

M and Z Estimators

Clément Lalanne

Notation for this lecture

Parameter and data. The parameter set is $\Theta \subset \mathbb{R}^p$ and the true parameter is θ_0 . The observations $X_1, \dots, X_n$ are i.i.d. copies of a random vector X of $\mathbb{R}^d$, of law $\mathcal{L}_{\theta_0}$ inside a model $\{\mathcal{L}_\theta; \theta \in \Theta\}$.

Criteria. $M_n : \Theta \rightarrow \mathbb{R}$ is a random function and $M : \Theta \rightarrow \mathbb{R}$ is its deterministic limit. For Z estimators, we use $Z_n : \Theta \rightarrow \mathbb{R}^q$ and $Z : \Theta \rightarrow \mathbb{R}^q$.

Random function means that, for every θ , $M_n(\theta)$ is a random variable, and that all the $M_n(\theta)$, $\theta \in \Theta$, are defined on one probability space.

Subscripts. $\mathbb{E}_\theta$ is the expectation computed as if the parameter were θ , while $\mathbb{E}_X = \mathbb{E}_{\theta_0}$ is the real one. Suprema carry their index set, as in $\sup_{\theta \in \Theta}$.

Derivatives. $J\phi(\theta)$ denotes the Jacobian matrix of ϕ at θ , and $\nabla\phi(\theta)$ the gradient when ϕ is real valued.

Part I. Two new ways to define an estimator

M estimators

Definition. Let $(M_n)_{n \in \mathbb{N}}$ be a sequence of random functions from Θ to $\mathbb{R}$. An **M-estimator** is a sequence $(\hat{\theta}_n)_{n \in \mathbb{N}}$ of random vectors taking values in Θ such that, for every n ,

$$\hat{\theta}_n \in \arg \max_{\theta \in \Theta} M_n(\theta) .$$

The letter M stands for maximizer. Since replacing M_n by $-M_n$ turns a maximization into a minimization, we lose nothing by always maximizing.

In practice, the almost sure requirement is too strong for anything computed numerically. This is why it is relaxed twice, first to holding with probability going to 1, and then to maximizing up to an additive $o_{\mathbb{P}}(1)$.

Exercise, the empirical mean

Exercise. Given random vectors $X_1, \dots, X_n$ of $\mathbb{R}^d$, express the empirical mean $\bar{X}_n = \frac{1}{n} \sum_{i \leq n} X_i$ as an M-estimator.

Solution. We take $\Theta = \mathbb{R}^d$ and

$$M_n(\theta) = -\frac{1}{n} \sum_{i \leq n} \|\theta - X_i\|^2.$$

Then $\nabla M_n(\theta) = -\frac{2}{n} \sum_{i \leq n} (\theta - X_i)$ vanishes only at $\theta = \bar{X}_n$. Furthermore, since M_n is strictly concave and tends to $-\infty$ at infinity, it has a unique maximizer, which is thus $\bar{X}_n$.

Empirical risk minimization

Definition. Given a loss $\ell : \mathbb{R}^d \times \Theta \rightarrow \mathbb{R}$, the **empirical risk** and the **risk** of a parameter are

$$\widehat{R}_S(\theta) = \frac{1}{n} \sum_{i \leq n} \ell(X_i, \theta) ,$$

$$R(\theta) = \mathbb{E}_X[\ell(X, \theta)] ,$$

and the **empirical risk minimizer** is the M-estimator of $M_n = -\widehat{R}_S$, whose limit criterion is $M = -R$.

Languages from two communities. The two names come from two communities rather than from two objects, with $\ell = -m$ and $M = -R$. In supervised learning, $X = (x, y)$ and $\ell(X, \theta) = \ell(y, f_\theta(x))$ for a predictor f_θ .

Maximum likelihood

Assumption. For every θ , $\mathcal{L}_\theta$ has a density f_θ with respect to a reference measure ν .

The **likelihood** and the **log-likelihood** of the sample are $L_n(\theta) = \prod_{i \leq n} f_\theta(X_i)$ and $\log L_n(\theta) = \sum_{i \leq n} \log f_\theta(X_i)$.

