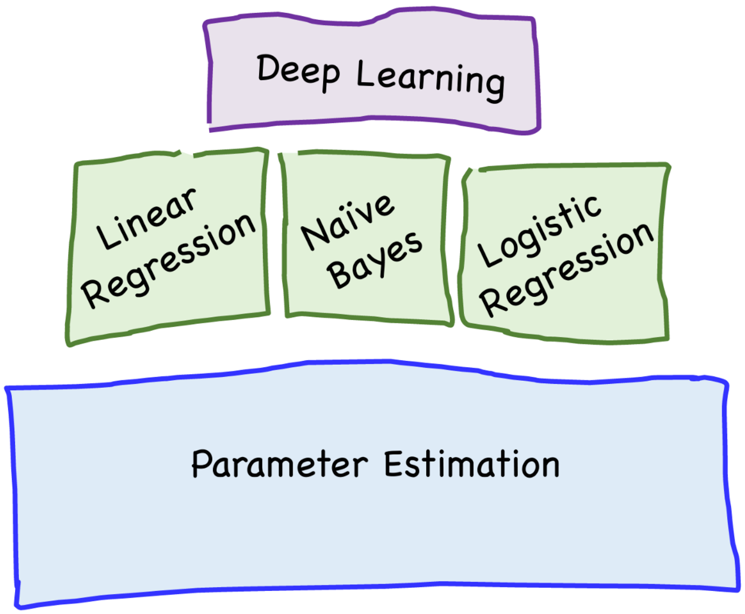
Computer Science

CSC380: Principles of Data Science

Statistics 2

Credit:
- Jason Pacheco,
- Kwang-Sung Jun,
- Chicheng Zhang
- Xinchen yu

Our path



- We don't know the true parameter.
- But we have observations.
- We assume each i.i.d observation follows a probability distribution with unknown parameters, and we build model.
  - e.g., Naive bayes model (X~Bernoulli)
- Compute estimator to estimate true parameter
- Many types of estimators with different properties
  - consistency
  - efficiency (mean squared error)
  - unbiasedness

Review

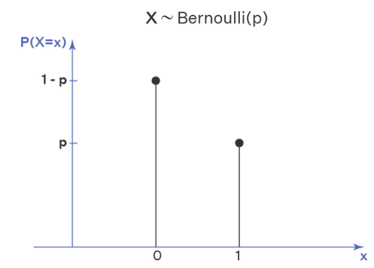
Law of Large Numbers:

$$\lim_{N\to\infty} \hat{\mu}_n = \mu$$

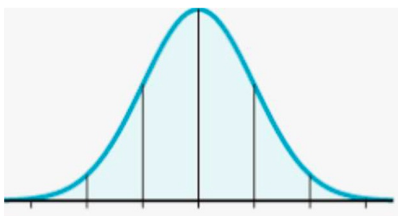
[1, 0, 1, 0, 0, …, 1, 1, 0]

(1+0+1+0+0+...+1+1+0)/N

Draw from a distribution with unknown mean



Central Limit Theorem:



$$\lim_{N\to\infty} \bar{X}_N \to \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\lim_{N\to\infty} \frac{\sqrt{N}}{\sigma}(\bar{X}_N - \mu) \to \mathcal{N}(0,1)$$

[1, 0, 1, 0, 0, …, 1, 0, 1] $\bar{X}_N$ for sample 1

[1, 0, 0, 0, 0, …, 1, 1, 0] $\bar{X}_N$ for sample 2

[1, 1, 1, 0, 1, …, 0, 1, 0] $\bar{X}_N$ for sample 3

……

[0, 0, 1, 1, 1, …, 0, 0, 0] $\bar{X}_N$ for sample k

If N is very large, and we draw the distribution of $\bar{X}_N$ from all the samples, it follows normal distribution.

Intuition Check

Suppose that we toss a coin 100 times. We observe 73 heads and 27 tails…

<u>Question</u> Let θ be the coin bias (probability of heads). What is a more likely estimate? What is your reasoning?

A: $\hat{\theta} = 0.73$, strong preference for heads

Why sample mean?

