Parameter Estimation

Data Intelligence and Learning (DIAL) Lab
Prof. Jongwuk Lee

Bayes' Theorem

Motivation: Partial Information

- We have assumed we **know nothing about the outcome of our experiment.**

- Sometimes, we have **partial information** that may affect the **likelihood** of a given event.
  - Experiment: you **roll a die**.
  - Partial information: you are told that **the number is odd**.

  - Experiment: we predict the **weather tomorrow**.
  - Partial information: we know that the **weather today** is **rainy**.

Incorporating Partial Information

- Knowing about event B (e.g., "it is raining today") changes our beliefs about event A (e.g., "will it rain tomorrow?").

- How to update our probability law to incorporate **this new knowledge**?

- Introduce a **conditional probability**.

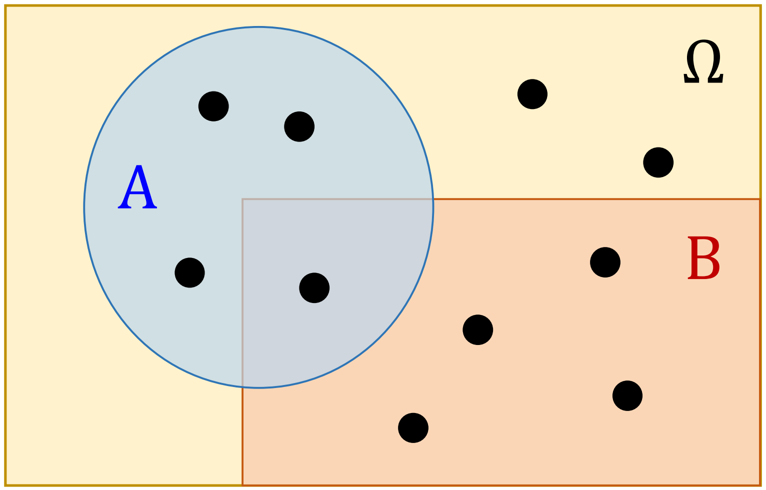
What is Conditional Probability?

- **Original problem**
  - What is the probability of some event A?
    - What is the probability that we roll a number less than 4?
  - This is given by our probability law.

- **New problem**
  - **Given event $\boldsymbol{B}$**, what is the probability of event A?
    - Given that the number rolled is an odd number, what is the probability that it is less than 4?
  - We call this the **conditional distribution of $\boldsymbol{A}$ given $\boldsymbol{B}$**.
  - We write this as $\boldsymbol{P(A \mid B)}$.
    - Read | as **given** or **conditioned on the fact that**.
  - Our **conditional probability** is still describing "**the probability of something**", so we expect it to behave like a **probability distribution**.

Idea of Conditioning

- $P(A \mid B)$ = "Probability of A, given that B occurred"



Usually, Ω is ignored.

$$P(A \mid \Omega) = \frac{P(A \cap \Omega)}{P(\Omega)}$$

$$\Downarrow$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

defined only if $P(B) > 0$

Bayes' Theorem

- Let $A_1, A_2, \ldots, A_n$ be a partition of the sample space.
- Let B be any set. Then, for each $i = 1, 2, \ldots, n$

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

- It is useful for **inferring hidden causes** from **our observation**.

Thomas Bayes (1701-1761).
English statistician, philosopher and Presbyterian minister

Typical Bayes Rule Example

- Considering testing for some latent (hidden/unobservable) disease, it will not be symptomatic until a future time point.
- We can directly observe the outcome of the test.
- Assuming the test is not 100% accurate, we cannot directly observe whether we have the disease.

- Two possible hidden causes for a positive test result.
  - We have the disease, and the test is correct.
  - We don’t have the disease, and the test is a false positive.

- Inferring which hidden cause underlies our observation

Example: Disease Testing

- **Assume that the disease affects 2% of the population.**
  - The false positive rate is **1%**.
  - The false negative rate is **5%**.
  - We take the test, and the result is **positive**.

- **Given that you tested positive, what is the probability you have the disease?**

Example: Disease Testing

- **Assume that the disease affects 2% of the population.**
  - The false positive rate is **1%**.
  - The false negative rate is **5%**.
  - We take the test, and the result is **positive**.