Definition. The **maximum likelihood estimator** is the M-estimator associated with $M_n = \frac{1}{n} \log L_n$, that is, with

$$M_n(\theta) = \frac{1}{n} \sum_{i \leq n} \log f_\theta(X_i) .$$

Equivalently, it is the empirical risk minimizer for the **log loss** $\ell(x, \theta) = -\log f_\theta(x)$, since $M_n = -\widehat{R}_S$ for that loss.

Idea. We choose the member of the model that makes the observed sample most probable. Note that the factor $\frac{1}{n}$ does not change the maximizer, and that it makes M_n converge to something finite.

Exercise, the gaussian model

Exercise. What is the maximum likelihood estimator in the gaussian model $\Theta = \mathbb{R} \times (0, \infty)$, $\theta = (\mu, \sigma^2)$, $\mathcal{L}_\theta = \mathcal{N}(\mu, \sigma^2)$?

Solution. With $f_{(\mu, \sigma^2)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, we obtain

$$M_n(\mu, \sigma^2) = \frac{1}{n} \sum_{i \leq n} \left\{ -\log \sigma - \frac{1}{2} \log(2\pi) - \frac{(X_i - \mu)^2}{2\sigma^2} \right\}.$$

$\partial_\mu M_n = \frac{1}{n\sigma^2} \sum_{i \leq n} (X_i - \mu)$ vanishes only at $\mu = \bar{X}_n$, and

$$\partial_{\sigma^2} M_n = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \frac{1}{n} \sum_{i \leq n} (X_i - \mu)^2,$$

which vanishes at $\mu = \bar{X}_n$ exactly when $\sigma^2 = \frac{1}{n} \sum_{i \leq n} (X_i - \bar{X}_n)^2$.

We recover the pair of the method of moments of the previous lecture, although the two methods differ in general.

Z estimators

Definition. Let $(Z_n)_{n \in \mathbb{N}}$ be a sequence of random functions from Θ to $\mathbb{R}^q$. A **Z-estimator** is a sequence $(\hat{\theta}_n)_{n \in \mathbb{N}}$ of random vectors taking values in Θ such that, for every n , almost surely

$$Z_n(\hat{\theta}_n) = 0 .$$

The letter Z stands for zero.

Every smooth M-estimator is a Z-estimator. Indeed, if Θ is open and M_n is differentiable, a maximizer in the interior satisfies $\nabla M_n(\hat{\theta}_n) = 0$, so that $\hat{\theta}_n$ is a Z-estimator with $Z_n = \nabla M_n$ and $q = p$.

The method of moments is a Z-estimator, with $Z_n(\theta) = e_n - e(\theta)$ in the notation of the previous lecture.

Part II. Consistency

The two ingredients of consistency

We want to show that $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$, a statement that mixes two things of completely different natures.

A statement about the criterion. M_n is close to M , uniformly in θ . All the randomness lies there.

A statement about the geometry. M separates θ_0 from the rest of Θ . There is no randomness in it at all.

Independence of the two statements. Each can hold while the other fails.

Stage one, the criterion error

Proposition. Let $\Delta_n = \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$ and assume that $\hat{\theta}_n$ nearly maximizes M_n , meaning

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) + o_{\mathbb{P}}(1) .$$

Then

$$\sup_{\theta \in \Theta} M(\theta) - M(\hat{\theta}_n) \leq 2\Delta_n + o_{\mathbb{P}}(1) .$$

Proof. For every fixed $\theta \in \Theta$, we have

$$\begin{aligned} M(\hat{\theta}_n) &\geq M_n(\hat{\theta}_n) - \Delta_n && \text{by definition of } \Delta_n \\ &\geq M_n(\theta) - \Delta_n + o_{\mathbb{P}}(1) && \text{by near-maximization} \\ &\geq M(\theta) - 2\Delta_n + o_{\mathbb{P}}(1) . && \text{by definition of } \Delta_n \end{aligned}$$

Since the left side does not depend on θ , taking the supremum over θ on the right concludes the proof.