B: $\hat{\theta} = 0.50$, fair coin (we observed unlucky outcomes)

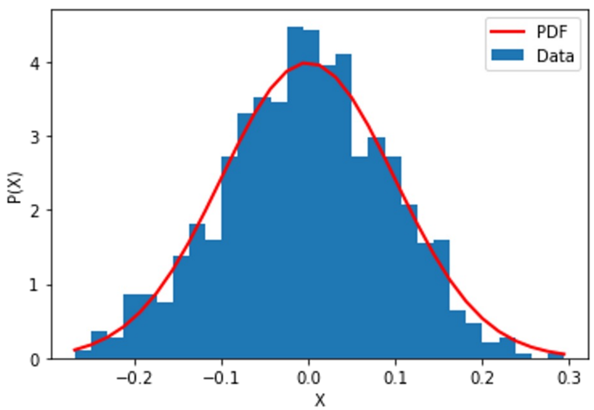
Likelihood (informally) Probability/density of the observed outcomes from a particular model.

Likelihood (Intuitively)

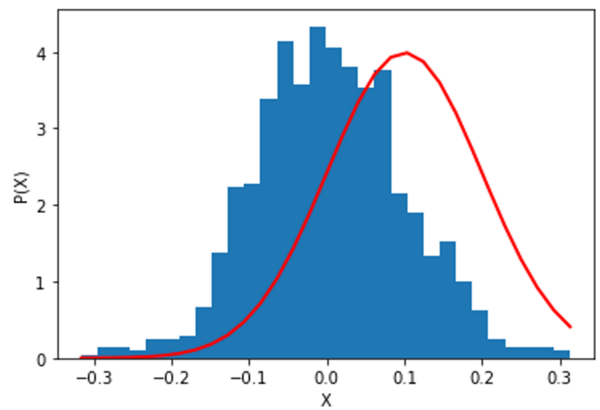
Suppose we observe N data points from a Gaussian model $\mathcal{N}(\mu,\sigma^2)$, and wish to estimate both μ and σ^2.

Say we only need to choose from the following three Gaussians…

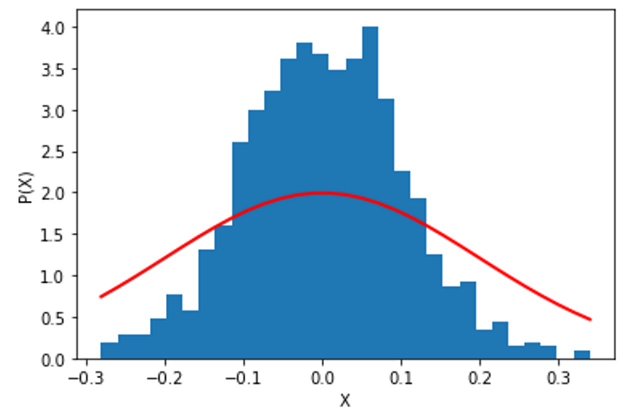
High Likelihood



Low Likelihood (mean)



Low Likelihood (variance)



Likelihood Principle: *Given a statistical model, the <u>likelihood function</u> describes all evidence of a parameter that is contained in the data.*

Likelihood Function

Suppose $x_i \sim p(x;\theta)$, then what is the **joint probability** over N *independent identically distributed* (iid) observations $x_1, \ldots, x_N$?

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

what appears after ; are parameters, not random variables.

- We call this the **likelihood function**, often denoted $\mathcal{L}_N(\theta)$
- It is a function of the parameter θ, the data are fixed
- Describes how well parameter θ describes data (*goodness of fit*)

How could we use this to estimate a parameter θ*?*

Likelihood Function

Suppose $x_i \sim p(x;\theta)$, then what is the **joint probability** over N *independent identically distributed* (iid) observations $x_1, \ldots, x_N$?

$$p(x_1, \ldots, x_N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

what appears after ; are parameters, not random variables.