- **Given that you tested positive, what is the probability you have the disease?**

- **Let T be the event "tests positive" and D be the event "has disease."**
  - $P(D) = 0.02,\ P(T \mid D^C) = 0.01,\ P(T^C \mid D) = 0.05$

Example: Disease Testing

- Given that you tested positive, what is the probability you have the disease?

- What is $P(D \mid T)$? Bayes' rule gives us:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^C)P(D^C)}$$

- We get from the **conditional probability of an observation given a hidden cause** (which we usually know) to the **conditional probability of a hidden cause given an observation** (which we usually care about!)

Example: Disease Testing

- What is $P(D \mid T)$? Bayes' rule gives us:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid D^C)P(D^C)}$$

- So, let's plug in the numbers. Recall
  - $P(D) = 0.02,\ P(T \mid D^C) = 0.01,\ P(T^C \mid D) = 0.05$
  - So, $P(T \mid D) = 0.95, P(D^C) = 0.98$

$$P(D \mid T) = \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.01 \times 0.98} = \frac{0.019}{0.0288} = 0.66$$

Bayes' Theorem in ML

- It is useful for **inferring hidden causes** from **our observation.**

Posterior probability **Likelihood** **Prior probability**

$$P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)} \propto P(X \mid \theta)P(\theta)$$

θ: parameter, X: data

- It is also commonly used for parameter estimation methods.
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori estimation (MAP)

Bayes' Theorem in ML

- **Notations**
  - Posterior is the probability of the parameters $\boldsymbol{\theta}$ given $\boldsymbol{X}$.
  - Prior encapsulates our subjective prior knowledge of the observed (latent) variable $\boldsymbol{\theta}$ before observing any data.
  - Likelihood is the function of $\boldsymbol{\theta}$ given fixed $\boldsymbol{X}$.

Likelihood **Prior**

$$\boldsymbol{P(\theta \mid X)} = \frac{\boldsymbol{P(X \mid \theta) P(\theta)}}{\boldsymbol{P(X)}} \propto \boldsymbol{P(X \mid \theta) P(\theta)}$$

Posterior **Evidence**

- **It is also commonly used for parameter estimation methods.**
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori estimation (MAP)

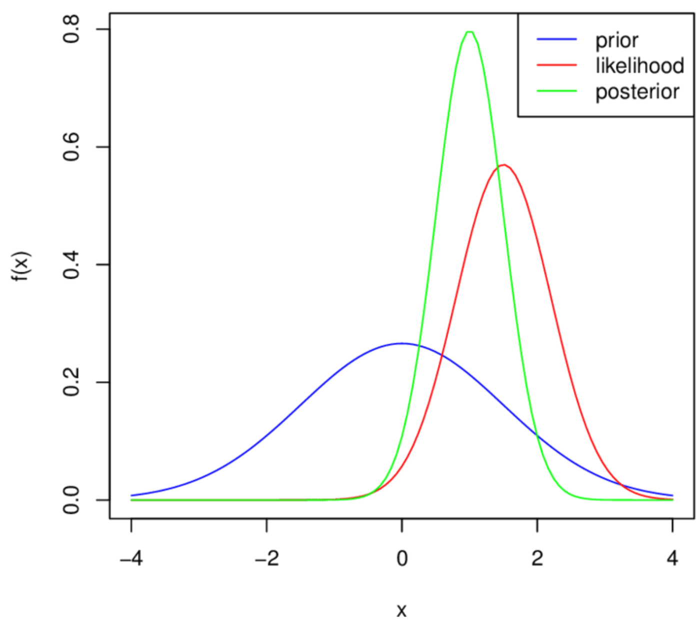
Bayes' Theorem in ML

➢ Intuition

- **Prior:** how plausible is the model a priori **before observing the data**?
  - The probability of being a head is 0.5.
- **Likelihood**: how well does **the model explain the data**?
  - For 4 out of 5 trials, the coin is head.
- **Posterior**: how plausible is the model **after observing the data**?
  - For this coin, the probability of a head is 0.7.

Posterior Likelihood Prior

$$P(\theta \mid X) \propto P(X \mid \theta) P(\theta)$$



Bayes' Theorem: Model Version

- **Let M be model and E be evidence.**

- $P(M|E)$ **proportional to** $P(E|M) \times P(M)$

$$P(M \mid E) \propto P(E \mid M) P(M)$$

Posterior **Likelihood** **Prior**

- **Intuition**
  - **Prior** = how plausible is the event (model, theory) a priori before seeing any evidence?
  - **Likelihood** = how well does the model explain the data?

