

Parameter Estimation II: Bayesian

Dr. Edmund Lam

Department of Electrical and Electronic Engineering
The University of Hong Kong

ELEC2844: Probabilistic Systems Analysis
(Second Semester, 2019-20)

<http://www.eee.hku.hk/~elec2844>

Motivation

Suppose we have a possibly biased coin with unknown probability p that Head will show up. Our best strategy to estimate p is perhaps to experiment:

- What if we flip 1000 times and you get 500 heads?
- What if we flip 999 times and you get 500 heads?
- What if we flip once and you get head?
- What if we flip 10 times and you get 4 heads, *but the coin looks “normal” and you would normally think that it is a fair coin?*

In this lecture, we will explore how to incorporate **prior belief** (i.e., our view of the parameter that does not depend on the observations) in estimating probabilistic quantities based on observations.

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior
- 6 Sequential learning with MAP
- 7 Hypothesis testing

A probabilistic view of the unknown parameter

In ML, the parameter θ is assumed to be *fixed*, but unknown. This is the *classical* or *frequentist* approach.

But there is an alternative approach known as **Bayesian**: the parameter θ is a *random variable* in itself, with a certain distribution, and our estimate is the value that gives the largest value in the distribution.

A probabilistic view of the unknown parameter

In ML, the parameter θ is assumed to be *fixed*, but unknown. This is the *classical* or *frequentist* approach.

But there is an alternative approach known as **Bayesian**: the parameter θ is a *random variable* in itself, with a certain distribution, and our estimate is the value that gives the largest value in the distribution.

The general approach:

- 1 Start with a certain PMF or PDF
- 2 Observe data
- 3 Based on the observation, adjust the PMF or PDF
- 4 Make the estimate based on maximizing the adjusted PMF or PDF

Latin lesson

In probability, people like to use these two latin phrases:

a priori = “beforehand”
a posteriori = “afterwards”

- *A priori* knowledge is the “prior”, which usually comes from a mathematical model;
- *A posteriori* knowledge comes from observation or empirical evidence.

Hence, our task is to morph from the *a priori* distribution (**prior probability**) to the *a posteriori* distribution (**posterior probability**) based on observations.

Bayesian

In this Bayesian approach, the key equation is the Bayes' rule:

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (1)$$

Θ is the random variable to be estimated, and $X = x$ the observation.

Bayesian

In this Bayesian approach, the key equation is the Bayes' rule:

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (1)$$

Θ is the random variable to be estimated, and $X = x$ the observation.

The interpretation is:

$$\text{posterior} = \text{prior} \times \text{likelihood} \quad (2)$$

Bayesian

In this Bayesian approach, the key equation is the Bayes' rule:

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (1)$$

Θ is the random variable to be estimated, and $X = x$ the observation.

The interpretation is:

$$\text{posterior} = \text{prior} \times \text{likelihood} \quad (2)$$

A further interpretation:

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}) \quad (3)$$

So $\log(\text{prior})$ is an *extra term* when we try to maximize the posterior as opposed to maximizing the likelihood.

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples**
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior
- 6 Sequential learning with MAP
- 7 Hypothesis testing

Example 1

Assuming that a certain random variable Θ has the following distribution

$$f_{\Theta}(\theta) = \begin{cases} 3\theta^2 & 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, another random variable X is related to Θ , such that $(X|\Theta = \theta) \sim \text{Geo}(\theta)$. Given that we observe $X = 2$, find the MAP estimate of Θ .