Suppose X ~ Bernoulli(p), we have 5 observations [1, 1, 0, 1, 0]

- If true parameter is 0.6: fit the data better

$$p(1,\ 1,\ 0,\ 1,\ 0;\ .6) = p(1; .6) \cdot p(1; .6) \cdot p(0; .6) \cdot p(1; .6) \cdot p(0; .6) = 0.6^3 0.4^2$$

- If true parameter is 0.2:

$$p(1,\ 1,\ 0,\ 1,\ 0;\ .2) = p(1; .2) \cdot p(1; .2) \cdot p(0; .2) \cdot p(1; .2) \cdot p(0; .2) = 0.2^3 0.8^2$$

Maximum Likelihood

Maximum Likelihood Estimator (MLE) as the name suggests, maximizes the likelihood function.

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}_N(\theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

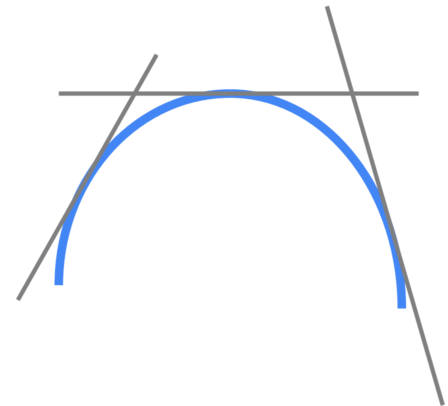
Question How do we find the MLE?

1. closed-form
2. iterative methods

How to find the maximum/maximizer of a function?

Option 1: finding the maximum/maximizer

Example: Suppose $f(\theta) = -a\theta^2 + b\theta + c$ with $a > 0$



It is a quadratic function.
=> finding the 'flat' point suffices

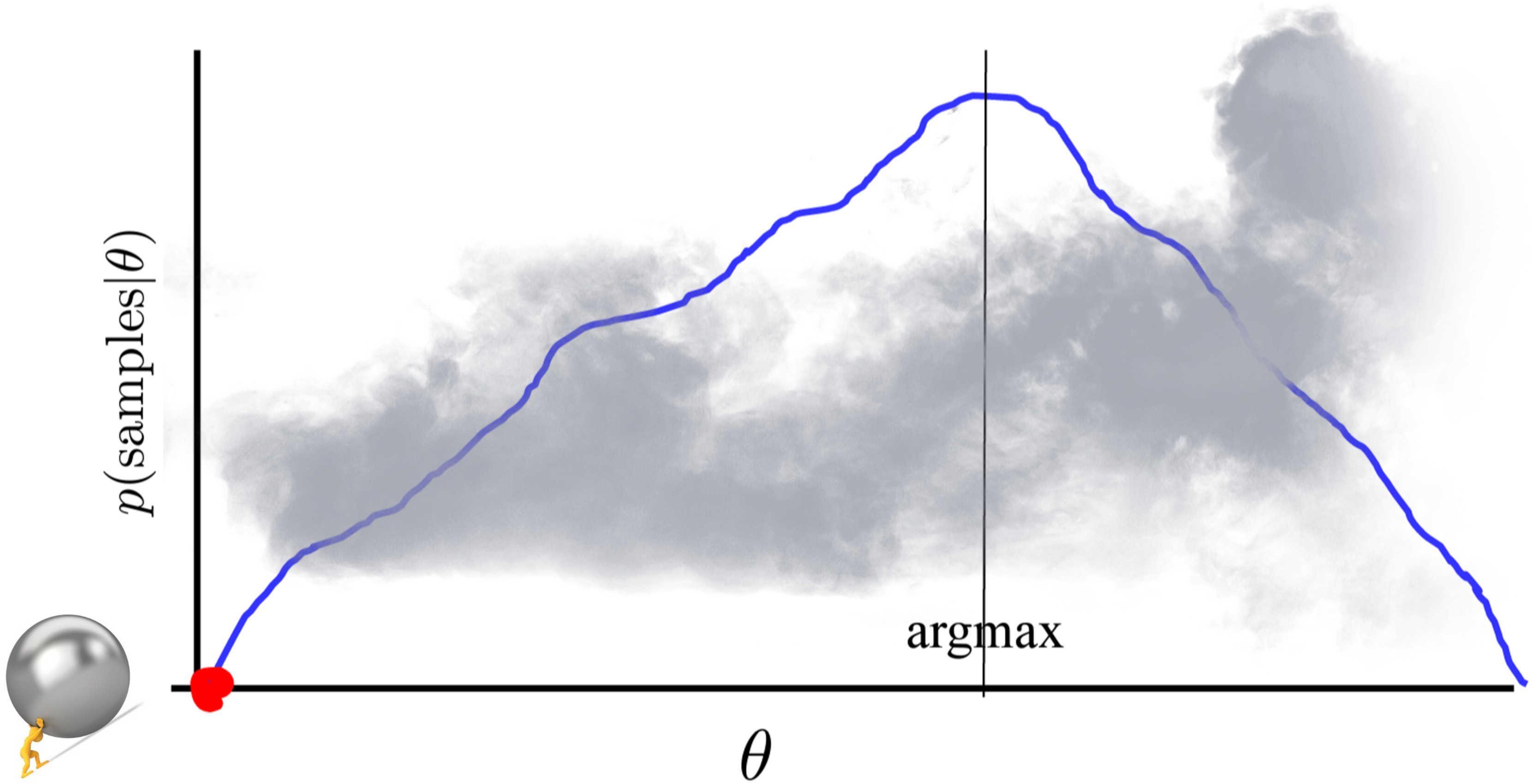
Compute the gradient and set it equal to 0

