MAXIMUM LIKELIHOOD ESTIMATION

Motivation example:

Given a biased coin whose probability of heads is <u>unknown</u>

$$P(X = H) = \theta$$
$$P(X = T) = 1 - \theta$$

And given these independence <u>observations</u> : H H H H T

<u>Problem:</u> Estimate θ

Solution:

- We find the probability of observing the dataset given some value of parameter θ. For example:
  - $P(HHHHT \mid \theta = 0.9) = \theta^4 (1-\theta)^1$
  - $P(HHHHT \mid \theta = 0.8) = \dots$
  - $P(HHHHT \mid \theta = 0.7) = \dots$

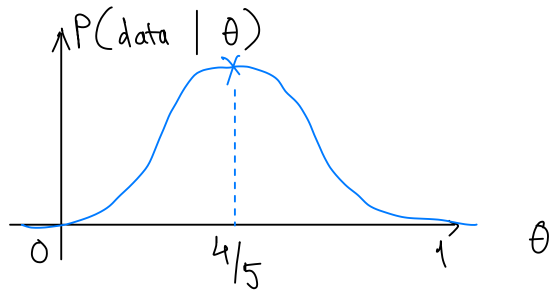
$$\downarrow$$

Continuing this pattern, we some function over θ

$$L(\theta) = P(HHHHT \mid \theta) \rightarrow \text{Likelihood Function}$$
$$= \theta^4 (1-\theta)^1$$

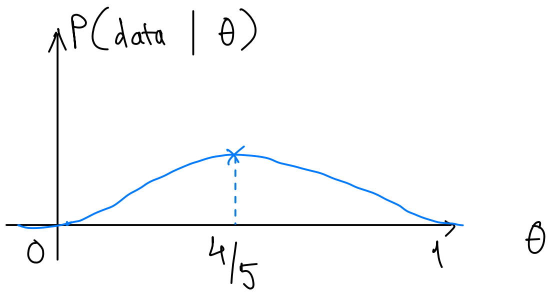
To best estimate θ, we maximize this function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) \rightarrow \text{Maximum Likelihood Estimation } L(\theta)$$
$$= \underset{\theta}{\operatorname{argmax}}\, \theta^4 (1-\theta)^1$$



In practice, we often maximize the log function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \log\left(\theta^4 (1-\theta)^1\right) \rightarrow \text{Log-liklihood Function } \ell(\theta)$$



Solve for best value of θ:

$$\nabla_\theta \left[\log\left(\theta^4 (1-\theta)\right)\right] = 0$$
$$\Leftrightarrow \nabla_\theta \left[4 \log \theta + \log(1-\theta)\right] = 0$$
$$\Leftrightarrow \frac{1}{\theta} 4 + \frac{-1}{1-\theta} = 0$$

$$\boxed{\Rightarrow \hat{\theta} = \frac{4}{5}}$$

Generalized approach of MLE:

- Given a set of observation / dataset $S = \{x_i\}_{i=1}^{n}$
- Independent and identically distributed (i.i.d.) following an unknown distribution, from a parametric family of distributions:

$$\{ p(.\mid\theta) : \theta \in \Theta \}$$

- Problem: Estimate Θ

Solution:

- Write down likelihood function:

$$\begin{aligned} L(\Theta) &= P(S = \{x_1, x_2 \dots x_n\} \mid \theta) && \text{<definition>} \\ &= P(x_1 \mid \Theta) \cdot P(x_2 \mid \Theta) \dots P(x_n \mid \theta) && \text{<i.i.d.>} \\ &= \prod_{i=1}^{n} P(x_i \mid \Theta) \end{aligned}$$

- Then write down the log-likelihood function:

$$\begin{aligned} \ell(\theta) &= \log L(\Theta) \\ &= \log\left(\prod_{i=1}^{n} P(x_i \mid \Theta)\right) \\ &= \sum_{i=1}^{n} \log P(x_i \mid \Theta) \end{aligned}$$

- Define maximum likelihood estimation (MLE):

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\ \ell(\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(x_i \mid \theta) \end{aligned}$$