Example 1

Because $X|\Theta$ is a geometric distribution,

$$f_{X|\Theta}(x|\theta) = (1 - \theta)^{x-1} \theta$$

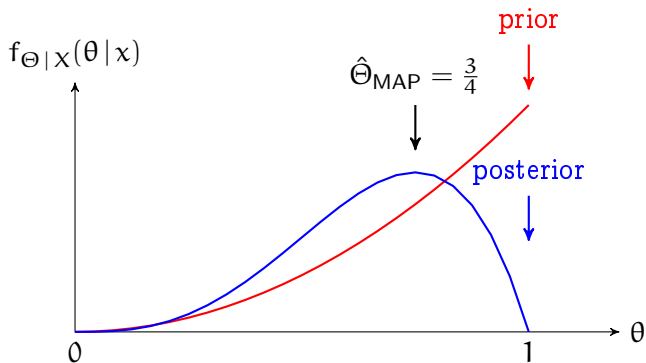
In MAP estimation,

$$\begin{aligned} f_{\Theta|X}(\theta|x=2) &\propto f_{\Theta}(\theta) p_{X|\Theta}(x=2|\theta) \\ &= (3\theta^2) [(1 - \theta)^{2-1} \theta] \\ &= 3\theta^3(1 - \theta) \end{aligned}$$

We need to maximize this expression, so we take the derivative w.r.t. θ and set to zero, i.e.,

$$\begin{aligned} 3\theta^2 - 4\theta^3 &= 0 \\ \theta &= \frac{3}{4} \end{aligned}$$

Example 1



Example 2: Poisson with exponential prior

Suppose random variables $X_1, X_2, \dots, X_n$ are independently generated according to a Poisson random variable with parameter θ , but I only observe their arithmetic mean, i.e.,

$$s = \frac{1}{n} \sum_{i=1}^n x_i$$

where each x_i are the generated numbers. Furthermore, I know that θ itself is randomly generated, according to an exponential distribution

$$f_{\Theta}(\theta) = \lambda e^{-\lambda\theta}.$$

Find the MAP estimate of Θ given observation s .

Example 2: Poisson with exponential prior

In the discussion of ML estimation, we have already noted that the log-likelihood is

$$\begin{aligned}
 \log p_X(x_1, \dots, x_n; \theta) &= \sum_{i=1}^n \log p_{X_i}(x_i; \theta) \\
 &= -n\theta + \sum_{i=1}^n x_i \log \theta - \sum_{i=1}^n \log(x_i!) \\
 &= -n\theta + ns \log \theta - \sum_{i=1}^n \log(x_i!)
 \end{aligned}$$

The log-prior is $\log f_{\Theta}(\theta) = \log \lambda - \lambda\theta$.

The log-posterior is the sum of the two, i.e., (C denoting what does not involve θ)

$$\log f_{\Theta|S}(\theta | s) = C - n\theta + ns \log \theta - \lambda\theta$$

Example 2: Poisson with exponential prior

To maximize this, we take the derivative w.r.t. θ and set to zero,

$$-n + \frac{ns}{\theta} - \lambda = 0$$

$$\theta = \frac{ns}{n + \lambda} = \frac{s}{1 + \frac{\lambda}{n}}$$

So the MAP estimate is $\Theta_{\text{MAP}} = \frac{s}{1 + \frac{\lambda}{n}}$.

In the earlier discussion on ML estimate, we already know that the ML estimate of a Poisson distribution is $\Theta_{\text{ML}} = s$. Here, we note two situations:

- If λ is small, the prior becomes flat, and MAP approaches ML.
- If n is large, we have more and more trust on the data, and MAP approaches ML.

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations**
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior
- 6 Sequential learning with MAP
- 7 Hypothesis testing

Example 1: additive noise

A signal X is transmitted through a certain wireless channel, which is corrupted by an additive noise W . Suppose we know that both are Gaussian distributed with zero mean, i.e.,

$$X \sim \mathcal{N}(0, \sigma_X^2)$$

$$W \sim \mathcal{N}(0, \sigma_W^2)$$

and that they are independent of each other. The received signal is $Y = X + W$. Given that we observe $Y = y$, find:

- 1 the ML estimate of X
- 2 the MAP estimate of X

Example 1: additive noise

In thinking about the ML estimate, we assume that X has taken on a value x , which is fixed but unknown, and we try to estimate this value based on our observation.

