A short note on M-estimation

Tong Zhou

January 2, 2021

This note is mainly aiming to provide an overview of asymptotic properties of M-estimators.

1 Maximum Likelihood Estimation

The MLE enjoys many attractive features: it is *consistence, equivariant, asymptotically Normal, asymptotically optimal (efficient)* and *approximately Bayes estimator*, under certain regularity conditions. Those regularity conditions dictate smoothness conditions on $p_\theta(X)$. Unless otherwise stated, we shall tacitly assume that these conditions hold.

For a generic dominated family $\mathcal{P} = \{p_\theta \colon \theta \in \Theta\}$, suppose $X_1, \cdots, X_n \overset{iid}{\sim} p_\theta$. The **maximum likelihood estimator (MLE)** is

$$\begin{aligned} \widehat{\theta}_n &\in \arg\max_{\theta\in\Theta} p_\theta(X) \\ &= \arg\max_{\theta\in\Theta} \ell_n(\theta; X) \end{aligned}$$

where $\ell_n(\theta; X) = \log p_\theta(X) = \sum_{i=1}^n p_\theta(X_i)$.

Remark.

1. The maximizer may not exist.
2. The maximizer may not be unique.

1.1 Consistency of MLE

Suppose θ_0 is the true value of the model. We hope to have

$$\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta_0.$$

Recall the KL-divergence:

$$\mathsf{KL}(p_{\theta_0} \,||\, p_\theta) = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)}\right]$$

By Jensen's inequality:

$$\begin{aligned}
-\mathsf{KL}(p_{\theta_0} \,||\, p_\theta) &= \mathbb{E}_{\theta_0}\left[\log \frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right] \\
&\leqslant \log \mathbb{E}_{\theta_0}\left[\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)}\right] \\
&= \log \int p_{\theta_0}(x)\frac{p_\theta(x)}{p_{\theta_0}(x)}\,\mathrm{d}\mu(x) \\
&= 0
\end{aligned}$$

Unless $p_\theta = p_{\theta_0}$, we have

$$-\mathsf{KL}(p_{\theta_0} \,||\, p_\theta) < 0$$

Let $\ell(\theta; X_i) = \log p_\theta(X_i)$ and let $\bar{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^n (\ell(\theta; X_i) - \ell(\theta_0; X_i))$. Then

$$\max_{\theta\in\Theta} \ell_n(\theta) \iff \max_{\theta\in\Theta} \bar{W}_n(\theta).$$

Also, for any $\theta \in \Theta$, by the law of large numbers, letting $W_i(\theta) = \ell(\theta; X_i) - \ell(\theta_0; X_i)$,

$$\bar{W}_n(\theta) \xrightarrow{\mathsf{P}} \mathbb{E}_{\theta_0}[W_i(\theta)] = -\mathsf{KL}(p_{\theta_0}||p_\theta) < 0.$$

To prove consistency, we need this convergence to be uniform over $\theta \in \Theta$.

We split the proof into 3 steps:

1. Prove a general result on consistency of MLE.
2. Prove consistency under the compactness of Θ.
3. Prove consistency without relying on the compactness of Θ.

1.1.1 A General Consistency Result

The result relies on two conditions:

Theorem 1. *Let θ_0 denote the true value of θ. Define*

$$\bar{W}_n(\theta) = \frac{1}{n}\sum_{i=1}^n W_i(\theta).$$

where $W_i(\theta) = \ell(\theta; X_i) - \ell(\theta_0; X_i)$.

Define $W(\theta) = -\mathsf{KL}(p_{\theta_0} \,||\, p_\theta)$.

Suppose that

1.

$$\sup_{\theta\in\Theta}\left|\bar{W}_n(\theta) - W(\theta)\right| \xrightarrow{\mathsf{P}} 0$$

2. *For any* $\varepsilon > 0$*,*

$$\sup_{\theta:\ |\theta-\theta_0|\geqslant\varepsilon} W(\theta) < W(\theta_0)$$

Then the MLE $\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta_0$.

Proof. We need to show $\forall\varepsilon > 0, \mathbb{P}\left(\left|\widehat{\theta}_n - \theta_0\right| \geqslant \varepsilon\right) \to 0$.