- Solve the optimization problem, there are 2 ways:
  - Closed form if we have "nice form" function
  - Numerical algorithm otherwise

    a common algorithm is Gradient Ascent

Example: Gaussian Distribution

- Given $\{x_i\}_{i=1}^{n}$ i.i.d. drawn from Gaussian Distribution $N(\mu, \sigma^2)$:

$$p(x \mid \theta) = \frac{1}{\sqrt{(2\pi)}\ \sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad \theta = \{\mu, \sigma\}$$

- Problem: Estimate θ

Solution:

- Write down log-likelihood function:

$$\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{n} \log p(x_i \mid \theta) \\
&= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi}\ \sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right] \\
&= \sum_{i=1}^{n}\left[\log\left(\frac{1}{\sqrt{2\pi}\ \sigma}\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\
&= n \cdot \log\left(\frac{1}{\sqrt{2\pi}\ \sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
&= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

- Define maximum likelihood estimation (MLE):
  - Since $\theta = \{\mu, \sigma\}$:
  - Fist, find $\hat{\mu}$:

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\ \ell(\theta)$$

Solve using closed form:

$$\begin{aligned}
& \nabla_\mu \ell(\theta) = 0 \\
\Leftrightarrow\ & \nabla_\mu\left[-\sum_{i=1}^{n}(x_i-\mu)^2\right] = 0 \\
\Leftrightarrow\ & 2\left[\sum_{i=1}^{n}(x_i-\hat{\mu})\right] = 0 \\
\Leftrightarrow\ & \sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \\
\Leftrightarrow\ & \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\hat{\mu} = 0 \\
\Leftrightarrow\ & \sum_{i=1}^{n}x_i - n \cdot \hat{\mu} = 0 \\
\Leftrightarrow\ & \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i \quad \text{(mean value of dataset } S\text{)}
\end{aligned}$$

  - Second, find $\hat{\sigma}$:

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}}\ \ell(\theta)$$

Also solve using closed form:

$$\nabla_{\sigma} \quad l(\theta) = 0$$

$$\Leftrightarrow \quad \nabla_{\sigma}\left[-n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] = 0$$

$$\Leftrightarrow \quad -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\cdot\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$ (mean variance of dataset S)

- Conclusion :

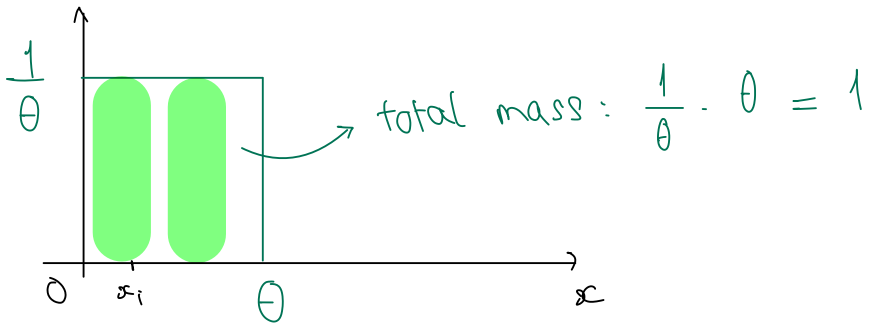
MLE gives
$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \end{cases}$$

Example: Uniform Distribution

- Given $\{x_i\}_{i=1}^n$ i.i.d. drawn from a uniform distribution $\text{Uniform}([0, \theta])$

$$p(x \mid \theta) = \begin{cases} \frac{1}{\theta} & \text{if } x \in [0, \theta] \\ 0 & \text{otherwise} \end{cases}$$

<u>Visualize:</u>



- <u>Problem:</u> Estimate θ

<u>Solution:</u>

- Since some probability distribution is 0, using likelihood function is easier in this case (because likelihood function use products) :