Principles for Estimating Parameters

➢ Maximum likelihood estimation (MLE)

- Choose θ that maximizes the **likelihood** for **observed data** X

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, P(X \mid \theta)$$

➢ Maximum a posteriori (MAP)

- Choose θ given **prior of $\boldsymbol{\theta}$** and the **likelihood for observed data** X

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid X)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(X \mid \theta)\, P(\theta)$$

Maximum Likelihood Estimation (MLE)

Estimation in Statistics

- Use **sample statistics** to estimate **population parameters.**
  - E.g., Sample means are used to estimate population means.

- A **point estimate** of a population parameter is **a single value** of a statistic.

- An **interval estimate** is defined by two numbers between which a population parameter is said to lie.
  - $a < x < b$ is an interval estimate of the population mean μ.

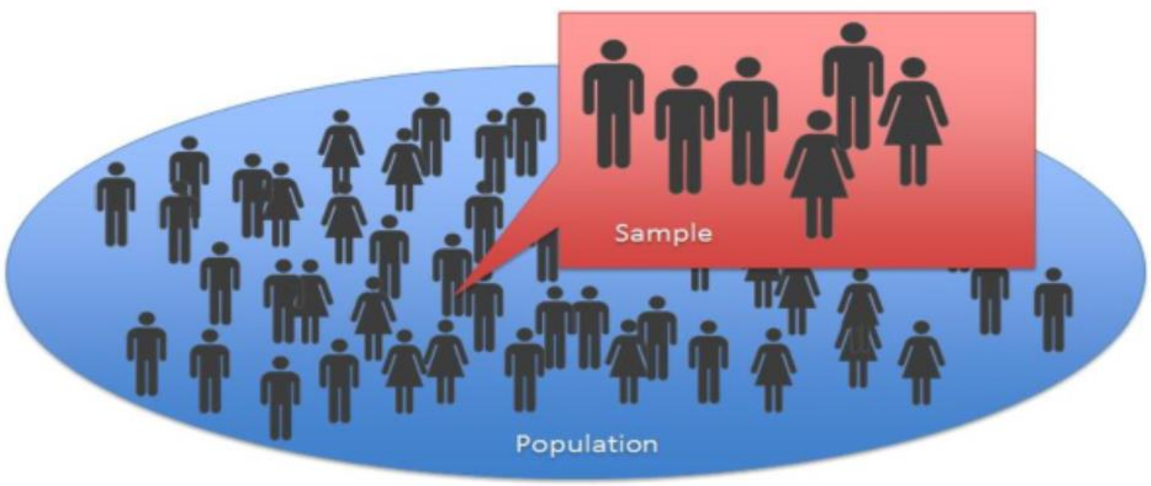
Example: Cilantro-Haters

- Some people taste cilantro like soap.

- How many percent p of people taste cilantro like soap?

Example: Cilantro-Haters

- **Experiment**: Ask n random people to taste cilantro.
- **Model**: $X_i \sim Bernoulli(p)$ is whether the i-th person says it taste like soap.
- **Data**: $x_1, \ldots, x_n$ are the results of the experiment.
- **Inference**: Estimate p from the data.



Example: Cilantro-Haters

- Asking 100 people to taste cilantro and 65 people say that it tastes like soap.

- Use this data to estimate p the fraction of all people for whom it tastes like soap.

- So, p is the **parameter of interest**.

What is Likelihood?

- For a given value of p the probability of getting 65 'successes' is the binomial probability.

The likelihood $P(data \mid p) = \binom{100}{65} p^{65}(1-p)^{35}$

- Note: The likelihood takes the **data as fixed** and computes the probability of the data for a given p.

Maximum Likelihood Estimation (MLE)

- **It is a way to estimate the value of a parameter of interest.**

- **Finding the value of p that maximizes the likelihood**

- **There are different methods of finding the maximum**
  - Calculus: Solve $\frac{d}{dp}P(data \mid p) = 0$ for p.
    - We should also check that the **critical point is a maximum**.
  - Sometimes, the derivative is never 0.
    - It is at an endpoint of the allowable range.