$$f'(\theta) = -2a\theta + b \qquad \Rightarrow \qquad \theta = \frac{b}{2a}$$

Closed form!

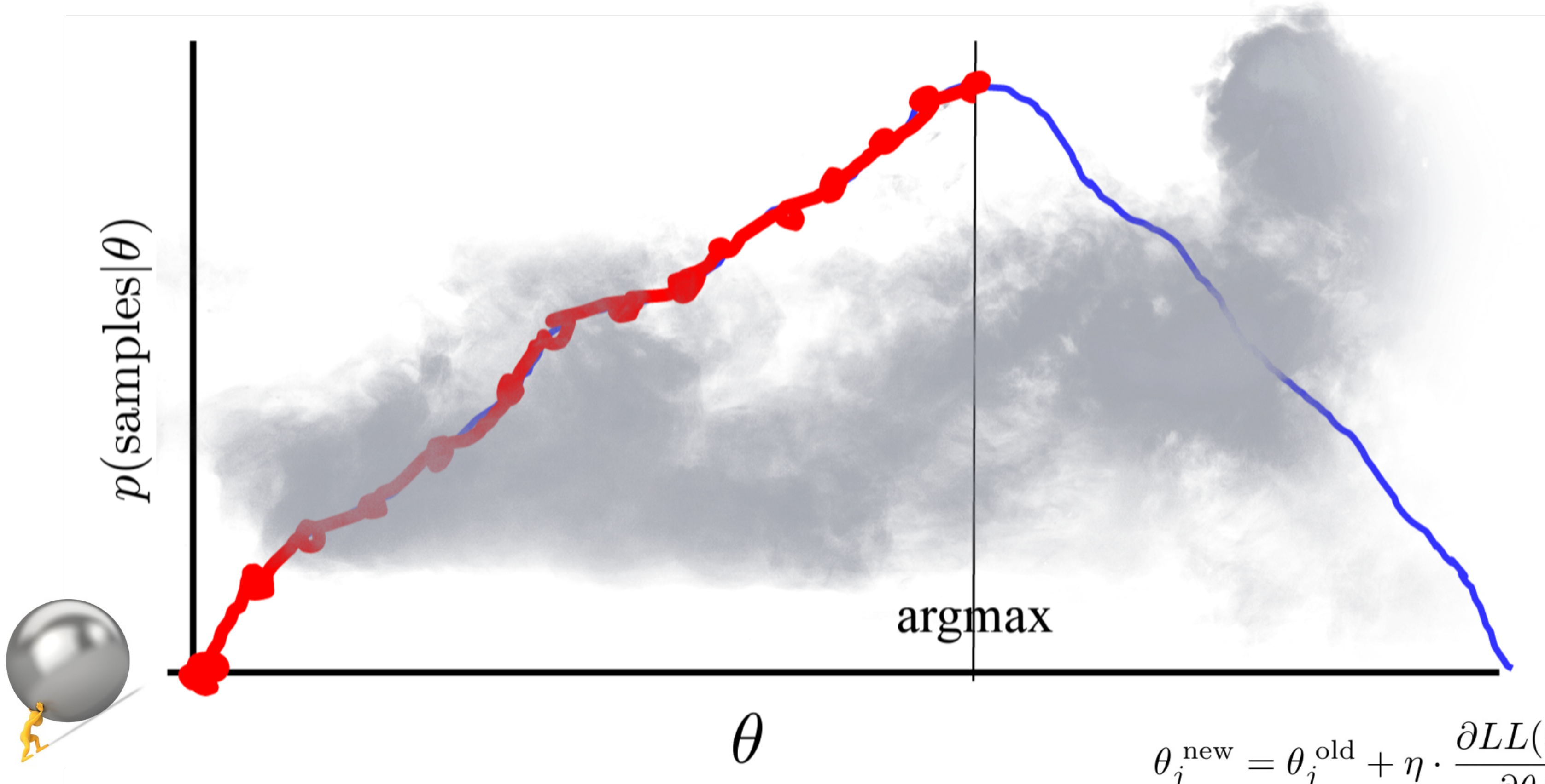
Q: Does this trick of grad=0 work for other functions?
⇒ Yes, **concave** functions!
⇒ Roughly speaking, functions that curves down only, never upwards