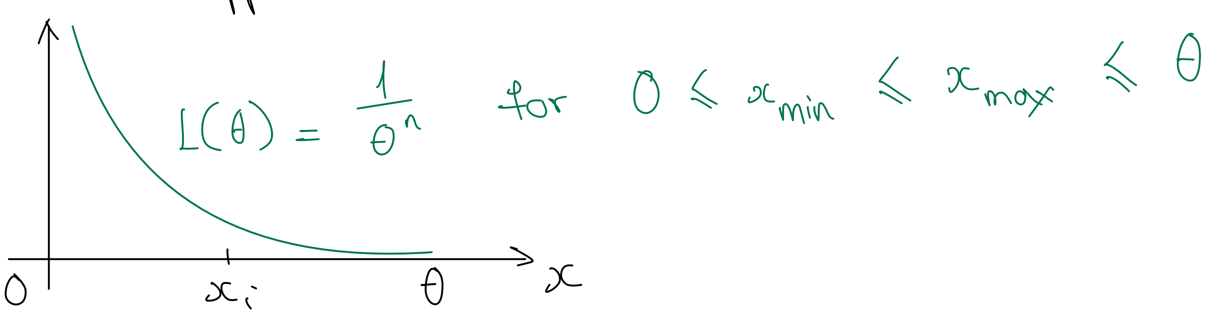
$$\begin{aligned} L(\theta) &= \prod_{i=1}^n p(x_i \mid \theta) \\ &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}(x_i \in [0, \theta]) \\ &= \left(\frac{1}{\theta}\right)^n \prod_{i=1}^n \mathbb{1}(x_i \in [0, \theta]) \\ &= \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } x_i \in [0, \theta] \ \forall i \\ 0 & \text{otherwise} \end{cases} \\ \text{or } &= \frac{1}{\theta^n} \quad \text{for } 0 \leqslant x_{min} \leqslant x_{max} \leqslant \theta \end{aligned}$$

- Try to solve using closed form:

$$\begin{aligned} & \nabla_\theta L(\theta) = 0 \\ \Leftrightarrow\ & \nabla_\theta \left(\frac{1}{\theta}\right)^n = 0 \\ \Leftrightarrow\ & -n \left(\frac{1}{\theta}\right)^{n+1} = 0 \end{aligned}$$

always $\neq 0$

- Try another approach, lets visualize $L(\theta)$:



$L(\theta)$ is a decreasing function of θ, so it doesn't have a max value.

But we can still find the "allowed" maximum, the "allowed" maximum occurs at the minimum θ (as showed in graph):

$$\max L(\theta) = \min \theta$$

And since the constraint on θ is:

$$\theta \geq x_{max}$$

Therefore:

$$\max L(\theta) = \min \theta = x_{max}$$

meaning: $\hat{\theta} = x_{max}$

or : $\hat{\theta} = \{x_1, \dots x_n\}$

- Conclusion:

The best estimate of θ is one that:

$$\begin{cases} \circ \text{ covers all data point} : & \theta \geq x_{max} \\ \circ \text{ as tight as possible} : & \theta = x_{max} \end{cases}$$

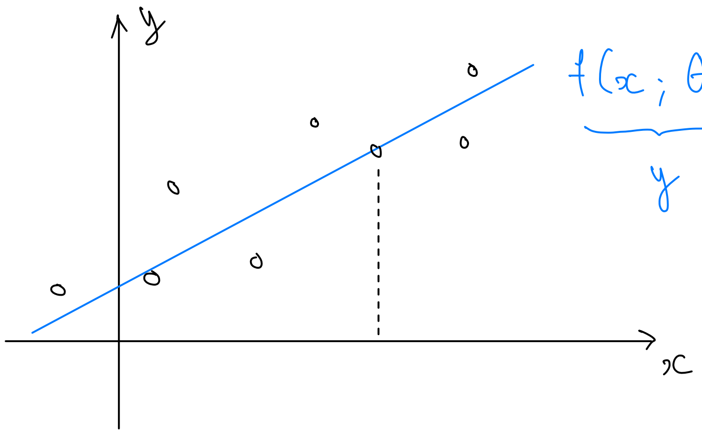
Example: MLE for (Linear) Regression

- Given $\{x_i, y_i\}_{i=1}^n$, where x_i, such that:

$$p(y \mid x, \theta) = \mathcal{N}(y \mid \underbrace{f(x;\theta)}_{\text{some line to fit the dataset}}, \underbrace{\sigma^2}_{\text{noise}}) \rightarrow \text{Why?}$$

We usually assume y_i follow Gaussian Distribution

Visualize:



$$\underbrace{f(x;\theta)}_{y} = \theta x + b$$

- Problem: Estimate θ and σ^2

Solution:

- Write down log-likelihood function:

$$\ell(\theta, \sigma) = \sum_{i=1}^n \log P(y_i \mid x_i, \theta)$$

$$= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left[y_i - f(x_i, \theta)\right]^2$$

- Define maximum likelihood estimation (MLE):
  - For $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta, \sigma)$$

$$= \arg\max_{\theta} \left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \left[y_i - f(x_i, \theta)\right]^2\right)$$

$$= \arg\min_{\theta} \left(\sum_{i=1}^n \left[y_i - f(x_i, \theta)\right]^2\right)$$

$$= \arg\min_{\theta} \| y - f(x, \theta) \|$$

$\rightarrow$ Linear Least Square

  - For $\hat{\sigma}$:

$$\hat{\sigma} = \arg\max_{\sigma} \ell(\theta)$$

Solve for $\hat{\sigma}$: $\nabla_{\sigma} \ell(\theta, \sigma) = 0$

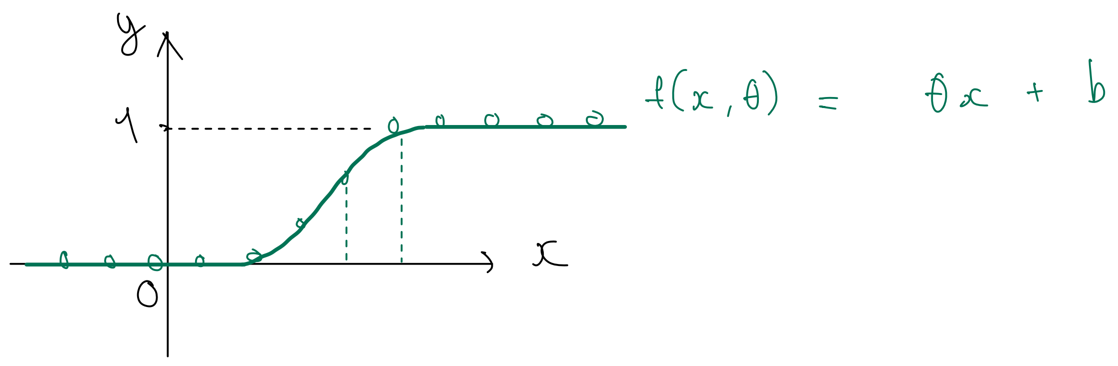
$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left[y_i - f(x_i, \hat{\theta})\right]^2 \rightarrow \text{mean variance of } y$$

Example: MLE for Logistic Regression

- Given $\{x_i, y_i\}_{i=1}^n$, where $y_i \in [0, 1]$ such that:

$$p(y \mid x, \theta) = \frac{\exp(y \cdot f(x, \theta))}{1 + \exp(f(x, \theta))}$$

Visualize:



- Problem: Estimate θ

Solution:

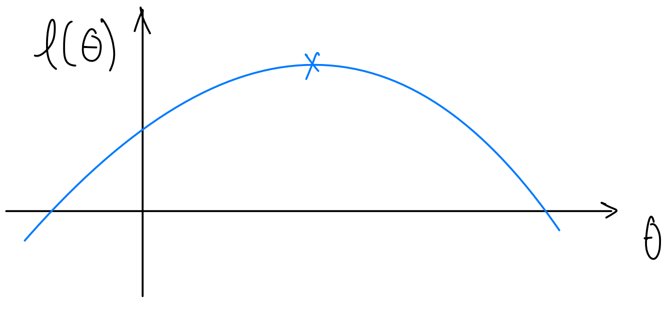
- Write down log-likelihood function:

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^n \log P(y_i \mid x_i, \theta) \\ &= \sum_{i=1}^n \log \left[\frac{\exp(y_i \, f(x_i, \theta))}{1 + \exp(f(x_i, \theta))} \right] \\ &= \sum_{i=1}^n \left[y_i \, f(x_i, \theta) - \log(1 + \exp(f(x_i, \theta))) \right]\end{aligned}$$

- Define maximum liklihood estimation:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \; \ell(\theta)$$

Since $\ell(\theta)$ has a term with log, this function curves exponentially, so you won't find a nice solution using closed form method. Instead, we converge slowly to the optimal solution using gradient "ascent", this is called numerical method.



Gradient update rule:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta \ell(\theta)$$