Computing MLE

- **The MLE is computed by calculus.**

$$\frac{dP(\, data \mid p)}{dp} = \binom{100}{65}\left(65p^{64}(1-p)^{35} - 35p^{65}(1-p)^{34}\right) = 0$$

- **A sequence of algebraic steps gives:**

$$65p^{64}(1-p)^{35} = 35p^{65}(1-p)^{34}$$

$$65(1-p) = 35p$$

$$65 = 100p$$

$$\hat{p} = \frac{65}{100}$$

Log Likelihood

- **Because the log function turns multiplication into addition, it is convenient to use the log of the likelihood function.**

$$\textbf{Log likelihood} = \ln(\mathbf{likelihood}) = \ln(\boldsymbol{P}(\boldsymbol{data} \mid \boldsymbol{p}))$$

- **Example**

The likelihood is $P(data \mid p) = \binom{100}{65} p^{65}(1-p)^{35}$

The log likelihood is $\ln \binom{100}{65} + 65 \ln p + 35 \ln(1-p)$

Computing MLE with Log Likelihood

- **The MLE is computed by calculus.**

$$\frac{dP(\,data\,|\,p)}{dp} = \ln\binom{100}{65} + 65\ln p + 35\ln(1-p) = 0$$

- **A sequence of algebraic steps gives:**

$$\frac{65}{p} - \frac{35}{1-p} = 0$$

$$65(1-p) = 35p$$

$$65 = 100p$$

$$\hat{p} = \frac{65}{100}$$

Discussion

- Our data was 10 people, and 6 out of 10 people tasted cilantro like soap. Is it okay?

- Intuitively, we need a large enough sample size to make a conclusion. How large?

- Note: we need mathematical modeling to understand the accuracy of this procedure.

MLE vs. MAP

Maximum Likelihood Estimation (MLE)

- Estimate the **maximum likelihood** given **independent** observations $x_1, x_2, \ldots, x_n$.

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

- As the function of θ, what θ **maximizes the likelihood** of the observed data?

$$\frac{\partial}{\partial \theta} \mathcal{L}(\theta; x_1, \ldots, x_n) = 0$$

Maximum Likelihood Estimation (MLE)

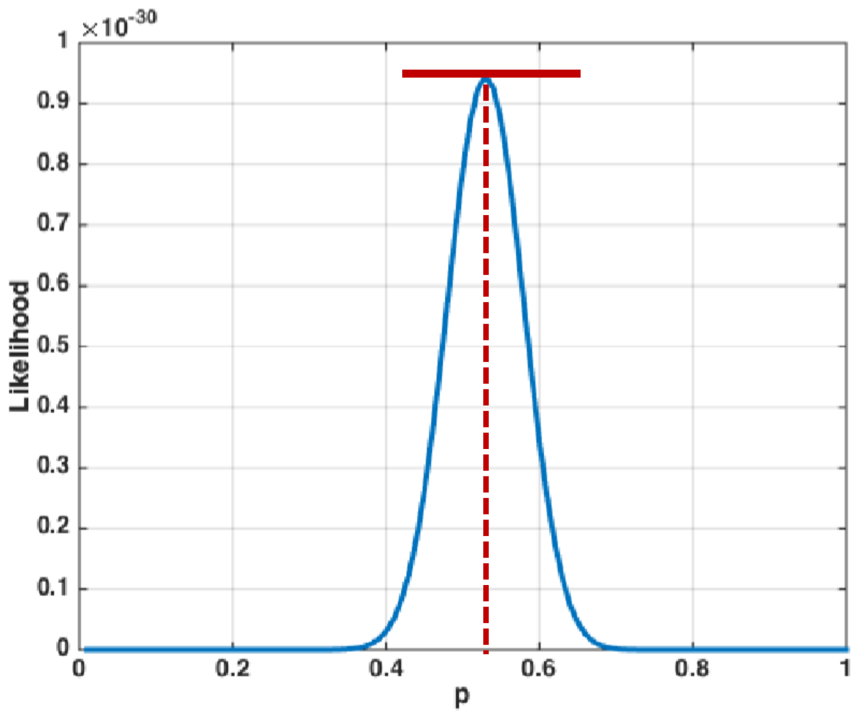
- We take a **derivative** and set it to zero.

$$\frac{\partial}{\partial\theta}\log\sum_{i}\boldsymbol{P}(\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathbf{0}$$

