., and Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*.