Option 2: finding the maximum/maximizer



Walk uphill and you will find a local maxima (if your step size is small enough)

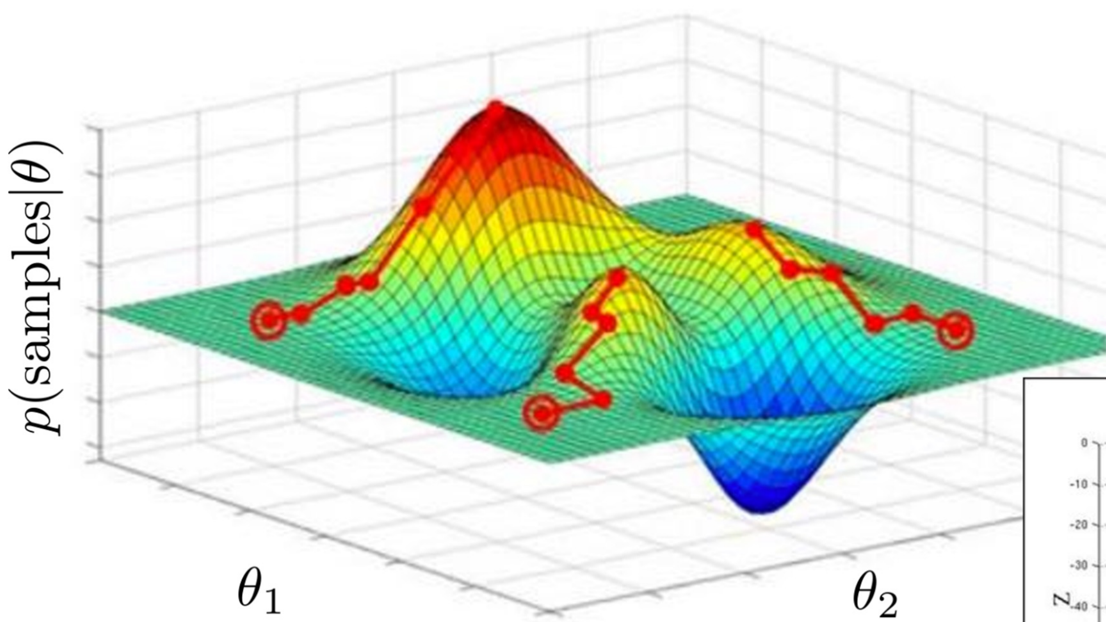
Option 2: finding the maximum/maximizer



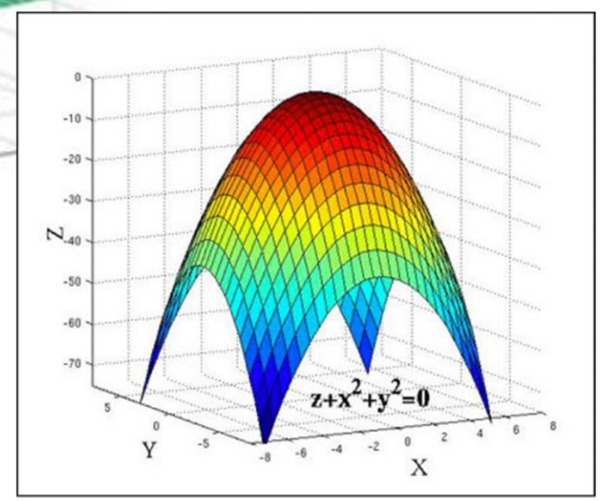
$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{ old}})}{\partial \theta_j^{\text{old}}}$$

Walk uphill and you will find a local maxima (if your step size is small enough)

Option 2: finding the maximum/maximizer



Especially good for concave function



$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{ old}})}{\partial \theta_j^{\text{old}}}$$

Walk uphill and you will find a local maxima (if your step size is small enough)

Option 2: finding the maximum/maximizer

What if there is no closed form solution?

Example: $f(\theta) = \frac{1}{2}x(ax - 2\log(x) + 2)$

$$f'(\theta) = ax - \log(x)$$

No known closed form for $ax = \log(x)$

Iterative methods
- for **concave** functions
  => Will find the global maximum
- for **nonconcave**,
  => usually find a local maximum but could also get stuck at *stationary point*.

Iterative methods:
- Gradient ascent (or *descent* if you are minimizing):
- Newton's method
- Etc. (beyond the scope of our class)

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$

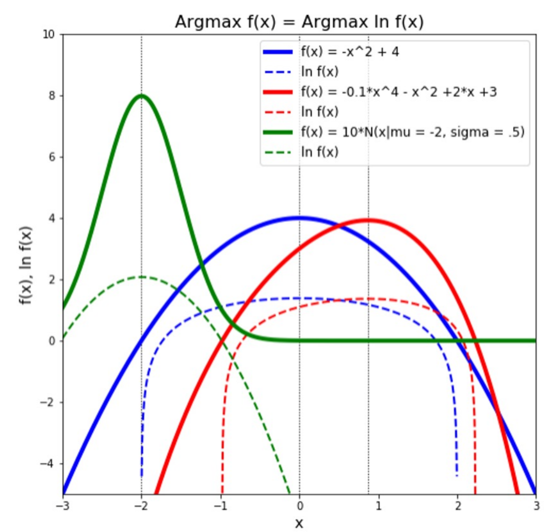
Maximum Likelihood

