



Computer
Science

CSC380: Principles of Data Science

Statistics 2

Xinchen Yu

Announcement

- HW04 has been out on D2L
 - You can work in pairs
 - due next Wednesday by 11:59 pm
 - Points will be deducted if not including code in the PDF
- Disclaimer: all derivatives in the HWs or exams are the most basic ones
 - $\frac{d}{dx}(x) = 1, \frac{d}{dx}(\ln(x)) = \frac{1}{x}, \frac{d}{dx}(c) = 0, (cf(x))' = c(f'(x))$
 $(f(x) + g(x))' = f'(x) + g'(x), \frac{d}{dx}(f(g(x))) = f'(g(x))g'(x)$

Calculus Review

Question: $\frac{d}{dx} \ln(1-x)$?

$$g(x) = 1 - x, \quad f(z) = \ln(z), \quad f(g(x)) = \ln(1 - x)$$

$$\frac{d}{dx}(1 - x) = \frac{d}{dx}(1) + \frac{d}{dx}(-1 \cdot x) = 0 + (-1) = -1$$

$$\frac{d}{dx} \ln(1 - x) = \frac{1}{1 - x} \cdot (-1) = \frac{1}{x - 1}$$

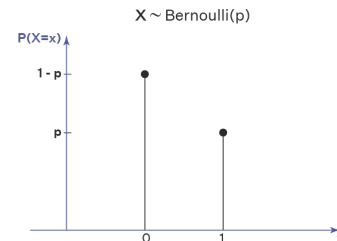
Law of Large Numbers:

[1, 0, 1, 0, 0, ..., 1, 1, 0]

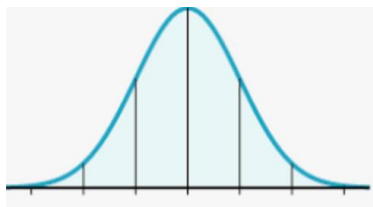
$(1+0+1+0+0+\dots+1+1+0)/N$

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

Draw from a
distribution with
unknown mean



Central Limit Theorem:



[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

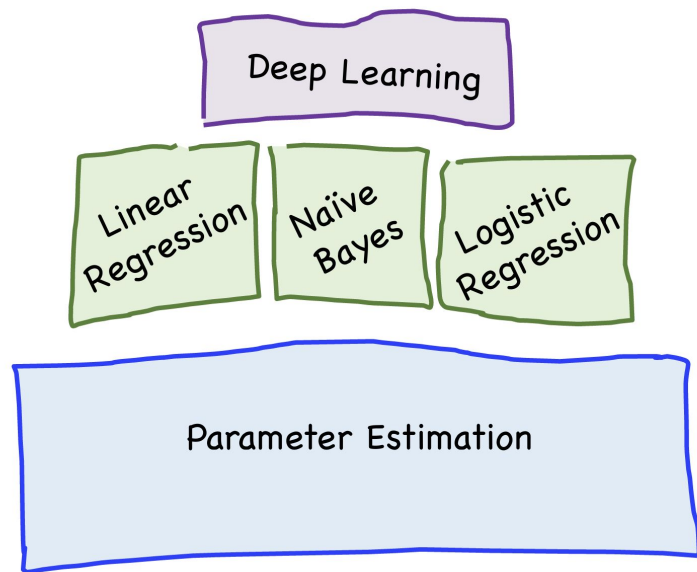
.....

[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.

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

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



- 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 ($X \sim \text{Bernoulli}$)
- Compute estimator to estimate true parameter
- Many types of estimators with different properties
 - consistency
 - efficiency (mean squared error)
 - unbiasedness

*All models are wrong
but some are useful*



George E.P. Box

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

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

What model to choose?

Why Bernoulli(0.73) is better than Bernoulli(0.50)?

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

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

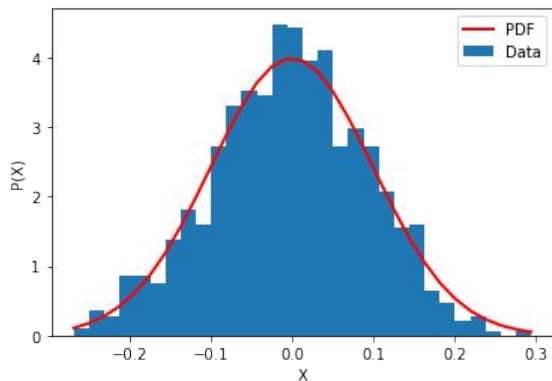
Likelihood (informally) Probability/density of the observed outcomes from a particular model.