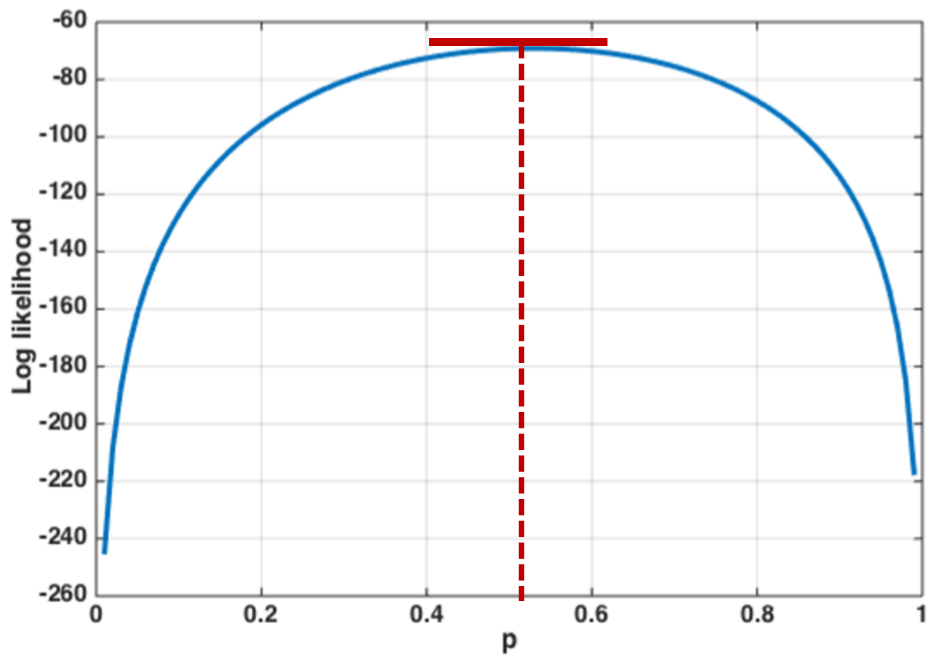
- Solving for $\frac{\partial}{\partial\theta}\log\sum_i \boldsymbol{P}(\boldsymbol{X}_i;\boldsymbol{\theta}) = \mathbf{0}$

Likelihood vs. Log Likelihood

- **Log-likelihood is a monotonic function of the likelihood.**
  - The maximum is achieved at the same point.
  - In most cases, log-likelihood requires much less computation.



Likelihood



Log-likelihood

Coin Toss Problem

➢ **What is a best estimate of θ given k heads of n tosses?**

➢ **Flip it repeatedly, observing that**

- It turns up heads k times.
- It turns up tails $n - k$ times.

$$P(X \mid \theta) = \theta^k (1 - \theta)^{n-k} \quad \text{where } \theta = P(X = head)$$

➢ **Consider the maximization problem.**

$$\max_{0 \le \theta \le 1} P(X \mid \theta) = \max_{0 \le \theta \le 1} \theta^k (1 - \theta)^{n-k}$$

How to Estimate $\theta = P(X = heads)$

$$\underset{\theta \in [0,1]}{\operatorname{argmax}} \theta^k (1-\theta)^{n-k}$$

- **Observe that we have 5 heads out of 5 tosses.**
  - What is the best estimate of θ?
- **Observe that we have 0 heads out of 5 tosses.**
  - What is the best estimate of θ?
- **Observe that we have 4 heads out of 5 tosses.**
  - What is the best estimate of θ?

Parameter Estimation for Coin Toss

- **Assuming that sample $x_1, x_2, \dots, x_n$ is from a parametric distribution $P(X \mid \theta)$, estimate a parameter θ.**

- **Given a sample $HHTHH$ of coin flips, estimate θ.**
  - θ: Probability that the coin turns up heads

- **$P(X \mid \theta)$: a probability function with a parameter θ**

Likelihood: Relative Values of Interest

- **$P(X \mid \theta)$: Probability of event X given a parameter θ**

- **Given fixed θ, it is the function of X**
  - This is a ***probability***, $\sum_X P(X \mid \theta) = 1$