By assumption 3, for any $\varepsilon > 0$, there exists $\delta > 0$ such that for any $\theta \in \{\theta \in \Theta:\ |\theta - \theta_0| \geqslant \varepsilon\}$

$$W(\theta) + \delta < W(\theta_0) \equiv 0.$$

We can construct a relation between two events:

$$\{\theta \in \Theta:\ |\theta - \theta_0| \geqslant \varepsilon\} \subset \{\theta \in \Theta:\ W(\theta) + \delta < W(\theta_0)\}$$

This monotonicity implies :

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta_0\right| \geqslant \varepsilon\right) \leqslant \mathbb{P}\left(W(\theta_0) - W\left(\widehat{\theta}_n\right) > \delta\right) \tag{1}$$

So it suffices to bound the RHS to be arbitrarily small. Observe that

$$\begin{aligned} W(\theta_0) - W(\widehat{\theta}_n) &= W(\theta_0) - \bar{W}_n(\theta_0) + \bar{W}_n(\theta_0) - W(\widehat{\theta}_n) \\ &\leqslant W(\theta_0) - \bar{W}_n(\theta_0) + \bar{W}_n(\widehat{\theta}_n) - W(\widehat{\theta}_n) \\ &\leqslant \underbrace{W(\theta_0) - \bar{W}_n(\theta_0)}_{\xrightarrow{\mathsf{P}} 0 \text{ by LLN}} + \underbrace{\sup_{\theta\in\Theta}\left|\bar{W}_n(\theta) - W(\theta)\right|}_{\xrightarrow{\mathsf{P}} 0 \text{ by assumption 2}} \end{aligned}$$

It implies:

$$W(\theta_0) - W(\widehat{\theta}_n) \xrightarrow{\mathsf{P}} 0$$

or equivalently in Equation (1)

$$\mathbb{P}\left(W(\theta_0) - W\left(\widehat{\theta}_n\right) > \delta\right) \to 0,\ \ n \to \infty.$$

It follows that

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta_0\right| \geqslant \varepsilon\right) \to 0, n \to \infty$$

or

$$\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta_0$$

$\square$

Remark.

1. The identifiability is not in need.
2. ULLN is needed.
3. Assumption 2 restricts the behavior of $W(\theta)$ outside the neighborhood $B_\varepsilon(\theta_0)$.

1.1.2 Consistency with compact Θ

2 M-estimation

In this part, we stand at a high level to discuss the asymptotic properties of general M-estimators. Before diving into that, we first introduce uniform law of large numbers for a generic class of functions.

Definition: Let $\mathscr{F}$ be a collection of functions $f : \mathcal{X} \to \mathbb{R}$. Then $\mathscr{F}$ satisfies a uniform law of large numbers (ULLN) for distribution P if

$$\|P_n - P\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} |P_n f - P f| \xrightarrow{\mathsf{P}} 0,$$

where $Pf = \int f \,\mathrm{d}P$ and $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution of sample $\{X_1, \cdots, X_n\}$.

Remark.

The MLE fits this framework, in Theorem 1:

- $P_n f = \frac{1}{n}\sum_{i=1}^n W_i(\theta)$.
- $Pf = \mathbb{E}[W_i(\theta)] = -\mathsf{KL}(p_{\theta_0} \,\|\, p_\theta)$.
- $f = W_i(\theta) = \ell(\theta; X_i) - \ell(\theta_0; X_i)$.
- $\mathscr{F} = \{p_\theta : \theta \in \Theta\}$.

Theorem 2. *If $\mathscr{F} = \{\ell_\theta\}_{\theta \in \Theta}$ satisfies a ULLN and the sequence of estimators $\left\{\widehat{\theta}_n\right\}_n$ satisfies*

$$R_n\left(\widehat{\theta}_n\right) \leqslant \inf_{\theta \in \Theta} R(\theta) + o_{\mathbb{P}}(1).$$

Also, for all $\varepsilon > 0$, there exists some $\delta > 0$, such that

$$R(\theta) \geqslant R(\theta^*) + \delta, \quad \textit{whenever } d(\theta, \theta^*) \geqslant \varepsilon.$$

Then $\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta^*$.

Proof. *Observe that we have the following monotonicity relation:*

$$\{\theta : d(\theta, \theta^*) \geqslant \varepsilon\} \subset \{\theta : R(\theta) \geqslant R(\theta^*) + \delta\}.$$

We only need to show:

$$\mathbb{P}\left(R(\widehat{\theta}_n) - R(\theta^*) \geqslant \delta\right) \to 0.$$

Further observe that

$$\begin{aligned}
\delta \leqslant R(\widehat{\theta}_n) - R(\theta^*) &= R(\widehat{\theta}_n) - R_n(\widehat{\theta}_n) + R_n(\widehat{\theta}_n) - R(\theta^*) \\
&\leqslant \sup_{\theta \in \Theta} |R(\theta) - R_n(\theta)| + \inf_{\theta \in \Theta} R(\theta) + o_{\mathbb{P}}(1) - R(\theta^*) \\
&\leqslant \sup_{\theta \in \Theta} |R(\theta) - R_n(\theta)| + o_{\mathbb{P}}(1) \\
&\xrightarrow{\mathsf{P}} 0
\end{aligned}$$

So for any $\varepsilon > 0$

$$\mathbb{P}\left(d(\widehat{\theta}_n, \theta^*) \geqslant \varepsilon\right) \to 0,$$

$$\widehat{\theta}_n \xrightarrow{\mathsf{P}} \theta^* m$$

$\square$