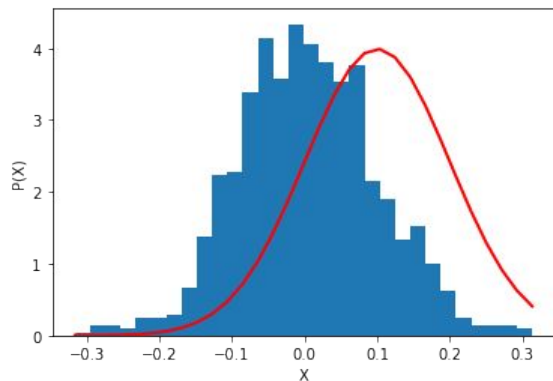
Suppose we observe N data points from a Gaussian model $\mathcal{N}(\mu, \sigma^2)$, and wish to estimate both μ and σ^2 .

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

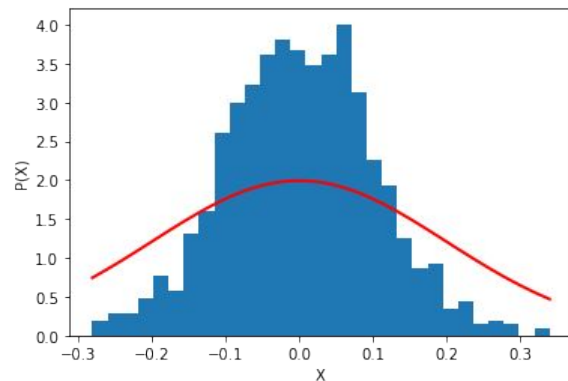
**High
Likelihood**



**Low
Likelihood (mean)**



**Low
Likelihood (variance)**



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

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

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

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 ?

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

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

what appears after ; are parameters, not random variables.

Suppose $X \sim \text{Bernoulli}(p)$, 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 Estimator (MLE) as the name suggests, maximizes the likelihood function.

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

Question How do we find the MLE?

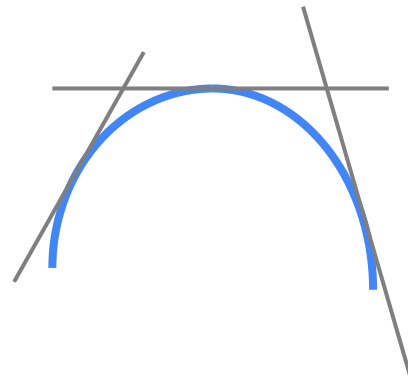
1. closed-form
2. iterative methods

How to find the maximum/maximizer of a function?

