Summary: MLE as a Random Variable

- MLE is actually a random variable (as a function of the random data)

Why?

- Given some random data i.i.d. drawn from some distribution:

$$\{x_i\} \overset{\text{i.i.d}}{\sim} p(\cdot \mid \theta^*)$$ true parameter

- Then the $\hat{\theta}$ resulted from maximizing the likelihood function, is actually a function of the randomly drawn data:

$$\hat{\theta} = \hat{\Theta}(x_1 \dots x_n)$$

- Since MLE is a random variable, we can use metrics like:

  Bias, Variance, Mean Squared Error (MSE)

  to understand its behavior

What those metrics mean?

- $\underline{\text{Bias}(\hat{\theta})}$ $= \underbrace{E_{\theta^*}[\hat{\Theta}(x_1 \dots x_n)]}_{\text{expected MLE}} - \underbrace{\theta^*}_{\text{true parameter}}$

$$= \int \underbrace{\hat{\theta}}_{\text{MLE}} \cdot \underbrace{f(\hat{\theta})}_{\text{PDF of MLE}} \, d\hat{\theta} - \theta^*$$

Expected value continuous r.v.

$$\boxed{E(X) = \int x \cdot \text{PDF}(x) \, dx}$$

- $\underline{\text{Variance}(\hat{\theta})}$ $= E_{\theta^*}\left[\left(\hat{\Theta}(x_1 \dots x_n) - E_{\theta^*}[\hat{\Theta}(x_1 \dots x_n)]\right)^2\right]$

$$(\text{MLE} - \text{Expected MLE})^2$$

or: $E_{\theta^*}[\hat{\Theta}(x_1 \dots x_n)^2] - E_{\theta^*}[\hat{\Theta}(x_1 \dots x_n)]^2$

$$\text{Expected MLE}^2 - (\text{Expected MLE})^2$$

Variance of r.v.

$$\boxed{\begin{array}{l} \text{Var}(X) = E[(X - EX)^2] \\ \text{or } \text{Var}(X) = E(X^2) - (EX)^2 \end{array}}$$

- $\underline{\text{MSE}(\hat{\theta})}$ $= E_{\theta^*}\left[\left(\underbrace{\hat{\Theta}(x_1 \dots x_n)}_{\text{MLE}} - \underbrace{\theta^*}_{\text{true parameter}}\right)^2\right]$

Relationship between Bias, Variance, Mean Squared Error (MSE)

$$\boxed{MSE = Bias^2 + Variance}$$

Proof:

$$
\begin{aligned}
MSE(x) &= E\left[(X - X^*)^2\right] \\
&= E\left[(X - EX + EX - X^*)^2\right] \\
\langle (x+y)^2 = x^2 + 2xy + y^2 \rangle \quad &= E\left[(X-EX)^2 + (EX - X^*)^2 + 2(X-EX)(EX - X^*)\right] \\
\langle \text{Linearity of expectation} \rangle \quad &= E\left[(X-EX)^2\right] + E\left[(EX - X^*)^2\right] + E\left[2(X-EX)(EX-X^*)\right] \\
&= Var(X) + Bias(X) + 0
\end{aligned}
$$

why?

$$E\left[2(X-EX)\underbrace{(EX - X^*)}_{\text{constant}}\right]$$

constant since:

- EX is a constant
- X^* is a constant

$$
\begin{aligned}
&= 2(EX - X^*) \cdot E\left[(X - EX)\right] \\
&= 2(EX - X^*) \cdot \left[EX - E(EX)\right] \\
&= 2(EX - X^*) \cdot \left[EX - EX\right] \\
&= 2(EX - X^*) \cdot 0
\end{aligned}
$$

What happen under the hood MLE:

The goal is to come up with a value of $\hat{\theta}$ that is as close to θ^* as possible.