With a fixed x , we can see that Y is a Gaussian distribution with mean x and variance σ_W^2 . The likelihood function is

$$f_Y(y; x) = \frac{1}{\sqrt{2\pi}\sigma_W} e^{-(y-x)^2/2\sigma_W^2}$$

We can maximize the log-likelihood, i.e.,

$$\log f_Y(y; x) = \log \frac{1}{\sqrt{2\pi}\sigma_W} - \frac{(y-x)^2}{2\sigma_W^2}$$

which we can see that it happens when $x = y$. So, our ML estimate is

$$\hat{X}_{ML} = y$$

Example 1: additive noise

In thinking about the MAP estimate, the posterior PDF of X is

$$\begin{aligned} f_{X|Y}(x|y) &\propto f_X(x) f_{Y|X}(y|x) \\ &= c e^{-x^2/2\sigma_X^2} e^{-(y-x)^2/2\sigma_W^2} \end{aligned}$$

where c is a certain constant that does not depend on the value of x . Note that given $X = x$, Y is a Gaussian distribution with mean x and variance σ_W^2 .

We need to maximize the *logarithm* of the above expression, which is equivalent to minimizing

$$\frac{x^2}{2\sigma_X^2} + \frac{(y-x)^2}{2\sigma_W^2}$$

Example 1: additive noise

We take the derivative (w.r.t. x) and set to zero, i.e.,

$$\begin{aligned}\frac{2x}{2\sigma_X^2} - \frac{2(y-x)}{2\sigma_W^2} &= 0 \\ \sigma_W^2 x &= \sigma_X^2 (y-x) \\ x &= \frac{\sigma_X^2}{\sigma_W^2 + \sigma_X^2} y\end{aligned}$$

So, our MAP estimate is

$$\hat{X}_{\text{MAP}} = \left(\frac{\sigma_X^2}{\sigma_W^2 + \sigma_X^2} \right) y$$

Unlike the ML estimate, the MAP estimate is saying that if the noise is large, we will simply estimate X to be close to zero (remember it's a zero-mean Gaussian) irrespective of what we have observed.

Example 2a: flipping a coin

A coin has an unknown probability θ to result in head. *We model the prior PDF as a uniform distribution, where*

$$f_{\Theta}(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We toss the coin n times, and let $X = k$ be the number of heads observed. Find the MAP estimate.

Example 2a: flipping a coin

A coin has an unknown probability θ to result in head. *We model the prior PDF as a uniform distribution, where*

$$f_{\Theta}(\theta) = \begin{cases} 1 & 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We toss the coin n times, and let $X = k$ be the number of heads observed. Find the MAP estimate.

ANS: The number of heads given $\Theta = \theta$ is given by

$$p_{X|\Theta}(k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \quad k = 0, 1, 2, \dots, n$$

From Bayes' rule, the **posterior probability** of Θ is

$$f_{\Theta|X}(\theta|k) \propto f_{\Theta}(\theta) p_{X|\Theta}(k|\theta) = c \theta^k (1-\theta)^{n-k}$$

where c is a constant independent of θ .

Example 2a: flipping a coin

We obtain the MAP estimate by maximizing

$$f_{\Theta|X}(\theta|k) = c\theta^k(1-\theta)^{n-k}$$

Example 2a: flipping a coin

We obtain the MAP estimate by maximizing

$$f_{\Theta|X}(\theta|k) = c\theta^k(1-\theta)^{n-k}$$

Taking the derivative w.r.t. θ and setting it to zero,

$$0 = ck(1-\theta)^{n-k}\theta^{k-1} - c(n-k)(1-\theta)^{n-k-1}\theta^k$$

which yields the MAP estimator $\hat{\Theta} = \frac{k}{n}$.

*This is exactly the same as the ML estimator! The reason is because we have an **uninformative prior**.*

Example 2b: flipping a coin

A coin has an unknown probability θ to result in head. *We model the prior PDF as*

$$f_{\Theta}(\theta) = \begin{cases} 2 - 4 \cdot \left| \frac{1}{2} - \theta \right| & 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We toss the coin n times, and let $X = k$ be the number of heads observed. Find the MAP estimate.