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

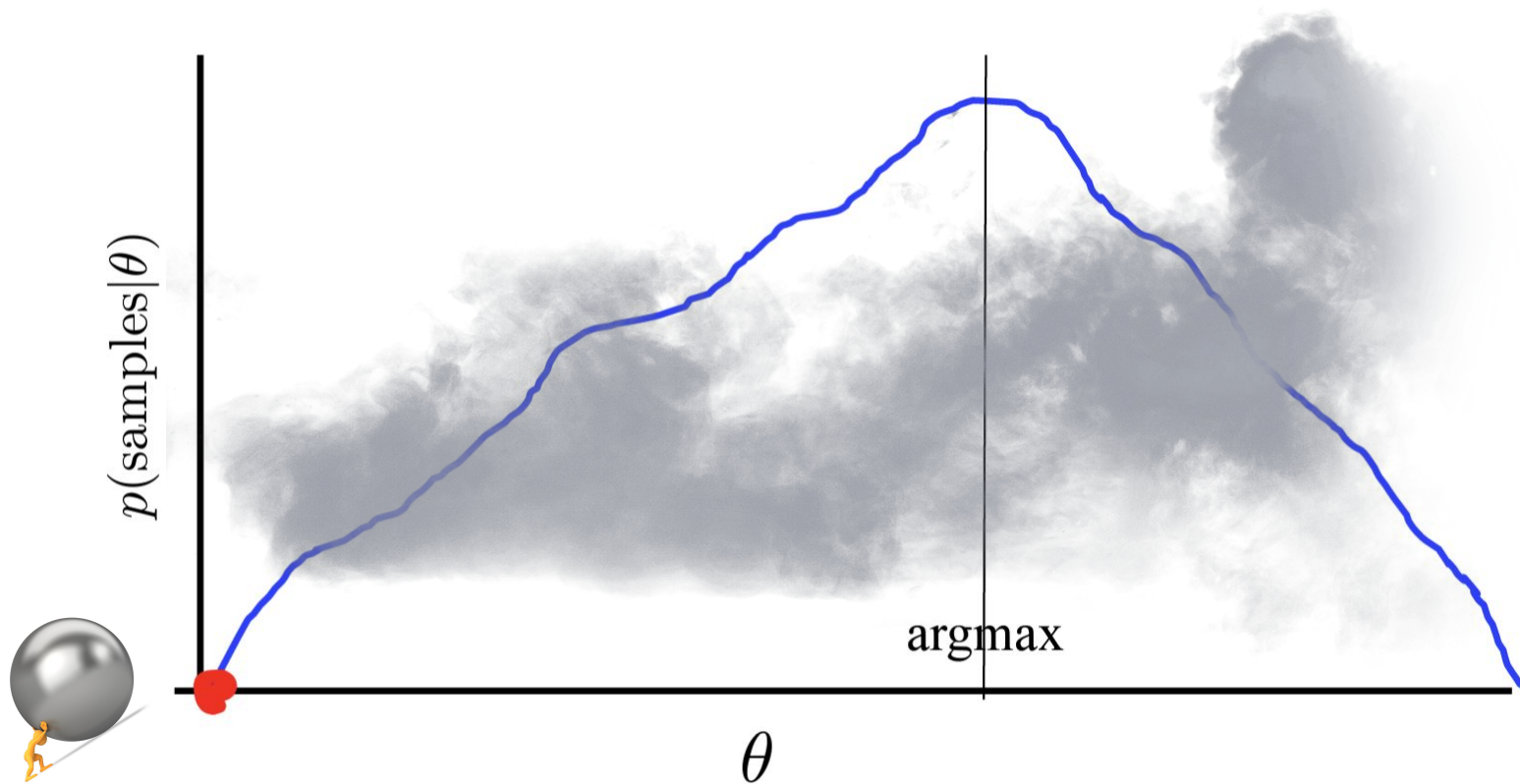
$$f'(\theta) = -2a\theta + b \quad \Rightarrow \quad \theta = \frac{b}{2a}$$

Closed form!

