Stat210B: Theoretical Statistics **Lecture Date: February 15, 2007**

Applications of ULLNs: Consistency of M-estimators

Lecturer: Michael I. Jordan *Scribe: Blaine Nelson*

1 M and Z-estimators (van der Vaart, 1998, Section 5.1, p. 41–54)

Definition 1 (M-estimator). An estimator $\hat{\theta}_n$ defined as a maximizer of the expression:

$$M_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i) \tag{1}$$

for some function $m_\theta(\cdot)$. If there is a unique solution, the estimator can be expressed simply as

$$\hat{\theta}_n = \text{argmax}_{\theta\in\Theta} M_n(\theta) \ .$$

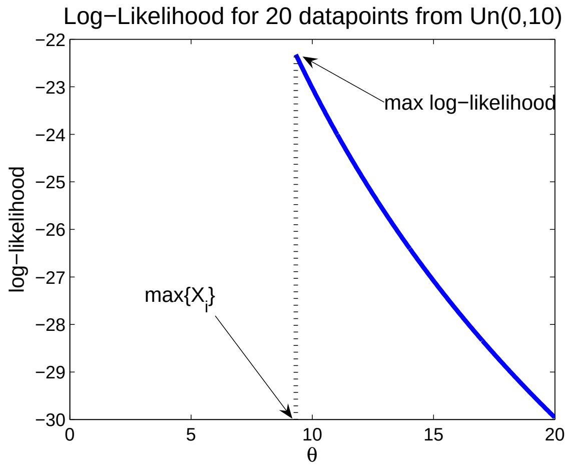
Definition 2 (Z-estimator (estimating equations)). An estimator $\hat{\theta}_n$ that can be expressed as the *root* of the expression:

$$\Phi_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} \phi_\theta(X_i)$$

for some function $\phi_\theta(\cdot)$; that is, a solution to

$$\Phi_n\left(\hat{\theta}_n\right) = 0$$

M-estimators first were introduced in the context of robust estimation by Peter J. Huber as a generalization of the *maximum likelihood estimator* (MLE): $m_\theta(x) = \log p_\theta(x)$. In the literature, they are often confused with Z-esimators because of the relationship between optimization and differentiation. In fact under certain conditions, they are equivalent via the relationship $\phi_\theta(x) = \nabla_\theta[m_\theta(x)]$. If m_θ is everywhere differentiable w.r.t. θ then the M-estimator is a Z-estimator. A simple example where this fails is the estimation of the parameter θ for the distribution $\text{Un}(0,\theta)$. In this model, the log-likelihood is discontinuous in θ but the MLE is well defined as $\hat{\theta}_n = \max\{X_i\}_{i=1}^n$, which occurs at this discontinuity as show in the following figure:



As is clear, the log-likelihood is $-\infty$ before the MLE and decreasing after it. Hence, the maximum of the log-likelihood occurs at this point of discontinuity even though the derivative is not 0 there (it is not defined).

2 Consistency of M-estimators (van der Vaart, 1998, Section 5.2, p. 44–51)

Definition 3 (Consistency). An estimator is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta_0$ (alternatively, $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$) for any $\theta_0 \in \Theta$, where θ_0 is the true parameter being estimated.

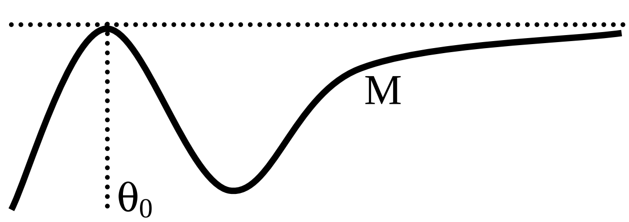
Theorem 4. *(van der Vaart, 1998, Theorem 5.7, p. 45) Let M_n be random functions and M be a fixed function such that $\forall\ \epsilon > 0$:*

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0 \tag{2}$$

$$\sup_{\{\theta \mid d(\theta, \theta_0) \geq \epsilon\}} M(\theta) < M(\theta_0) \tag{3}$$

Then, any sequence $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$ converges in probability to θ_0.

Notice, condition (2) is a restriction on the random functions M_n, whereas condition (3) ensures that θ_0 is a *well-separated* maximum of M; i.e., only θ close to θ_0 achieve a value $M(\theta)$ close to the maximum (See figure below):



Finally it is worth noting that sequences $\hat{\theta}_n$ that *nearly maximize* M_n (i.e., $M_n(\hat{\theta}_n) \geq sup_\theta M_n(\theta) - o_p(1)$) meet the above requirement on $\hat{\theta}_n$.