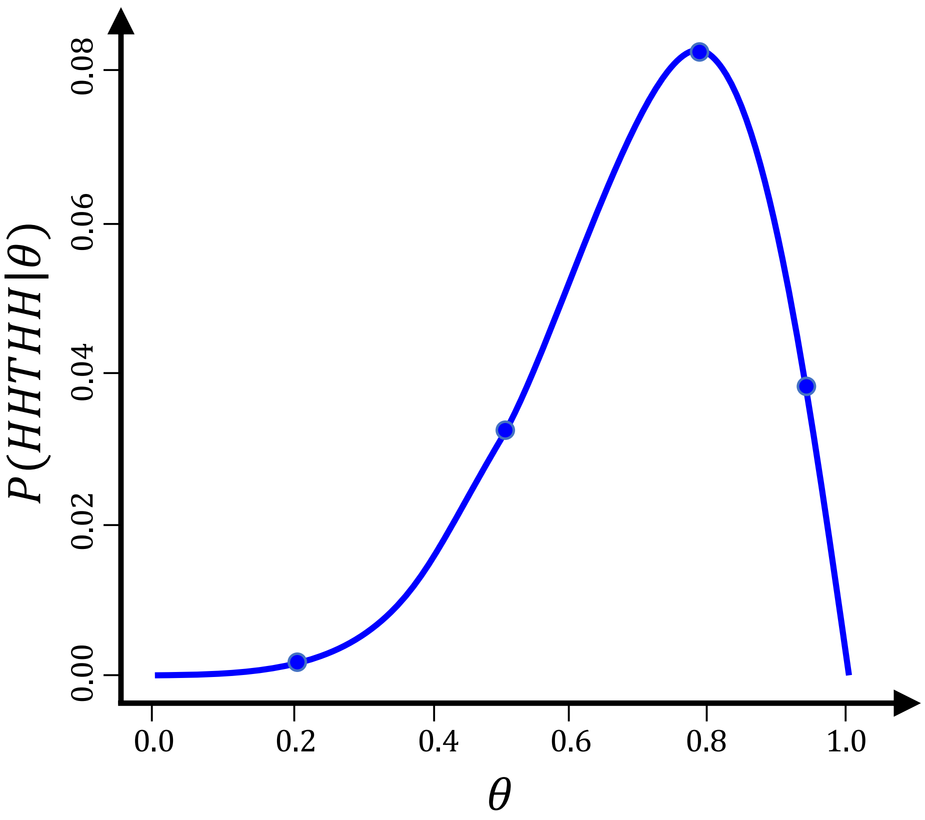
- **Given fixed X, it is the function of θ.**
  - This is a ***likelihood***, $\sum_\theta P(X \mid \theta)$ can be **anything**.
  - **Relative values of interest**
    - An event $HHTHH$ is more likely when $\theta = 0.6$ than $\theta = 0.5$
    - E.g., $P(HHTHH \mid \theta = 0.6) > P(HHTHH \mid \theta = 0.5)$

- **What θ makes $HHTHH$ *most likely*?**

Example: Likelihood Function

- Distribution for the probability of HHTHH, given P(H) = θ

θ	$\theta^4(1-\theta)$
0.2	0.0013
0.5	0.0313
…	…
0.8	0.0819
…	…
0.95	0.0407

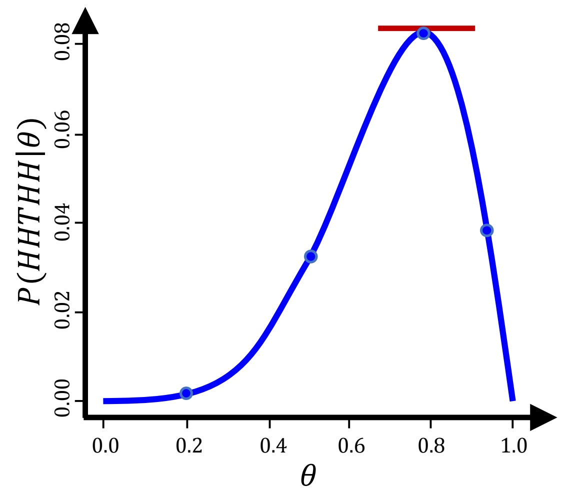


https://www.desmos.com/calculator

Log Likelihood Estimation

- Given n coin flips $x_1, x_2, \dots, x_n$ with k heads and $\mathrm{n} - k$ tails,
- θ = probability of heads

$$\frac{\partial L}{\partial \theta} = 0$$



$$\mathcal{L}(\theta; x_1, \dots, x_n) = \theta^k (1-\theta)^{n-k}$$

$$\ln \mathcal{L}(\theta; x_1, \dots, x_n) = k \ln \theta + (n-k) \ln(1-\theta)$$

- Setting to zero and solving

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; x_1, \dots, x_n) = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

$$\hat{\theta} = \frac{k}{n}$$

Coin Toss with MLE

- Each flip yields a Boolean value for X

$$X \sim Bernoulli: P(X) = \theta$$

- Toss flips produce ones and zeros.

$$P(X \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

$$\Downarrow$$

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

Estimiating $\theta = P(X = heads)$

- Test A: For 100 flips, 79 heads, 21 tails

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{79}{79 + 21} = \frac{79}{100}$$

- Test B: For 3 flips, 3 heads, 0 tails

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{3}{3 + 0} = 1$$

- Are they fair? If not, why?