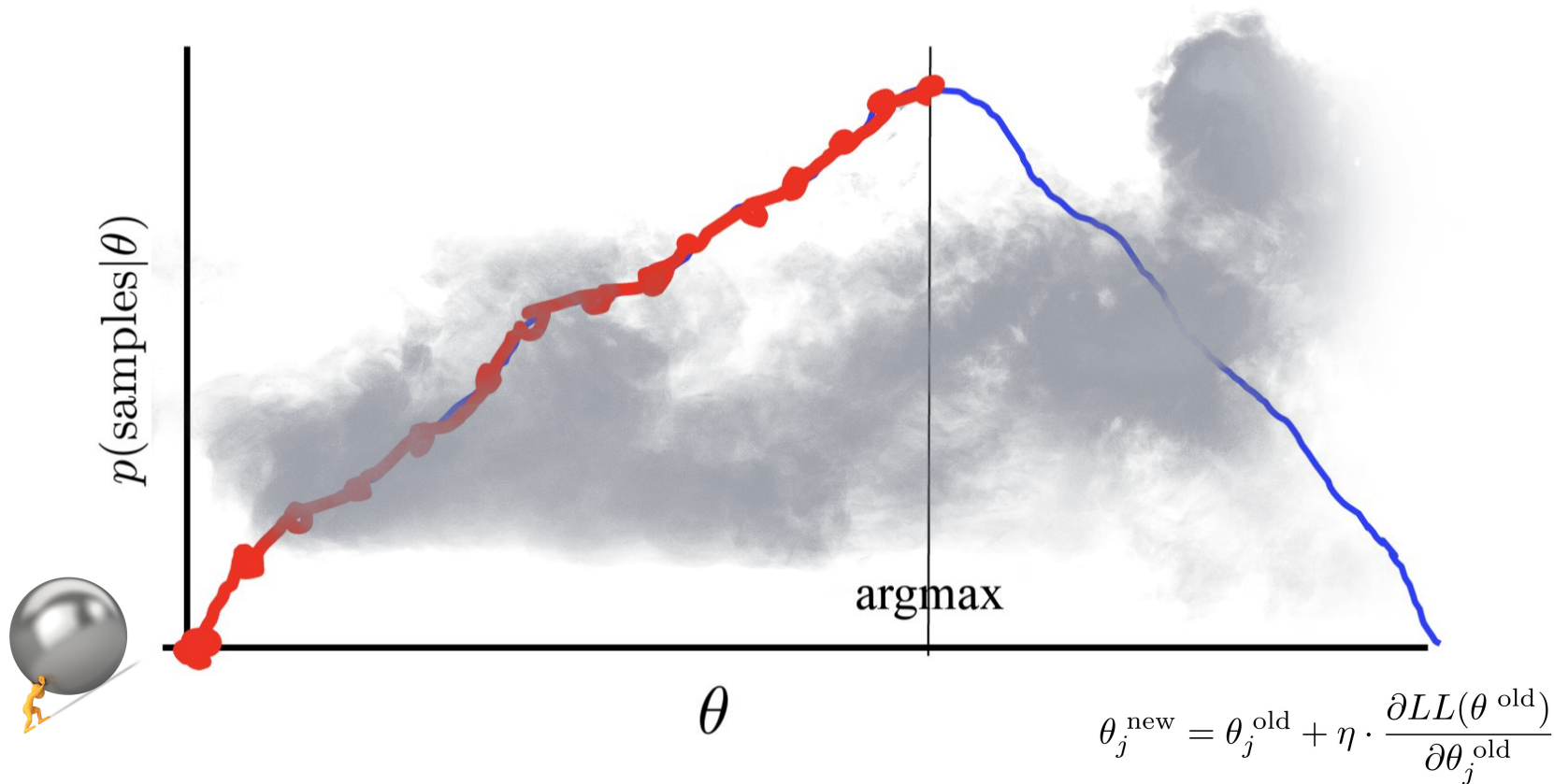
Q: Does this trick of $\text{grad}=0$ work for other functions?

⇒ Yes, **concave** functions!

⇒ Roughly speaking, functions that curves down only, never upwards



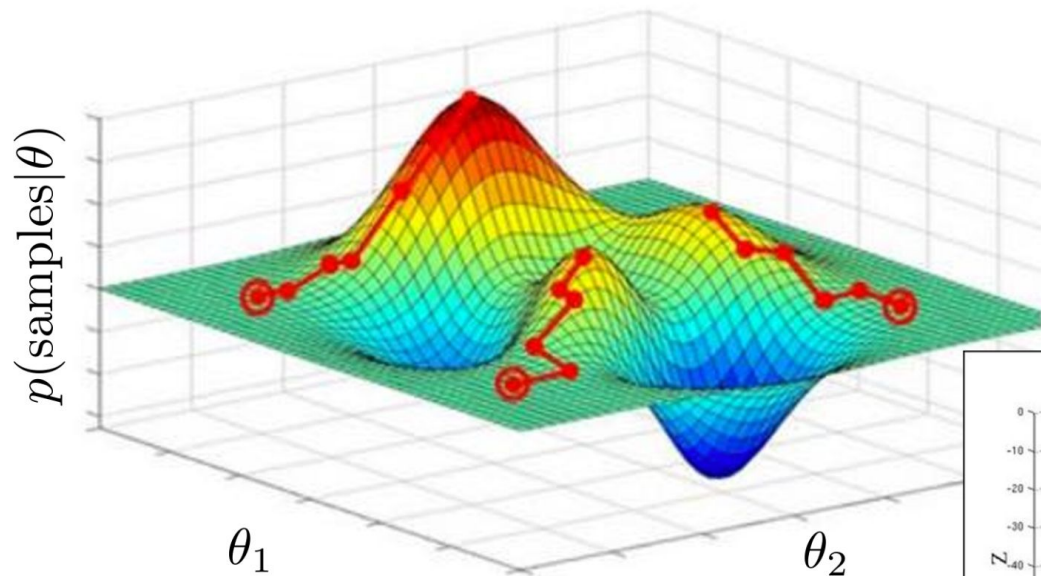
Walk uphill and you will find a local maxima (if your step size is small enough)



The gradient of a function gives you the direction that increase the function most.