Proof. We are assuming that our $\hat{\theta}_n$ satisfies, $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_p(1)$. Then, uniform convergence of M_n to M implies

$$\begin{aligned}
\Rightarrow \quad & M_n(\theta_0) \xrightarrow{P} M(\theta_0) \\
\Rightarrow \quad & M_n(\hat{\theta}_n) \geq M(\theta_0) - o_p(1) \\
\Rightarrow \quad & M(\theta_0) \leq M_n(\hat{\theta}_n) + o_p(1) \\
\Rightarrow \quad & M(\theta_0) - M(\hat{\theta}_n) \leq M_n(\hat{\theta}_n) - M(\hat{\theta}_n) + o_p(1) \\
& \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + o_p(1) \\
& \xrightarrow{P} 0 \quad \text{(by condition (2))}
\end{aligned}$$

Now, by condition (3), $\forall\ \epsilon > 0$, $\exists\ \eta$ such that $M(\theta) < M(\theta_0) - \eta$ is satisfied $\forall\ \theta : d(\theta, \theta_0) \geq \epsilon$. Thus $\{d(\hat{\theta}_n, \theta_0) \geq \epsilon\} \subseteq \{M(\hat{\theta}_n) < M(\theta_0) - \eta\}$.

$$\Rightarrow P\left(d(\hat{\theta}_n, \theta_0) \geq \epsilon\right) \leq \underbrace{P\left(M(\hat{\theta}_n) < M(\theta_0) - \eta\right)}_{\xrightarrow{P} 0 \quad \text{(as shown above)}}$$

□

The primary drawback of this approach is that it requires the metric entropy to achieve condition (2).

3 Consistency of the MLE (non-parametric)

We assume that we have n i.i.d. samples from some (unknown) distribution P; i.e., $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. Further, we assume P has a density $p_0 = \frac{dP}{d\mu}$. For the family of densities, $\mathcal{P}$, we will consider the *maximum likelihood estimator* (MLE) amongst $\mathcal{P}$ as

$$\hat{p}_n = \text{argmax}_{p \in \mathcal{P}} \int \log p dP_n$$

where $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$—the empirical distribution. To further formalize this, we consider the following definitions.

Definition 5 (**Kullback-Leibler (KL)-divergence).** The Kullback-Leibler divergence between two densities is defined as,

$$K(p_0, p) = \int \log \frac{p_0(x)}{p(x)} dP(x) \ .$$

(Recall, $K(p_0, p)$ is always non-negative and is 0 if and only if $p_0(x) = p(x)$ almost everywhere.)

Definition 6 (**Maximum Likelihood Estimator (MLE)).** The maximum-likelihood estimator $\hat{p}_n$ is the minimizer of

$$\int \log \frac{p_0(x)}{\hat{p}_n(x)} dP(x)$$

where P has a density p_0. This implies

$$\int \log \frac{\hat{p}_n}{p_0} dP_n \leq 0 \tag{4}$$

Given these definitions, we now derive a bound on the KL-divergence between the true density p_0 and the MLE $\hat{p}_n$:

$$
\begin{aligned}
\Rightarrow \quad & \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) \leq 0 \\
\Rightarrow \quad & \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) - K\left(p_0, \hat{p}_n\right) + K\left(p_0, \hat{p}_n\right) \leq 0 \\
\Rightarrow \quad & K\left(p_0, \hat{p}_n\right) \leq \left| \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP_n(x) - \int \log \frac{p_0(x)}{\hat{p}_n(x)} dP(x) \right| \\
& = \left| \int \log \frac{\hat{p}_n(x)}{p_0(x)} d(P_n - P)(x) \right| .
\end{aligned}
$$

Thus, we need a ULLN for the family of functions: $\mathfrak{F} = \{\log \frac{p}{p_0} \{p_0 > 0\} \mid p \in \mathcal{P}\}$. To this end, we use the following distance measure:

Definition 7 (**Hellinger Distance**).

$$
h\left(p_1, p_2\right) = \left(\frac{1}{2} \int \left(p_1^{1/2}(x) - p_2^{1/2}(x) \right)^2 d\mu(x) \right)^{\frac{1}{2}}
$$

Unlike the KL-divergence, Hellinger distance is a proper distance metric (non-negative, symmetric, transitive, and 0 if and only if $p_1 = p_2$ almost everywhere). Moreover, Hellinger is appealing as the square-root of a density lies in $\mathcal{L}_2$. Further we have the following:

Lemma 8.

$$
h^2(p_1, p_2) \leq \frac{1}{2} K(p_1, p_2)
$$

Proof. We use the inequality $\log(x) \leq x - 1$ in the form $\frac{1}{2}\log(v) \leq v^{1/2} - 1$. This gives the following:

$$
\begin{aligned}
\Rightarrow \quad & \frac{1}{2} \log \frac{p_2(x)}{p_1(x)} \leq \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} - 1 \\
\Rightarrow \quad & \frac{-1}{2} K(p_1, p_2) \leq \int_{p_1 > 0} \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} p_1(x) \mu(dx) - 1 \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \underbrace{\frac{1}{2}}_{\frac{1}{2}\int_{p_1>0} p_1(x)\mu(dx)} + \underbrace{\frac{1}{2}}_{\frac{1}{2}\int_{p_1>0} p_2(x)\mu(dx)} - \int_{p_1 > 0} \frac{p_2^{1/2}(x)}{p_1^{1/2}(x)} p_1(x) \mu(dx) \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \int_{p_1 > 0} \frac{1}{2} p_1(x) - p_1^{1/2}(x) p_2^{1/2}(x) + \frac{1}{2} p_2(x) \mu(dx) \\
\Rightarrow \quad & \frac{1}{2} K(p_1, p_2) \geq \underbrace{\frac{1}{2} \int \left(p_1^{1/2}(x) - p_2^{1/2}(x) \right)^2 \mu(dx)}_{= h^2(p_1, p_2)}
\end{aligned}
$$

□

Unfortunately, though, $\mathfrak{F}$ is hard to work with (p's are not bounded away from 0). Instead we will work with the family

$$
\mathfrak{G} \triangleq \{ \frac{1}{2} \log \frac{p + p_0}{2p_0} \{p_0 > 0\} \mid p \in \mathcal{P} \}
$$

which is bounded below by $\frac{1}{2}\log\frac{1}{2}$.

Lemma 9.
$$h^2\left(\frac{\hat{p}_n+p_0}{2}, p_0\right) \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} d(P_n - P)$$

Proof. Concavity of the logarithm implies

$$\begin{aligned}
\Rightarrow \quad & \log\frac{\hat{p}_n+p_0}{2} \geq \frac{1}{2}\log\hat{p}_n + \frac{1}{2}\log p_0 \\
\Rightarrow \quad & \log\frac{\hat{p}_n+p_0}{2} - \log p_0 \geq \frac{1}{2}\log\hat{p}_n - \frac{1}{2}\log p_0 \\
\Rightarrow \quad & \log\frac{\hat{p}_n+p_0}{2p_0}\{p_0>0\} \geq \frac{1}{2}\log\frac{\hat{p}_n}{p_0}\{p_0>0\}
\end{aligned}$$

Now, by the definition of the MLE (Eq. (4)):

$$\begin{aligned}
\Rightarrow \quad & 0 \leq \int_{p_0>0} \frac{1}{4}\log\frac{\hat{p}_n}{p_0} dP_n \\
\Rightarrow \quad & 0 \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} dP_n \\
& = \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} d(P_n-P) + \underbrace{\int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} dP}_{=-\frac{1}{2}K\left(p_0, \frac{\hat{p}_n+p_0}{2}\right)} \\
& \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} d(P_n-P) - h^2\left(\frac{\hat{p}_n+p_0}{2}, p_0\right) \quad \text{(by Lemma 8)} \\
\Rightarrow \quad h^2\left(\frac{\hat{p}_n+p_0}{2}, p_0\right) & \leq \int_{p_0>0} \frac{1}{2}\log\frac{\hat{p}_n+p_0}{2p_0} d(P_n-P)
\end{aligned}$$

□

Thus, elements of our family $\mathfrak{G}$ have Hellinger distance 0 that goes to 0. To connect this back to our orginal family $\mathfrak{F}$, we have the following Lemma:

Lemma 10.
$$h^2(p, p_0) \leq 16h^2(\bar{p}, p_0)$$
where $\bar{p} \triangleq \frac{p+p_0}{2}$.

Finally, we arrive at the following Theorem:

Theorem 11. *Let* $\mathfrak{G} = \{\frac{1}{2}\log\frac{\bar{p}}{p_0}\{p_0>0\} \mid p \in \mathcal{P}\}$ *and let* $G = \sup_{g\in\mathfrak{G}}|g|$. *Assume that* $\int G dP < \infty$ *and* $\forall\epsilon>0 \quad \frac{1}{n}H_1(\epsilon, P_n, \mathfrak{G}) \xrightarrow{P} 0$, *then*
$$h(\hat{p}_n, p_0) \xrightarrow{a.s.} 0$$

Example 12 (Logistic Regression for nonparameteric links). We are given data pairs: (Y_i, Z_i) and we assume the conditional distribution of Y follows a particular functional form:

$$P(Y = 1|Z = z) = F_{\theta_0}(z)$$

where F_θ is an increasing function of z for every $\theta \in \Theta$ and $\theta_0 \in \Theta$ is the true parameter.

Let μ be (counting measure on $\{0, 1\}) \times Q$ where Q is the distribution of Z. Now, the family of joint densities we obtain is

$$\mathcal{P} = \{p_\theta(y, z) = yF_\theta(z) + (1 - y)(1 - F_\theta(z))\}$$

which has the following properties:

- $\sup_{p \in \mathcal{P}} p \leq 1$.
- $H_B(\epsilon, \mu, \mathcal{P}) \leq A\epsilon^{-1}$ (for increasing functions).

Hence we have

$$h(\hat{p}_n, p_0) \xrightarrow{P} 0$$

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.