STAT 460/560 Class 11: Consistency of M- and Z-estimators

Ben Bloem-Reddy

Reading: Chapter 5.1-5.2, [van98].

1. M- and Z- estimators

A common way to estimate a parameter θ (or functional $\theta(P)$) from an i.i.d. sample $X_1, \dots, X_n \sim P$ is to maximize an objective

$$\theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) \ , \tag{11.1}$$

where m_θ is a known, fixed $\mathbb{R}$-valued function. An estimator $\hat{\theta}_n$ that maximizes $M_n(\theta)$ over Θ is called an **M-estimator**. An example that we've already seen is maximum likelihood, where $m_\theta = \log f_\theta$.

When appropriate derivatives exist, maximizing M_n often is equivalent to solving the system of equations

$$\Psi_n(\theta) := \sum_{i=1}^{n} \psi_\theta(X_i) = 0 \ . \tag{11.2}$$

Here ψ_θ is vector-valued, typically one dimension for each component of θ. In the example of maximum likelihood, ψ_θ is the score function.

Abstracting away from derivatives, *any* estimator $\hat{\theta}_n$ that is obtained by solving a system of equations like (11.2) is called a **Z-estimator** (for zero). van der Vaart [van98] has some examples other than maximum likelihood of each type.

Throughout, θ_0 will denote the true underlying parameter/functional.

Activity 11.1. Assume that $X_i \in \mathbb{R}$. Show that $\theta(P) = E[X]$ (the sample mean) and $\theta(P) =$ the sample median can be written as Z-estimators. Show that both functions are nonincreasing in θ.

2. Random functions

Denote by $\hat{P}_n$ the **empirical measure**, $\hat{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A)$, which is just the generalization of the empirical CDF to more general spaces. Using operator notation for expectations, $Pf = \int f(x) dP(x)$, we can write

$$M_n(\theta) = \hat{P}_n m_\theta \ , \quad \text{and} \quad \Psi_n(\theta) = \hat{P}_n \psi_\theta \ .$$

Viewed as functions of θ, these are **random functions**. The randomness comes from the sample $X_1, \dots, X_n$. Fig. 1 shows an example. By the law of large numbers, they converge in probability pointwise,

$$M_n(\theta) \xrightarrow{\text{P}} M(\theta) = P m_\theta \ , \quad \text{and} \quad \Psi_n(\theta) \xrightarrow{\text{P}} \Psi(\theta) = P \psi_\theta \ , \quad \theta \in \Theta \ .$$

The main objective today is to prove that under appropriate conditions, M- and Z-estimators are consistent, i.e., $\hat{\theta}_n \xrightarrow{\text{P}} \theta_0$. Since an M-estimator $\hat{\theta}_n$ maximizes the random function $\theta \mapsto M_n(\theta)$, the estimator implicitly depends on *the entire function*. Hence, pointwise consistency $M_n(\theta) \xrightarrow{\text{P}} M(\theta)$ is not strong enough.

Intuitively, if there are values of θ where $M_n(\theta)$ is (erroneously) large and for which $M_n(\theta)$ also converges very slowly, then we may not be able to guarantee that $M_n(\hat{\theta}_n) \xrightarrow{\text{p}} M(\hat{\theta}_n)$.

3. Consistency of M-estimators: uniform convergence

Assuming that Θ is a metric space[1] with metric d. We want to show consistency of M-estimators $\hat{\theta}_n$, i.e., that $d(\hat{\theta}_n, \theta_0) \xrightarrow{\text{p}} 0$.

We'll start with a result that is very similar to our proof of the consistency of MLEs. (In fact, it's essentially the same.) But it's worth looking at again and thinking carefully about why/how it works.

Theorem 11.1. *Let M_n be random functions and M a fixed function of θ such that for every $\epsilon > 0$,*

1. $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$.
2. $\sup_{\theta : d(\theta, \theta_0) \geq \epsilon} M(\theta) < M(\theta_0)$.

If $\hat{\theta}_n$ is a sequence of estimators such that $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$, then $\hat{\theta}_n$ converges in probability to θ_0.

What are the assumptions? The first is a *uniform law of large numbers* (ULLN), which as we will see shortly is a generally useful assumption/property. It is a *stochastic* property, having to do with the set of functions $\{m_\theta : \theta \in \Theta\}$ and the underlying distribution P. Proving that a sequence of estimators satisfies a ULLN can be quite challenging, and often requires techniques from empirical process theory, which is the subject of Ch. 19 of [van98]. Ch. 4 of [Wai19] takes a related but different approach to ULLNs. All of that is beyond the scope of this course, so be aware that there is a lot packed in to the first assumption.

The second assumption is purely a property of M, ensuring that the maximum is well separated, in the sense that $M(\theta)$ is close to $M(\theta_0)$ only if $d(\theta, \theta_0)$ is small.

Finally, the condition that $\hat{\theta}_n$ is a sequence of estimators such that $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ allows us to apply the theorem to estimators that *nearly* maximize M_n, rather than the exact maximization that we required in Class 8. (This is the main point of difference in the proof.)