Generality of stage one

If $\Delta_n \xrightarrow{\mathbb{P}} 0$, the proposition takes the form

$$\sup_{\theta \in \Theta} M(\theta) - M(\hat{\theta}_n) \xrightarrow{\mathbb{P}} 0 .$$

No topology, no identifiability, no compactness, no true parameter. Θ is an arbitrary set, M is an arbitrary function on it, and we have not even used θ_0 .

A statement about real numbers. The conclusion says that the value of the criterion at $\hat{\theta}_n$ approaches its optimal value, and nothing about where $\hat{\theta}_n$ is.

This is already the useful statement in many settings. In empirical risk minimization, the excess risk is this quantity, which generalization bounds control.

Empirical risk minimization as stage one

For the empirical risk minimizer of Part I, where $M = -R$ and $M_n = -\widehat{R}_S$, stage one takes the form

$$R(\hat{\theta}_n) - \inf_{\theta \in \Theta} R(\theta) \leq 2 \sup_{\theta \in \Theta} \left| \widehat{R}_S(\theta) - R(\theta) \right| + o_{\mathbb{P}}(1) .$$

Names of the terms. The left side is the estimation error, Δ_n is the uniform deviation, and the $o_{\mathbb{P}}(1)$ is the optimisation error (i.e. the gap between what the optimizer returns and the true empirical minimum).

Statistical learning theory as a whole is the estimation of Δ_n , by Rademacher complexities, VC dimension or metric entropy.

Irrelevance of stage two in learning

The parameter is not identifiable, on purpose. Permuting the units of a hidden layer, or rescaling one layer against the next, leaves f_θ and hence M unchanged. As a result, the label switching counterexample of Part II is the normal situation here rather than a pathology, and $\delta \equiv 0$ below the distance between two symmetric copies.

Over-parameterized models have a continuum of exact empirical minimizers, so that $\hat{\theta}_n$ is not well defined as a point and no statement about its convergence could hold.

The model is misspecified deliberately. There is no θ_0 with $\mathcal{L}_{\theta_0} = \mathbb{P}$, and $\inf_\theta R(f_\theta) - R^\star$ is the approximation error, i.e. a deterministic quantity that the estimator cannot affect.

What is asked of a learning algorithm is the risk of the predictor it returns

Now Stage Two: Well separated maximum

Definition. A point $\theta_0 \in \Theta$ is a **well separated maximizer** of M if $M(\theta_0) = \sup_{\theta \in \Theta} M(\theta)$ and $\delta(\epsilon) > 0$ for every $\epsilon > 0$, where

$$\delta(\epsilon) := M(\theta_0) - \sup_{\theta \in \Theta: \|\theta - \theta_0\| \geq \epsilon} M(\theta) .$$

Reading δ . It measures, in criterion, the loss incurred by standing at distance ϵ from θ_0 in parameter, and the condition says that this loss is positive at every distance.

It is deterministic. Indeed, δ depends on M and on Θ and on nothing else (no sample, no n , no probability).

Stage two, from the criterion to the parameter

Proposition. Let θ_0 be a well separated maximizer of M , and let $(\hat{\theta}_n)$ satisfy $M(\theta_0) - M(\hat{\theta}_n) \xrightarrow{\mathbb{P}} 0$.
Then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. Let us fix $\epsilon > 0$. If $\|\hat{\theta}_n - \theta_0\| \geq \epsilon$, then $\hat{\theta}_n$ belongs to the set over which the supremum defining $\delta(\epsilon)$ is taken, so that

$$\begin{aligned} M(\hat{\theta}_n) &\leq \sup_{\theta \in \Theta: \|\theta - \theta_0\| \geq \epsilon} M(\theta) \\ &= M(\theta_0) - \delta(\epsilon) , \end{aligned}$$

that is, $M(\theta_0) - M(\hat{\theta}_n) \geq \delta(\epsilon)$.

Hence,

$$\mathbb{P} \left(\|\hat{\theta}_n - \theta_0\| \geq \epsilon \right) \leq \mathbb{P} \left(M(\theta_0) - M(\hat{\theta}_n) \geq \delta(\epsilon) \right) ,$$

and the right side tends to 0 since the criterion error tends to 0 in probability and $\delta(\epsilon)$ is a fixed positive number, which concludes the proof.