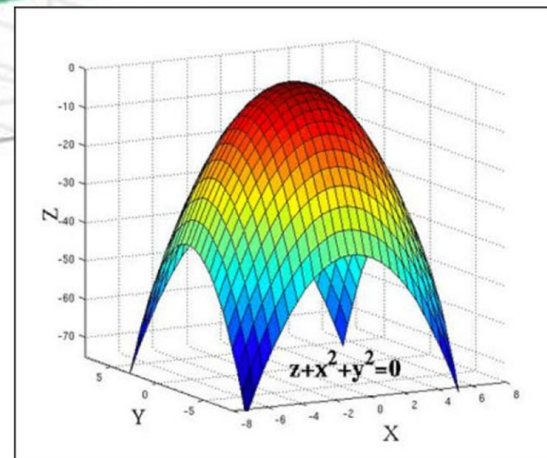
Option 2: finding the maximum/maximizer

16



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)

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*.

We only work with functions having closed-form solutions!

Iterative methods:

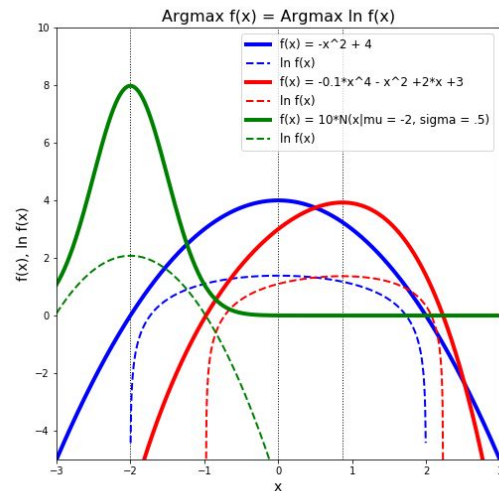
- Gradient ascent (or *descent* if you are minimizing): $\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$
- Newton's method
- Etc. (beyond the scope of our class)

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,

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



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 θ

5. Use an optimization algorithm to calculate argmax

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

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

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

$$\begin{aligned} \text{Likelihood: } \mathcal{L}_n(p) &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{x_1+\dots+x_n}(1-p)^{n-(x_1+\dots+x_n)} \\ &= p^S(1-p)^{n-S} \end{aligned}$$

$$\text{where: } S = \sum_i x_i$$

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

[Source: Wasserman, L. 2004]

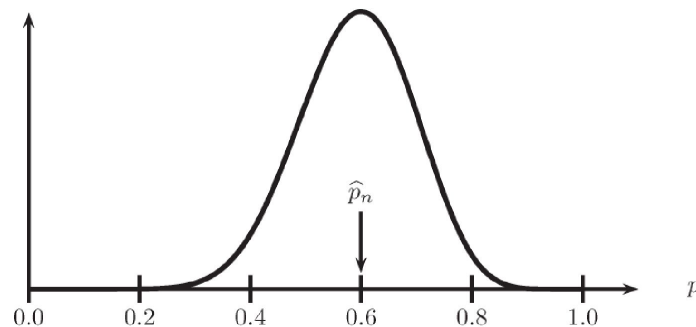
Set the derivative of $\mathcal{L}\mathcal{L}_n(p)$ to zero and solve,

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

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

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!

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

$$\mathcal{L}_n(\mu, \sigma) = \prod_i \frac{1}{\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \mu)^2 \right\}$$

$$= \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\}$$

Exercise: Show that

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

$$= \sigma^{-n} \exp \left\{ -\frac{nS^2}{2\sigma^2} \right\} \exp \left\{ -\frac{n(\bar{X} - \mu)^2}{2\sigma^2} \right\}$$

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

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 (not required)

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

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

Given:

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

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

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

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

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

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

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

Continuing, write log-likelihood as:

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

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:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

- 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.