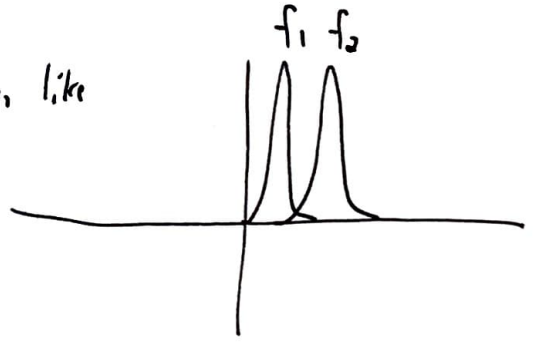
Lecture #9

- Having seen MLE in action on a few examples, we should ask: does it work well generally?
- You will see in the homework that MLE can be biased, even for simple examples.
- It is, however, consistent in a broad setting.
- To show this, we will need a notion of "distance between probability distributions"
- Of course, we could do the ℓ^p norm on the densities, i.e. $\frac{1}{2}\left(\int_{\mathbb{R}} |f_1(x) - f_2(x)|^p \, dx\right)^{1/p}$

is a perfectly sensible metric between densities f_1 and f_2, but does it really capture differences in the underlying random variables? Not really!

ex: Let X_1, X_2 be r.v. with densities like



Even though a typical draw from X_1 is close to a typical draw from X_2,

$\frac{1}{2}\int_{\mathbb{R}} |f_1(x) - f_2(x)|^p \Big]^{1/p} \approx 1$, i.e. as far apart as this metric allows!

- So, there is some subtlety in thinking about metrics to measure distance between r.v.
- The Kullback-Leibler distance is an alternative metric.

Def'n: Let $f, g : \mathbb{R} \to \mathbb{R}$ be two density functions. The Kullback-Leibler distance between f and g is $D_{KL}(f,g) := \int_{\mathbb{R}} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$.

- When is this small? When $f(x)$ different from $g(x) \Rightarrow f(x)$ is small! Basically, we need $f(x) \approx 0$ to cancel with $\log\left(\frac{f(x)}{g(x)}\right)$ large.

- What's the motivation for this? Suppose there are two densities, $f(x)$ and $g(x)$. I secretly pick one and sample $\{X_1, \dots, X_n\}$ i.i.d from it. If I know the formulae for $f(x)$ and $g(x)$, how can I use the observed data to guess which is the density selected?

Well, a natural thing is to choose the density which maximizes the likelihood.

So, I will choose $f(x)$ iff $\prod_{i=1}^n f(x_i) > \prod_{i=1}^n g(x_i)$

$$\Leftrightarrow \prod_{i=1}^n \frac{f(x_i)}{g(x_i)} > 1.$$

$$\Leftrightarrow \sum_{i=1}^n \log\left(\frac{f(x_i)}{g(x_i)}\right) > 0.$$

So, letting $R(x) = \log\left(\frac{f(x)}{g(x)}\right)$ consider the expectation under $f(x)$: $\int_{\mathbb{R}} f(x) \cdot \log\left(\frac{f(x)}{g(x)}\right) dx$. By LLN, $\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i)}{g(x_i)}\right) \to \int_{\mathbb{R}} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx$

when the $\{x_i\}_{i=1}^n$ are sampled from $f(x)$. So, as long as $D_{KL}(f,g) > 0$, $\frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i)}{g(x_i)}\right) > 0$ for n large enough with high probability, so our scheme for determining

The hidden density will work!

- Let's use KL distance to prove MLE is consistent.
- Let θ^* be the true optimal θ, i.e. we want to show $\hat{\theta}_{MLE} \xrightarrow{P} \theta^*$.
- So, our data is $\{x_i\}_{i=1}^n \overset{iid}{\sim} f(x_i;\theta^*)$, and we use MLE to estimate $\hat{\theta}_{MLE}$.
- Recall that $\hat{\theta}_{MLE} = \arg\max_\theta \log(\mathcal{L}_n(\theta))$

$$= \arg\max_\theta \sum_{i=1}^n \log(f(x_i;\theta))$$

- Let $M_n(\theta) = \frac{1}{n}\left[\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta^*)\right]$, so that $M_n(\theta)$ has the same maximizer as $\mathcal{L}_n(\theta)$, namely θ_{MLE}. But after some algebra,

$$M_n(\theta) = \frac{1}{n}\left[\sum_{i=1}^n \log(f(x_i;\theta)) - \sum_{i=1}^n \log(f(x_i;\theta^*))\right]$$

$$= \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^*)}\right).$$

- Letting $n \to \infty$, by LLN, $M_n(\theta)$

$$= \frac{1}{n}\sum_{i=1}^n \log\left(\frac{f(x_i;\theta)}{f(x_i;\theta^*)}\right)$$

$$\overset{\text{LLN}}{\approx} \int_{\mathbb{R}} \log\left(\frac{f(x;\theta)}{f(x;\theta^*)}\right) f(x;\theta^*)\,dx$$

(noting that our data is sampled i.i.d from $f(x;\theta^*)$)

$$= -D_{KL}\left(f(x;\theta), f(x;\theta^*)\right).$$

· So, since $M_n(\theta)$ is maximized at $\hat{\theta}_{MLE}$, so is $-D_{KL}(f(x;\theta), f(x;\theta^*))$. ④

But we note that: (1.) $D_{KL}(f\|g) \geq 0$, for all densities f, g
(HW)

(2.) $D(f\|f) = 0$, for all f densities

So, the minimizer of $D_{KL}(f(x;\theta), f(x;\theta^*))$ occurs at $\theta = \hat{\theta}_{MLE}$; thus, we can "conclude" as $n \to \infty$, $\hat{\theta}_{MLE} \to \theta^*$, since this is how we make

$D_{KL}(f(x;\theta), f(x;\theta^*)) = 0$.

· This is all at the intuitive level! We prove it next time.

[redacted]

ex: Let $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$

$g(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-[x-\delta]^2}{2}\right)$

be densities for $N(0,1)$ and $N(\delta, 1)$, respectively. Then

$$D_{KL}(f\|g) \overset{\text{def}}{=} \int_{\mathbb{R}} \log\left(\frac{f(x)}{g(x)}\right) f(x)\,dx$$

$$= \int_{\mathbb{R}} \log\left(\frac{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-[x-\delta]^2}{2}\right)}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx$$

$$= \int_{\mathbb{R}} \left[\frac{-x^2}{2} + \frac{[x-\delta]^2}{2}\right] \cdot \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx$$

$$= \frac{1}{2} \int_{\mathbb{R}} \left[-2x\delta + \delta^2\right] \cdot \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx$$

$$= \frac{-\delta}{\sqrt{2\pi}} \int_{\mathbb{R}} x \exp\left(\frac{-x^2}{2}\right) dx + \frac{\delta^2}{2} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx$$

0 by symmetry

1 since it $\int$ of a density

$$= \frac{\delta^2}{2}$$

. So when the Gaussians are "close" ($\delta \simeq 0$), the KL distance is small!