One Possible Heuristic

- **While keeping flipping, we want to design a single learning algorithm that gives a reasonable estimate after each flip.**

- **How to design the algorithm?**

$$\lambda \times \frac{\alpha_H}{\alpha_H + \alpha_T} + (1-\lambda) \times \frac{1}{2}$$

  - Your **belief** in flipping a coin is **1/2**.
  - λ is the parameter to control the belief of your observations.

Maximum a Posterior (MAP) Estimation

- **Similar to MLE, it is a parameter estimation method from a given training data.**
  - It incorporates a **prior distribution** that quantifies **additional information** through **prior knowledge of a related event**.

$$\hat{\theta}_{MAP}(x) = \underset{\theta}{\operatorname{argmax}}\, f(\theta \mid x)$$

Likelihood

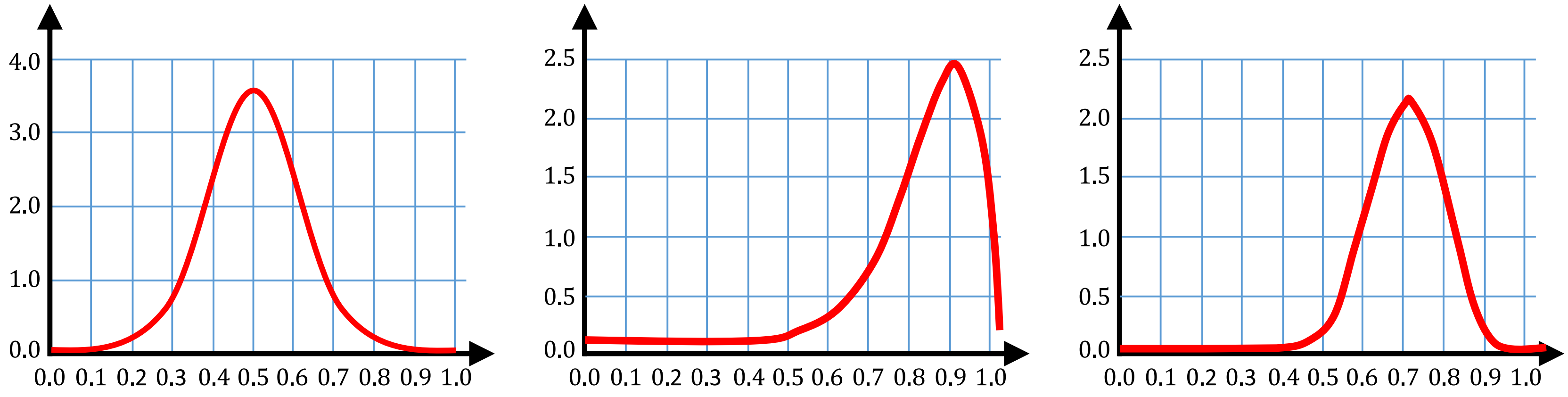
$$= \underset{\theta}{\operatorname{argmax}} \frac{f(x \mid \theta) f(\theta)}{f(x)} = \underset{\theta}{\operatorname{argmax}}\, f(x \mid \theta) f(\theta)$$

Prior

- **The MAP estimate of θ coincides with the ML estimate when the prior is uniform, i.e., it is constant.**

Example: Coin Tossing

- Before tossing coins, we assume that the probability of being head is 0.5.
- It is observed that 80 out of 100 are head.
- After 100 trials, our estimation can be changed to a more significant number than 0.5.