Unknown

True parameter: θ^*

Known

Observations: $\{x_i\}_{i=1}^{n}$

$\downarrow$ argmax $L(\theta)$

Estimated parameter (MLE): $\hat{\theta}$

MSE

Unbiased estimators vs Consistent estimators

- Unbiased estimators: $Bias(\hat{\theta}) = 0$ or $E(\hat{\theta}) = \theta$

  Expected MLE after many iterations equals true parameter

- Consistent estimators: $MSE(\hat{\theta}) \to 0$ as $n \to \infty$

  The more the data, the smaller the error

- Asymetic Unbiased: $Bias(\hat{\theta}) \to 0$ as $n \to \infty$

  The more the data, the smaller the bias

$$\begin{tikzcd}
Bias(\hat{\theta}) = 0 \arrow[rr, leftrightarrow, "\text{no imply}", "\times"'] & & MSE(\hat{\theta}) \downarrow \text{ as } n \uparrow \arrow[dll, "\text{imply}"] \\
Bias(\hat{\theta}) \downarrow \text{ as } n \uparrow & &
\end{tikzcd}$$

Unbiased $\hat{\theta}$ doesn't mean Consistent $\hat{\theta}$

- Since $MSE = Bias^2 + Variance$
- If $Bias = 0$, Variance might $\uparrow$ as $n \uparrow \infty$:

  So MSE doesn't $\downarrow 0$

Consistent $\hat{\theta}$ doesn't mean Unbiased $\hat{\theta}$

- Since $MSE = Bias^2 + Variance$
- If $MSE \downarrow 0$ as $n \uparrow \infty$, doesn't mean that $Bias = 0$

  because a finite dataset might contain biasness.

<u>BUT:</u> Consistent $\hat{\theta}$ does imply "Asymetic Unbiased"

Example 1: MLE of Gaussian is both Unbiased and Consistent

For $\{x_i\}_{i=1}^{n} \sim \mathcal{N}(\mu, \sigma^2)$, MLE is:

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \end{cases}$$

$$\begin{aligned}
\text{Bias}(\hat{\mu}) &= E(\hat{\mu}) - \mu \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] - \mu \\
&= \frac{1}{n}\sum_{i=1}^{n} E(x_i) - \mu \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu - \mu \\
&= 0
\end{aligned}$$

$\Rightarrow \hat{\mu}$ is Unbiased

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= E(\hat{\mu}^2) - E(\hat{\mu})^2 \\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] - \mu^2
\end{aligned}$$

why? Proved in $\text{Bias}(\hat{\mu})$

$$\begin{aligned}
&= E\left[\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2\right] - \mu^2 \\
&= E\left[\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^2 + 2\sum_{i<j}^{n} x_i x_j\right)\right] - \mu^2 \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n} E(x_i^2) + 2\sum_{i<j}^{n} E(x_i)E(x_j)\right] - \mu^2 \\
&= \frac{1}{n^2}\left[n\cdot(\sigma^2 + \mu^2) + 2\,\frac{n(n-1)}{2}\,\mu^2\right] - \mu^2
\end{aligned}$$

$(a+b+c)^2$
$= a^2+b^2+c^2$
$+ 2(ab + bc + ac)$

$E(x_i x_j)$
$= E(x_i)E(x_j)$
$= \mu^2$

$\text{Var}(x_i) = E(x_i^2) - E(x_i)^2$
$\Leftarrow \sigma^2 = E(x_i^2) - \mu^2$

$\binom{n}{2}$ number of pairs $E(x_i)$ and $E(x_j)$

$$= \frac{1}{n}\sigma^2$$

- $$\begin{aligned}
\text{MSE}(\hat{\mu}) &= \text{Bias}(\hat{\mu})^2 + \text{Var}(\hat{\mu}) \\
&= 0^2 + \frac{1}{n}\sigma^2 \\
&= \frac{1}{n}\sigma^2
\end{aligned}$$

So as $n \uparrow \infty$, $\text{MSE}(\hat{\mu}) \downarrow 0$ $\Rightarrow \hat{\mu}$ is Consistent

- Other conclusions on σ^2:
  - $\hat{\sigma}^2$ is Consistent and Asymetic Unbiased, but not Unbiased
  - However, $\hat{S}^2 = \frac{1}{n-1}\sum^{n}(x_i - \hat{\mu})^2$ is Unbiased