Lecture #10

- Today, we finally establish the consistency of MLE.
- Recall that for a parametric family of densities $\{f(x;\theta)\}_{\theta \in \Theta}$, and data $\{x_i\}_{i=1}^n$, the associated likelihood function is $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(x_i;\theta)$.

- Recall that for two densities f, g, the (non-symmetric) KL distance is $D_{KL}(f, g)$

$$= \int_{\mathbb{R}} f(x) \cdot \log\left(\frac{f(x)}{g(x)}\right) dx.$$

- Let $\theta_1, \theta_2 \in \Theta$. We will use $D_{KL}(\theta_1, \theta_2)$ as shorthand for $D_{KL}(f(x;\theta_1), f(x;\theta_2))$

$$= \int_{\mathbb{R}} f(x;\theta_1) \log\left(\frac{f(x;\theta_1)}{f(x;\theta_2)}\right) dx.$$

- Let us suppose our data is drawn iid from $f(x;\theta^*)$, i.e. θ^* is the (hidden/latent) true parameter. We want to show $\hat{\theta}_{MLE} \xrightarrow{\mathbb{P}} \theta^*$. This will establish consistency of MLE.

- Notice that maximizing $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(x_i;\theta)$

$$\Updownarrow$$

maximizing $\log(\mathcal{L}_n(\theta)) = \sum_{i=1}^n \log(f(x_i;\theta))$

$$\Updownarrow$$

maximizing $M_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^*)}\right)$, because θ^* is constant and $\log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^*)}\right)$

$= \log(f(x_i;\theta)) - \log(f(x_i;\theta^*))$.

This looks a lot a discrete version of a KL distance. We can now state and prove a result guaranteeing consistency of MLE.

Theorem (Consistency of MLE): Let $\{X_i\}_{i=1}^n \overset{iid}{\sim} f(x;\theta^*)$. Let $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^*)}\right)$.

Let $M(\theta) = -D_{KL}(\theta^*,\theta) = -\int_{\mathbb{R}} f(x;\theta^*) \log\left(\frac{f(x;\theta^*)}{f(x;\theta^*)}\right) dx$. Assume

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0 \quad \text{and} \quad \forall \varepsilon>0, \sup_{\{\theta \mid |\theta-\theta^*|\geq\varepsilon\}} M(\theta) < M(\theta^*)$$

Then if $\hat{\theta}_n$ is the MLE for $\hat{\theta}^*$, $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$, i.e. $\hat{\theta}_n$ is consistent for θ.

Proof: We want to show $\forall \varepsilon>0$, $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n-\theta^*|>\varepsilon)=0$. Well, notice that since $\hat{\theta}_n$ is the MLE, it maximizes $M_n(\theta)$. In particular, $M_n(\hat{\theta}_n) \geq M_n(\theta^*)$.

Thus,
$$\begin{aligned}
M(\theta^*) - M(\hat{\theta}_n) &= M_n(\theta^*) - M_n(\theta^*) + M(\theta^*) - M(\hat{\theta}_n) \\
&= \underbrace{M_n(\theta^*)}_{\leq} - M(\hat{\theta}_n) + M(\theta^*) - M_n(\theta^*) \\
&\leq \left[M_n(\hat{\theta}_n) - M(\hat{\theta}_n)\right] + \left[M(\theta^*) - M_n(\theta^*)\right] \\
&\leq 2\cdot \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)|
\end{aligned}$$

We assume $\sup_{\theta\in\Theta} |M_n(\theta)-M(\theta)| \xrightarrow{\mathbb{P}} 0$, so we conclude $\left[M(\theta^*) - M(\hat{\theta}_n)\right] \xrightarrow{\mathbb{P}} 0$.

In particular, for any $\delta>0$, $\lim_{n\to\infty} \mathbb{P}\left(M(\hat{\theta}_n) < M(\theta^*) - \delta\right) = 0$ (☆)

Now, for any $\varepsilon > 0$, we know/assumed $\sup_{\{\theta \mid |\theta - \theta_*| \geq \varepsilon\}} M(\theta) < M(\theta^*)$.

In particular, for any $\varepsilon > 0$, $\exists\, \delta > 0$ s.t. $M(\theta) < M(\theta^*) - \delta$ for all θ satisfying $|\theta - \theta^*| \geq \varepsilon$. We conclude $\forall \varepsilon > 0$, $\mathbb{P}(|\hat{\theta}_n - \theta_*| > \varepsilon)$

$$\leq \mathbb{P}(M(\hat{\theta}_n) < M(\theta^*) - \delta) \quad \text{for some } \delta > 0 \text{. independent of } n.$$

Throwing down limits and applying ✪ yields

$$\lim_{n \to \infty} \mathbb{P}(|\hat{\theta}_n - \theta^*| > \varepsilon)$$

$$\leq \lim_{n \to \infty} \mathbb{P}(M(\hat{\theta}_n) < M(\theta^*) - \delta)$$

$$= 0,$$

and we have thus shown $\forall \varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}(|\hat{\theta}_n - \theta^*| > \varepsilon) = 0$, i.e. $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$. ∎

<u>Remarks</u>: This theorem required two technical assumptions.

(I) $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$. This is basically about the discrete-to-continuum convergence of the discrete KL distance (M_n) to the continuous one (M).

(II) $\forall \varepsilon > 0$, $\sup_{\{\theta \mid |\theta - \theta^*| \geq \varepsilon\}} M(\theta) < M(\theta^*)$. Recalling that $M(\theta) = -D_{KL}(\theta^*, \theta)$,

This is essentially saying $\sup_{\{\theta \mid |\theta - \theta^*| \geq \varepsilon\}} D_{KL}(\theta, \theta^*) > D_{KL}(\theta^*, \theta^*) = 0$, i.e.

from the standpoint of D_{KL}, there is only one ~~[illegible]~~ density equal to $f(x; \theta^*)$.

- After all the work that went into showing MLE is consistent, showing it is equivariant is rather straightforward.

<u>Theorem (Equivariance of MLE)</u>: Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}_n$ be the MLE of θ. Then the MLE of τ is $\hat{\tau}_n = g(\hat{\theta}_n)$ when g is invertible.

<u>Proof</u>: Let $h = g^{-1}$, which we assumed to exist. Set $\hat{\tau}_n = g(\hat{\theta}_n)$

$$\Rightarrow h(\hat{\tau}_n) = \hat{\theta}_n.$$

Indeed, for any τ, $\tau = g(\theta)$

$$\Leftrightarrow h(\tau) = \theta.$$

So, we can plug this into the likelihood function:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

$$= \prod_{i=1}^{n} f(x_i; h(\tau))$$

This is the natural definition for likelihood w.r.t. τ

$$= \prod_{i=1}^{n} f(x_i; \tau)$$

$$= \mathcal{L}_n(\tau).$$

Thus, for any τ, $\mathcal{L}_n(\tau) = \mathcal{L}_n(\theta) \leq \mathcal{L}_n(\hat{\theta}_n) = \mathcal{L}_n(\hat{\tau})$. ■

- Almost a tautology... we are basically "renaming" θ in an invertible way.

ex: Let $X_1, \dots, X_n \sim \text{Unif}(0,\theta)$. What is the MLE for $\tau = \ln(\theta)$? Well,

$$\hat{\theta}_{MLE} = \max_{1 \le i \le n} X_i \xRightarrow{\text{equivariance}} \hat{\tau}_{MLE} = \ln\left(\max_{1 \le i \le n} X_i\right)$$

$$= \max_{1 \le i \le n} \ln(X_i), \text{ by monotonicity of } \ln(x).$$

Remark: g invertible is not strictly required.