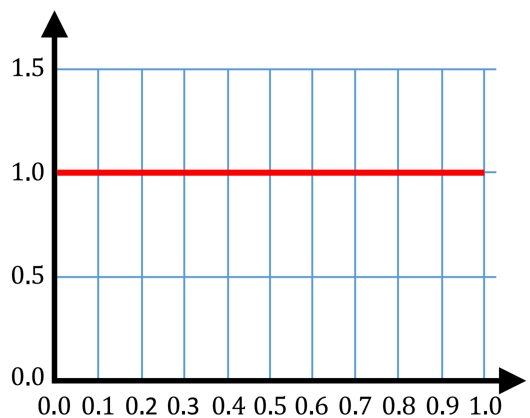
Prior Likelihood Posterior

Prior $P(\theta)$ for Coin Toss

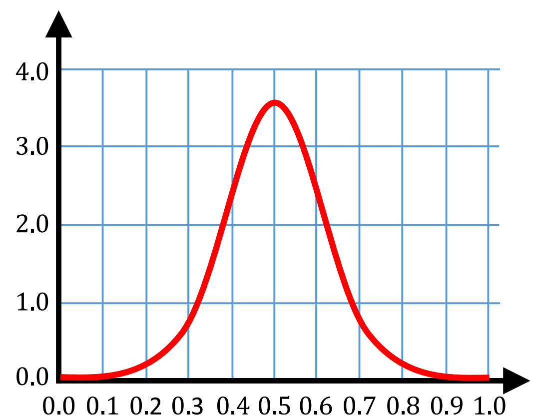
- **Beta distribution as the prior**

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

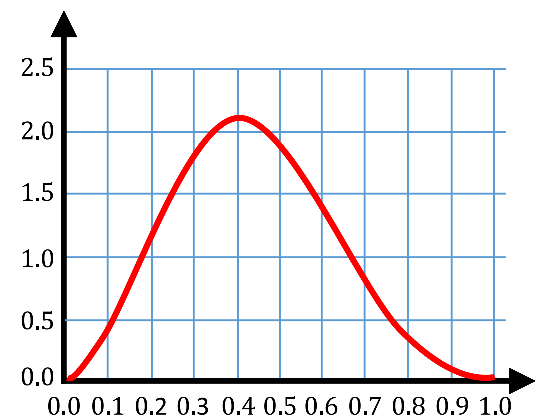
$\boldsymbol{Beta(\alpha, = 1\ \beta = 1)}$



$\boldsymbol{Beta(\alpha, = 10\ \beta = 10)}$



$\boldsymbol{Beta(\alpha, = 3\ \beta = 4)}$



http://keisan.casio.com/exec/system/1180573226

Coin Toss with MAP

- The likelihood is binomial.

$$P(X \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

- If the prior is the beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- MAP estimate is

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{(\alpha_H + \beta_H - 1) + (\alpha_T + \beta_T - 1)}$$

$$\hat{\theta}_{MAP} = P(\theta \mid X) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Detail: Coin Toss with MAP

- If the prior is the beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- We can derive the MAP as follows.

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \ln \theta^{\alpha_H}(1-\theta)^{\alpha_T} \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$$

$$\Downarrow$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \ln \frac{\theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_T - 1}}{B(\beta_H, \beta_T)}$$

Principles for Estimating Parameters

- **Principle 1: Maximum likelihood estimation (MLE)**
  - Choose a parameter that maximizes $P(X \mid \theta)$

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- **Principle 2: Maximum a Posteriori (MAP)**
  - Choose a parameter that maximizes $P(\theta \mid X)$

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#\, of\ hallucinated\_heads}{(\alpha_H + \#\, of\ hallucinated\_heads) + (\alpha_T + \#\, of\ hallucinated\_tails)}$$

Principles for Estimating Parameters

- **Which is better? MLE vs. MAP**

$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

vs.

$$\hat{\theta}_{MAP} = \frac{\alpha_H + \#\, of\;\; hallucinated\_heads}{(\alpha_H + \#\, of\;\; hallucinated\_heads) + (\alpha_T + \#\, of\;\; hallucinated\_tails)}$$

- **What if the number of samples is small?**
- **What if the number of samples is large?**

Q&A

How to Compute MAP?

- MAP estimates can be computed in several ways.

- Analytically, when the mode(s) of the posterior distribution can be given in a closed-form.
  - This is the case when conjugate priors are used.
- Use numerical optimization such as the conjugate gradient method or Newton's method.
  - This usually requires first or second derivatives, which are evaluated analytically or numerically.
- Use a modification of an expectation-maximization algorithm.
  - This does not require derivatives of the posterior density.
- Use a Monte Carlo method using simulated annealing.