The consistency theorem for M estimators

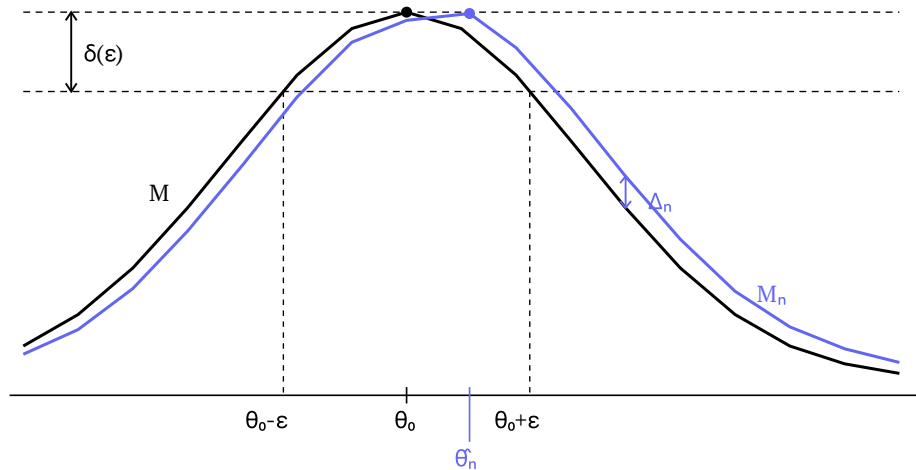
Theorem. Let (M_n) be random functions from $\Theta \subset \mathbb{R}^p$ to $\mathbb{R}$ and $M : \Theta \rightarrow \mathbb{R}$ deterministic. Assume

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$,
2. there is $\theta_0 \in \Theta$ such that for every $\epsilon > 0$, $\sup_{\theta \in \Theta: \|\theta - \theta_0\| \geq \epsilon} M(\theta) < M(\theta_0)$,
3. $(\hat{\theta}_n)$ satisfies $M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) + o_{\mathbb{P}}(1)$.

Then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. Hypotheses 1 and 3 yield stage one and hypothesis 2 is stage two, so that composing the two propositions concludes the proof.

The two stages, in one picture



Two gaps, two stages. The blue arrow is Δ_n (i.e. how far M_n sits from M uniformly), and stage one turns it into a bound on the vertical distance from the top. The black arrow is $\delta(\epsilon)$, the exchange rate between the axes, since leaving the interval lowers the criterion by $\delta(\epsilon)$, so that a criterion error below $\delta(\epsilon)$ keeps $\hat{\theta}_n$ inside it. A flat curve makes $\delta(\epsilon)$ zero and the exchange impossible.

Role of each hypothesis

Hypothesis 1 carries all the probability. It is a law of large numbers, uniform in θ , which does not mention θ_0 and to which the final part of the lecture is entirely devoted.

Hypothesis 2 carries all the geometry. It does not mention M_n , the sample, or n .

Hypothesis 3 is the definition of the estimator, weakened as far as possible.

A sufficient condition for well separation

Lemma. If Θ is compact, M is continuous on Θ , and θ_0 is the unique maximizer of M , then θ_0 is a well separated maximizer.

Proof. Let us fix $\epsilon > 0$ and let $K_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}$. If K_ϵ is empty, the supremum is $-\infty$ and there is nothing to prove. Otherwise, K_ϵ is closed in the compact Θ , hence compact, and the continuous M attains its supremum on it at some $\theta_\epsilon \neq \theta_0$, so that

$$\begin{aligned} \sup_{\theta \in K_\epsilon} M(\theta) &= M(\theta_\epsilon) \\ &< M(\theta_0) . \quad \text{since } \theta_0 \text{ is the unique maximizer} \end{aligned}$$

From separation to a rate

Let us suppose that the separation is polynomial, i.e. $\delta(\epsilon) \geq c\epsilon^\alpha$ for some $c, \alpha > 0$ and all small ϵ .