Proof. By assumption, $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$. The uniform convergence implies pointwise convergence, such that $M_n(\theta_0) \xrightarrow{\text{p}} M(\theta_0)$, so that $M_n(\hat{\theta}_n) \geq M(\theta_0) - o_P(1)$. Therefore,

$$\begin{aligned} M(\theta_0) - M(\hat{\theta}_n) &\leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_P(1) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + o_P(1) \xrightarrow{\text{p}} 0 \,, \end{aligned}$$

where the convergence follows from Assumption 1. Now, by Assumption 2, for every $\epsilon > 0$ there exists a $\delta > 0$ such that $M(\theta) < M(\theta_0) - \delta$ for every $d(\theta, \theta_0) \geq \epsilon$. Therefore, the condition $d(\theta, \theta_0) \geq \epsilon$ implies that $M(\theta) < M(\theta_0) - \delta$ and hence

$$P\{d(\hat{\theta}_n, \theta_0) \geq \epsilon\} \leq P\{M(\hat{\theta}_n) < M(\theta_0) - \delta\} \to 0 \,.$$

□

We can apply this result to Z-estimators by noting that a zero of Ψ_n maximizes the function $\theta \mapsto -\|\Psi_n(\theta)\|$. See Theorem 5.9 in [van98] for a precise statement.

4. Consistency of Z-estimators: Uniqueness of solution or monotonicity

The ULLN assumption often can be replaced, but as van der Vaart notes, there is not a "one-size-fits-all" approach. Here's an example that works in some interesting cases.

[1] A metric space is a set equipped with a metric, or distance, function d. Recall that $d\colon \Theta \times \Theta \mathbb{R}$ is a metric if for all $\theta, \theta', \theta'' \in \Theta$: i) $d(\theta, \theta) = 0$; ii) if $\theta \neq \theta'$ then $d(\theta, \theta') > 0$; iii) $d(\theta, \theta') = d(\theta', \theta)$; and iv) d satisfies the triangle inequality, $d(\theta, \theta'') \leq d(\theta, \theta') + d(\theta', \theta'')$.

Lemma 11.2. *Let Θ be a subset of $\mathbb{R}$, and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n(\theta) \xrightarrow{p} \Psi(\theta)$ for every $\theta \in \Theta$. Assume that each map $\theta \mapsto \Psi_n(\theta)$ is:*

1. *continuous and has exactly one zero $\hat{\theta}_n$; or*
2. *nondecreasing with $\Psi_n(\hat{\theta}_n) \xrightarrow{p} 0$.*

Let θ_0 be a point such that $\Psi(\theta_0 - \epsilon) < 0 < \Psi(\theta_0 + \epsilon)$ for every $\epsilon > 0$. Then $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Note that this applies to nonincreasing functions $\tilde{\Psi}_n$ because $\Psi_n = -\tilde{\Psi}_n$ is nondecreasing.

Proof. If Assumption 1 holds, then $\{\Psi_n(\theta_0 - \epsilon) < 0, \Psi_n(\theta_0 + \epsilon) > 0\}$ implies that $\{\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon\}$, and therefore

$$P(\Psi_n(\theta_0 - \epsilon) < 0, \Psi_n(\theta_0 + \epsilon) > 0) \leq P(\theta_0 - \epsilon \leq \hat{\theta}_n < \theta_0 + \epsilon) \, .$$

The LHS converges to 1 because $\Psi_n(\theta_0 \pm \epsilon) \xrightarrow{\text{p}} \Psi(\theta_0 \pm \epsilon)$ for each $\epsilon > 0$. Thus the RHS converges to 1, which implies that $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$.

Alternatively, if $\Psi_n(\theta)$ is nondecreasing and $\hat{\theta}_n$ is a zero, then the same argument applies.

Let Assumption 2 hold instead. Get ready for some thorny arguments. $\Psi_n(\theta_0 - \epsilon) < -\delta$ and $\hat{\theta}_n \leq \theta_0 - \epsilon$ imply that $\Psi_n(\hat{\theta}_n) < -\delta$. However, by assumption $\Psi_n(\hat{\theta}_n) = o_P(1)$, so $P(\Psi_n(\hat{\theta}_n) < -\delta) \to 0$. Let $L_n = \{\hat{\theta}_n > \theta_0 - \epsilon\} \setminus \{\Psi_n(\theta_0 - \epsilon) < -\delta\}$. (Note the direction of the inequality in the second event.) Then $P(L_n) = o(1)$. In words, since $\Psi_n(\theta_0 - \epsilon) < -\delta$ and $\hat{\theta}_n \leq \theta_0 - \epsilon$ together imply something that has probability approaching 0, then the event that one of them holds implies that the other one must *not* hold with probability approaching 1.