Example 2b: flipping a coin

ANS: From Bayes' rule, the posterior PDF of Θ is

$$\begin{aligned} f_{\Theta|X}(\theta|k) &\propto f_{\Theta}(\theta) p_{X|\Theta}(k|\theta) \\ &= c \left(2 - 4 \cdot \left| \frac{1}{2} - \theta \right| \right) \theta^k (1 - \theta)^{n-k} \end{aligned}$$

Example 2b: flipping a coin

ANS: From Bayes' rule, the posterior PDF of Θ is

$$\begin{aligned} f_{\Theta|X}(\theta|k) &\propto f_{\Theta}(\theta) p_{X|\Theta}(k|\theta) \\ &= c \left(2 - 4 \cdot \left| \frac{1}{2} - \theta \right| \right) \theta^k (1 - \theta)^{n-k} \end{aligned}$$

How do we maximize this w.r.t. θ ? We divide into three cases: (a) maximum is at $\theta < 1/2$; (b) maximum is at $\theta > 1/2$; (c) maximum is at $\theta = 1/2$.

For (a) and (b), we need the derivative to be zero; for (c), note that the function is not differentiable.

Example 2b: flipping a coin

(a) For $\theta < 1/2$: We have $f_{\Theta|X}(\theta|k) \propto \theta^{k+1}(1-\theta)^{n-k}$. Taking the derivative w.r.t. θ and setting it to zero, similar calculations as above would yield the MAP estimator

$$\hat{\Theta} = \frac{k+1}{n+1}.$$

Example 2b: flipping a coin

(a) For $\theta < 1/2$: We have $f_{\Theta|X}(\theta|k) \propto \theta^{k+1}(1-\theta)^{n-k}$. Taking the derivative w.r.t. θ and setting it to zero, similar calculations as above would yield the MAP estimator

$$\hat{\Theta} = \frac{k+1}{n+1}.$$

(b) For $\theta > 1/2$: We have $f_{\Theta|X}(\theta|k) \propto \theta^k(1-\theta)^{n-k+1}$. Taking the derivative w.r.t. θ and setting it to zero, similar calculations as above would yield the MAP estimator

$$\hat{\Theta} = \frac{k}{n+1}.$$

Example 2b: flipping a coin

(a) For $\theta < 1/2$: We have $f_{\Theta|X}(\theta|k) \propto \theta^{k+1}(1-\theta)^{n-k}$. Taking the derivative w.r.t. θ and setting it to zero, similar calculations as above would yield the MAP estimator

$$\hat{\Theta} = \frac{k+1}{n+1}.$$

(b) For $\theta > 1/2$: We have $f_{\Theta|X}(\theta|k) \propto \theta^k(1-\theta)^{n-k+1}$. Taking the derivative w.r.t. θ and setting it to zero, similar calculations as above would yield the MAP estimator

$$\hat{\Theta} = \frac{k}{n+1}.$$

(c) Situations other than (a) or (b) would be (c).

Example 2b: flipping a coin

Summarizing, the MAP estimate is given by

$$\hat{\Theta} = \begin{cases} \frac{k+1}{n+1} & \text{if } \frac{k+1}{n+1} < \frac{1}{2} \\ \frac{1}{2} & \text{if } \frac{k}{n+1} \leq \frac{1}{2} \leq \frac{k+1}{n+1} \\ \frac{k}{n+1} & \text{if } \frac{1}{2} < \frac{k}{n+1} \end{cases}$$

Example 2b: flipping a coin

Summarizing, the MAP estimate is given by

$$\hat{\Theta} = \begin{cases} \frac{k+1}{n+1} & \text{if } \frac{k+1}{n+1} < \frac{1}{2} \\ \frac{1}{2} & \text{if } \frac{k}{n+1} \leq \frac{1}{2} \leq \frac{k+1}{n+1} \\ \frac{k}{n+1} & \text{if } \frac{1}{2} < \frac{k}{n+1} \end{cases}$$

For example, if $n = 9$, then

k	0	1	2	3	4	5	6	7	8	9
$\hat{\Theta}$	0.1	0.2	0.3	0.4	0.5	0.5	0.6	0.7	0.8	0.9

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate**
- 5 Conjugate prior
- 6 Sequential learning with MAP
- 7 Hypothesis testing

MAP and LMS

Bayesian approach actually gives us another possible estimate. Consider the posterior probability

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta)$$

where Θ is the random variable to be estimated, and $X = x$ the observation.