If $\|\hat{\theta}_n - \theta_0\| = r$, then $\|\hat{\theta}_n - \theta_0\| \geq \epsilon$ for every $\epsilon \leq r$, so that the argument of stage two yields $M(\theta_0) - M(\hat{\theta}_n) \geq \delta(\epsilon) \geq c\epsilon^\alpha$ for every $\epsilon \leq r$. Hence, $M(\theta_0) - M(\hat{\theta}_n) \geq cr^\alpha$ and

$$\begin{aligned} \|\hat{\theta}_n - \theta_0\| &\leq \left(\frac{M(\theta_0) - M(\hat{\theta}_n)}{c} \right)^{1/\alpha} \\ &\leq \left(\frac{2\Delta_n + o_{\mathbb{P}}(1)}{c} \right)^{1/\alpha} . \quad \text{by stage one} \end{aligned}$$

The generic case is $\alpha = 2$. If M is twice differentiable at θ_0 with negative definite Hessian, the separation is quadratic near θ_0 , and the exponent $1/\alpha = 1/2$ loses a square root, since a typical Δ_n of order $n^{-1/2}$ only yields $n^{-1/4}$.

A pessimistic rate. The bound above uses the uniform Δ_n over all of Θ to control a local quantity. We latter show that replacing it by a modulus over a shrinking neighbourhood of θ_0 recovers the rate $n^{-1/2}$.

The same two stages for Z estimators

Proposition (stage one). If $\sup_{\theta \in \Theta} \|Z_n(\theta) - Z(\theta)\| \xrightarrow{\mathbb{P}} 0$ and $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1)$, then $\|Z(\hat{\theta}_n)\| = o_{\mathbb{P}}(1)$.

Proof. By the triangle inequality, $\|Z(\hat{\theta}_n)\| \leq \|Z_n(\hat{\theta}_n)\| + \sup_{\theta \in \Theta} \|Z_n(\theta) - Z(\theta)\|$, and both terms are $o_{\mathbb{P}}(1)$.

Proposition (stage two). If $Z(\theta_0) = 0$ and $\delta(\epsilon) := \inf_{\|\theta - \theta_0\| \geq \epsilon} \|Z(\theta)\| > 0$ for every $\epsilon > 0$, then $\|Z(\hat{\theta}_n)\| = o_{\mathbb{P}}(1)$ implies $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

Proof. $\|\hat{\theta}_n - \theta_0\| \geq \epsilon$ forces $\|Z(\hat{\theta}_n)\| \geq \delta(\epsilon)$, so $\mathbb{P}(\|\hat{\theta}_n - \theta_0\| \geq \epsilon) \leq \mathbb{P}(\|Z(\hat{\theta}_n)\| \geq \delta(\epsilon)) \rightarrow 0$.

The consistency theorem for Z estimators

Theorem. Let (Z_n) be random functions from $\Theta \subset \mathbb{R}^p$ to $\mathbb{R}^q$ and $Z : \Theta \rightarrow \mathbb{R}^q$ deterministic. Assume

1. $\sup_{\theta \in \Theta} \|Z_n(\theta) - Z(\theta)\| \xrightarrow{\mathbb{P}} 0$,
2. $Z(\theta_0) = 0$ and, for every $\epsilon > 0$, $\inf_{\theta \in \Theta: \|\theta - \theta_0\| \geq \epsilon} \|Z(\theta)\| > 0$,
3. $(\hat{\theta}_n)$ satisfies $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1)$.

Then $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.

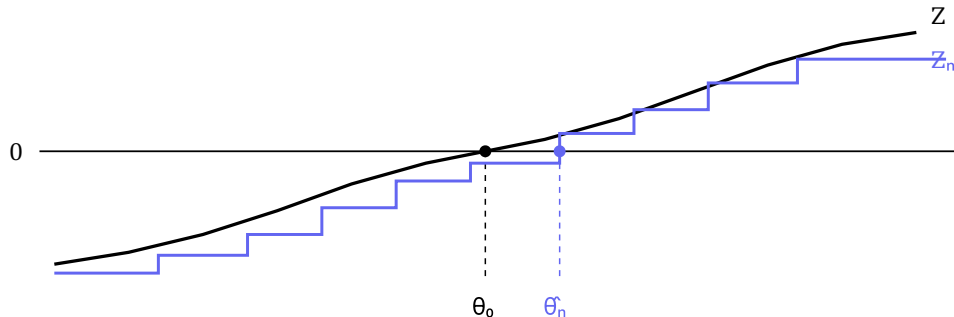
Remark. The course notes write hypothesis 2 as one chain, $\inf \|Z(\theta)\| > 0 = \|Z(\theta_0)\|$. It consists of two separate requirements, namely that the limit function vanishes at θ_0 and that it stays away from zero everywhere else.