Maximizing **log**-likelihood makes the math easier (as we will see) and doesn't change the answer (logarithm is an increasing function)

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta} \; \log \mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log p(x_i; \theta)$$

Derivative is a linear operator so,

$$\frac{d}{d\theta} \log \mathcal{L}_N(\theta) = \underbrace{\sum_{i=1}^{N} \frac{d}{d\theta} \log p(x_i; \theta)}_{\substack{\text{One term per data point} \\ \text{Can be computed in parallel} \\ \text{(big data)}}}$$



Review: maximum likelihood estimation

1. Decide on a model for the likelihood of your samples. This is often using a PMF or PDF.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4.Calculate the derivative of LL with respect to theta

5.Use an optimization algorithm to calculate argmax

Maximum Likelihood: Bernoulli

Example: Consider I.I.D. random variables: $X_1, X_2, X_3 \ldots X_n \sim Bernoulli\ (p)$
We don't know the coin bias p.

Probability Mass function: $p^{x_i}(1-p)^{1-x_i}$

Likelihood: $\mathfrak{L}_n(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{x_1+\ldots+x_n}(1-p)^{n-(x_1+\ldots+x_n)}$

$= p^S(1-p)^{n-S}$

$\boxed{S = \sum_i x_i}$

Log likelihood: $\mathfrak{L}\mathfrak{L}_n(p) = S \log p + (n-S)\log(1-p)$

Maximum Likelihood: Bernoulli

[Source: Wasserman, L. 2004]