Similarly, $\Psi_n(\theta_0 + \epsilon) > \delta$ and $\hat{\theta} \geq \theta_0 + \epsilon$ imply that $\Psi_n(\hat{\theta}_n) > \delta$, and a similar argument applies to the right tail. Hence,

$$P(\Psi_n(\theta_0 - \epsilon) < -\delta, \Psi_n(\theta_0 + \epsilon) > \delta) \leq P(\theta_0 - \epsilon < \hat{\theta}_n < \theta_0 + \epsilon) + o(1) \, .$$

For $\delta = \min\{-\Psi(\theta_0 - \epsilon), \Psi(\theta_0) + \epsilon\}/2$, the LHS converges to 1, and therefore $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$. □

Activity 11.2. Show that if the population median θ_0 is unique (i.e., $P(X < \theta - \epsilon) < 1/2 < P(X > \theta + \epsilon)$ for all $\epsilon > 0$) then the sample median converges in probability to θ_0.

References

[van98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

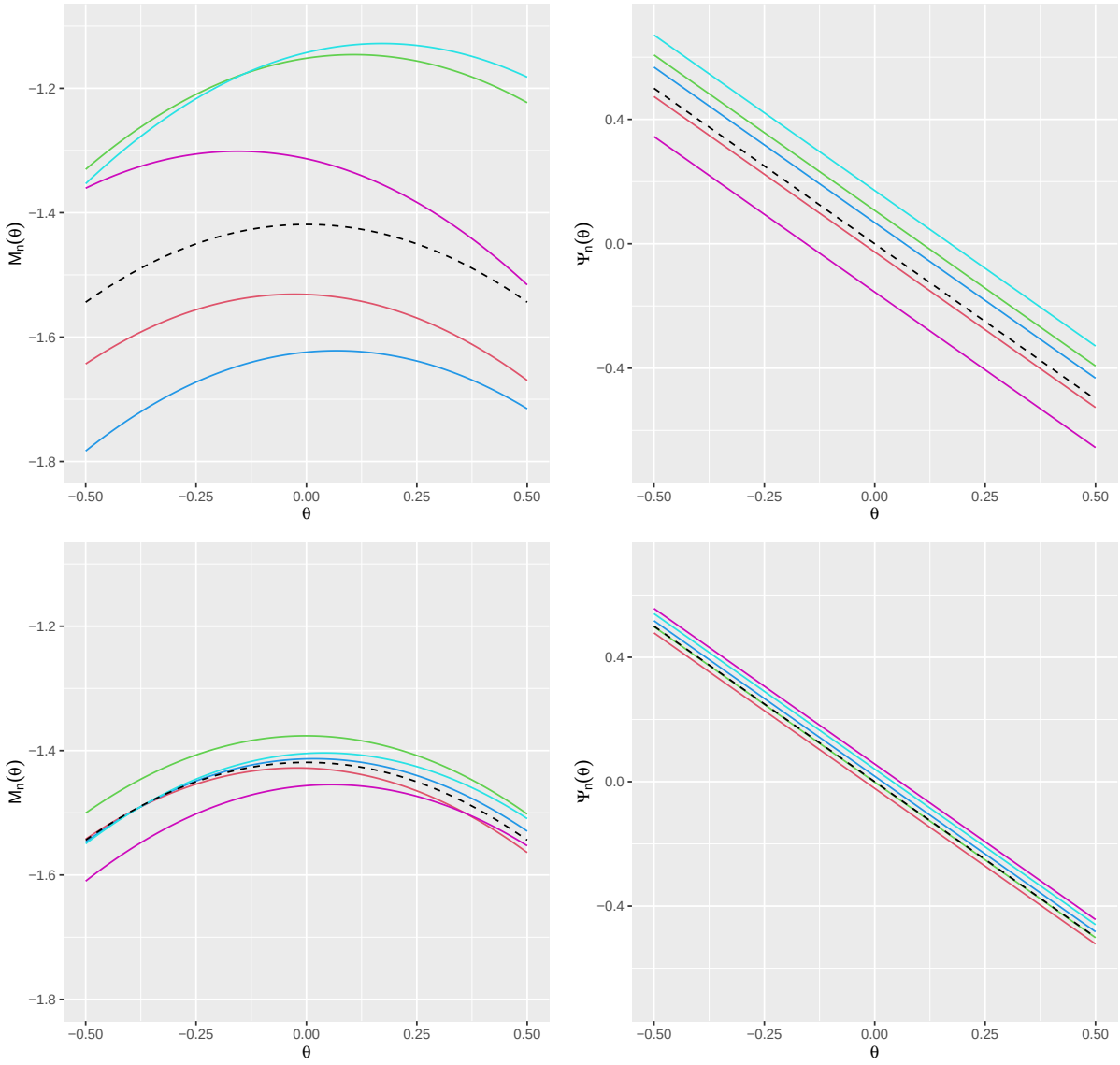

Figure 1: Random functions: Normal log-likelihood (fixed $\sigma = 1$) as a function of mean parameter θ (left) and its derivative with respect to θ (right). Each curve corresponds to a different set of samples (top: $n = 10$; bottom: $n = 1000$) generated from a standard normal distribution ($\theta = 0$). Dashed lines correspond to $M(\theta) = Pm_\theta$ and $\Psi(\theta) = P\psi_\theta$.