Easy case, a monotone function in one dimension

Theorem. Let $\Theta = \mathbb{R}$, let (Z_n) be random functions from $\mathbb{R}$ to $\mathbb{R}$ and $Z : \mathbb{R} \rightarrow \mathbb{R}$ deterministic. Assume

1. for every **fixed** θ , $Z_n(\theta) \xrightarrow{\mathbb{P}} Z(\theta)$, with no uniformity,
2. almost surely, Z_n is non-decreasing,
3. there is θ_0 such that $Z(\theta_0 - \epsilon) < 0 < Z(\theta_0 + \epsilon)$ for every $\epsilon > 0$.

Then any $(\hat{\theta}_n)$ with $Z_n(\hat{\theta}_n) = o_{\mathbb{P}}(1)$ satisfies $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$.



Proof in the monotone case

Let us fix $\epsilon > 0$. Splitting on the two sides, we have

$$\begin{aligned} & \mathbb{P} \left(|\hat{\theta}_n - \theta_0| \geq \epsilon \right) \\ &= \mathbb{P}(\hat{\theta}_n \leq \theta_0 - \epsilon) + \mathbb{P}(\hat{\theta}_n \geq \theta_0 + \epsilon) \\ &\leq \mathbb{P} \left(Z_n(\hat{\theta}_n) \leq Z_n(\theta_0 - \epsilon) \right) + \mathbb{P} \left(Z_n(\hat{\theta}_n) \geq Z_n(\theta_0 + \epsilon) \right) \quad \text{since } Z_n \text{ is non-decreasing} \\ &= \mathbb{P} (o_{\mathbb{P}}(1) \leq Z(\theta_0 - \epsilon) + o_{\mathbb{P}}(1)) + \mathbb{P} (o_{\mathbb{P}}(1) \geq Z(\theta_0 + \epsilon) + o_{\mathbb{P}}(1)) \quad \text{by hypotheses 3 and 1} \end{aligned}$$

The two numbers $Z(\theta_0 - \epsilon) < 0$ and $Z(\theta_0 + \epsilon) > 0$ are fixed, and each probability is the probability that a quantity tending to 0 in probability exceeds a fixed nonzero threshold in the wrong direction, so that both tend to 0, which concludes the proof.

The trade-off. Uniform convergence, the demanding hypothesis of the general theorem, is replaced by pointwise convergence at two points per ϵ . In exchange, the statement needs monotonicity and dimension one.

Exercise, the empirical median

Exercise. Let $X_1, X_2, \dots$ be i.i.d. copies of a real X with a continuous density bounded away from 0 on a neighbourhood of the population median. Show that the empirical median $\hat{\theta}_n$, defined by $\sum_{i \leq n} \text{sign}(\hat{\theta}_n - X_i) = 0$, is consistent.

Solution. We take $Z_n(\theta) = \frac{1}{n} \sum_{i \leq n} \text{sign}(\theta - X_i)$, so that $Z_n(\hat{\theta}_n) = 0$, and

$$Z_n(\theta) = \frac{1}{n} \sum_{i \leq n} \mathbf{1}\{X_i < \theta\} - \frac{1}{n} \sum_{i \leq n} \mathbf{1}\{X_i > \theta\}$$

$$\xrightarrow{\mathbb{P}} \mathbb{P}(X < \theta) - \mathbb{P}(X > \theta)$$

$$= 2F_X(\theta) - 1,$$

by the law of large numbers

since $\mathbb{P}(X = \theta) = 0$

which we denote $Z(\theta)$.

Hypothesis 2 holds since $\theta \mapsto \text{sign}(\theta - x)$ is non-decreasing for each x , and hypothesis 3 holds with θ_0 the median ($F_X(\theta_0) = 1/2$), since F_X is strictly increasing where the density is positive. Hence, the theorem applies.