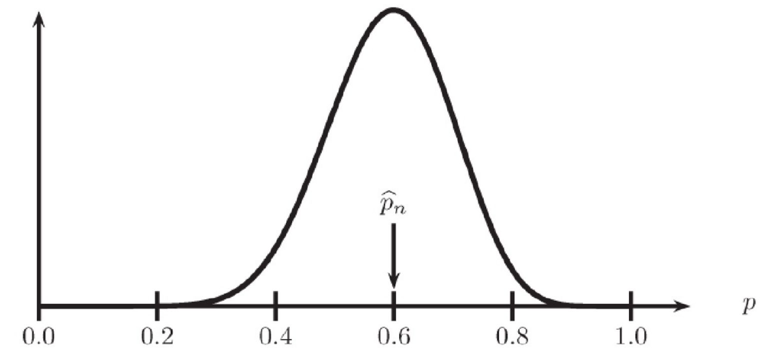
Set the derivative of $\mathcal{LL}_n(p)$ to zero and solve,

$$\mathcal{LL}_n(p) = S \log p + (n - S)\log(1 - p)$$

$$\frac{\partial \mathcal{LL}_n(p)}{\partial p} = S\frac{1}{p} + (n - S)\frac{-1}{1 - p} = 0$$

We get:

$$p_{MLE} = \frac{S}{n} = \frac{1}{n}\sum_i x_i \quad \boxed{S = \sum_i x_i}$$



Likelihood function for Bernoulli with n=20 and $\sum_i x_i = 12$ heads

Isn't that the same as the sample mean?

Yes, for Bernoulli

⇒ this showcases how MLE is aligned to our intuition!

Maximum Likelihood: Gaussian

Example Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with parameters $\theta = (\mu, \sigma^2)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned}
\mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(X_i - \mu)^2\right\} \\
&= \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2\right\} \\
&= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right\}
\end{aligned}$$

Exercise: Show that

$$\sum_i (X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$$

$$\boxed{e^{x+y} = e^x e^y}$$

Where $\bar{X} = \frac{1}{n}\sum_i X_i$ and $S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ are sample mean and sample variance, respectively.

Maximum Likelihood: Gaussian

Given:

$$\bar{X} = \frac{1}{n}\sum_i X_i$$

$$S^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$$

$$\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2 = \sum_i \left[(X_i - \overline{X})^2 + 2(X_i - \overline{X})(\overline{X} - \mu) + (\overline{X} - \mu)^2\right]$$

$$= \sum_i \left[(X_i - \overline{X})^2 + 2\left(X_i\overline{X} - X_i\mu - \overline{X}^2 + \overline{X}\mu\right) + \left(\overline{X}^2 - 2\overline{X}\mu + \mu^2\right)\right]$$

$$= \sum_i \left[(X_i - \overline{X})^2 + 2X_i\overline{X} - 2X_i\mu - 2\overline{X}^2 + 2\overline{X}\mu + \overline{X}^2 - 2\overline{X}\mu + \mu^2\right]$$

$$= \sum_i \left[(X_i - \overline{X})^2 + 2X_i(\overline{X} - \mu) - \overline{X}^2 + \mu^2\right]$$

$$= \sum_i (X_i - \overline{X})^2 + \sum_i 2X_i(\overline{X} - \mu) - n\overline{X}^2 + n\mu^2)$$

$$= \sum_i (X_i - \overline{X})^2 + 2n\overline{X}(\overline{X} - \mu) - n\overline{X}^2 + n\mu^2$$

$$= \sum_i (X_i - \overline{X})^2 + n\left(\overline{X}^2 - 2\overline{X}\mu + \mu^2\right) = \sum_i (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

Maximum Likelihood: Gaussian

Continuing, write log-likelihood as:

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}.$$

Solve zero-gradient conditions:

$$\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0,$$

To obtain maximum likelihood estimates of mean / variance:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i} (X_i - \hat{\mu})^2$$

MLE Review: Probability/Density vs Likelihood

- The **probability/density** of data given parameter is mathematically the same object as **likelihood** of a parameter given data

- The difference is the point of view!
  - From the *probabilistic perspective*, the parameter is fixed and **PMF/PDF** is viewed as a function of the possible data

  - From the *statistical perspective*, the data is given (thus fixed) and we view **likelihood** as a function of the parameter.

- Statistics is inherently about reverse engineering.