- MAP estimate:

$$\hat{\Theta}_{\text{MAP}} = \arg \max_{\Theta} f_{\Theta|X}(\theta|x)$$

- Least mean squares (LMS) estimate:

$$\hat{\Theta}_{\text{LMS}} = \mathbb{E}(\Theta|X) \quad (4)$$

MAP and LMS

Comparison of MAP and LMS:

- MAP finds the *maximum*, or mode of the posterior distribution
- LMS finds the *conditional expectation* of the posterior distribution. It minimizes the mean squared error over all estimators.
- The conditional variance $\text{var}(\Theta|X)$ also gives us information about how reliable the estimator is.
- If the posterior distribution is symmetric around its conditional mean and unimodal, then $\text{MAP} = \text{LMS}$.

LMS example

Assuming that a certain random variable Θ has the following distribution

$$f_{\Theta}(\theta) = \begin{cases} 3\theta^2 & 0 < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, another random variable X is related to Θ , such that $(X|\Theta = \theta) \sim \text{Geo}(\theta)$. Given that we observe $X = 2$, find the LMS estimate of Θ .

LMS example

We had looked into this problem with MAP estimation, and note that the posterior is (except for the constant)

$$f_{\Theta|X}(\theta|x=2) \propto \theta^3(1-\theta)$$

Let K be the proportionality constant. Since this is a PDF, it must integrate to unity,

$$\int_0^1 K\theta^3(1-\theta)d\theta = 1$$

$$K \left[\frac{1}{4}\theta^4 - \frac{1}{5}\theta^5 \right]_0^1 = 1$$

$$K = 20$$

LMS example

Hence, the LMS estimate (conditional expectation) is

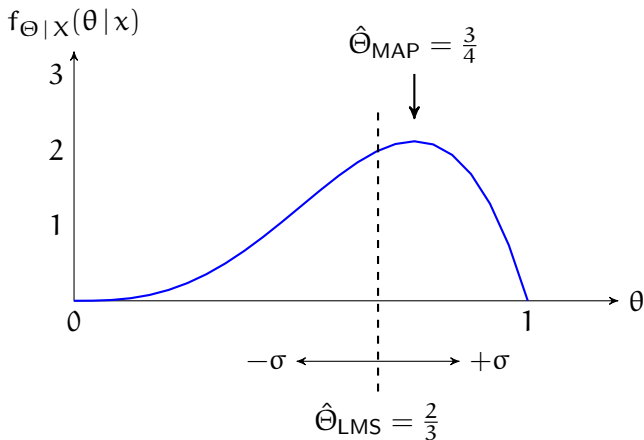
$$\begin{aligned}\mathbb{E}(\Theta | X = 2) &= \int_0^1 \theta \cdot 20\theta^3(1 - \theta) d\theta \\ &= 20 \left[\frac{1}{5}\theta^5 - \frac{1}{6}\theta^6 \right]_0^1 = \frac{2}{3}\end{aligned}$$

Also, for the conditional variance,

$$\begin{aligned}\mathbb{E}(\Theta^2 | X = 2) &= \int_0^1 \theta^2 \cdot 20\theta^3(1 - \theta) d\theta \\ &= 20 \left[\frac{1}{6}\theta^6 - \frac{1}{7}\theta^7 \right]_0^1 = \frac{10}{21} \\ \text{var}(\Theta | X) &= \frac{10}{21} - \left(\frac{2}{3} \right)^2 = \frac{2}{63}\end{aligned}$$

LMS example

Putting things in perspective:



- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior**
- 6 Sequential learning with MAP
- 7 Hypothesis testing

Example

A coin has an unknown probability θ to result in head. *We model the prior PDF as a beta distribution, where*

$$f_{\Theta}(\theta) \sim \text{Beta}(\theta; \alpha, \beta) = \begin{cases} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} & 0 < \theta < 1 \\ 0 & \text{otherwise,} \end{cases}$$

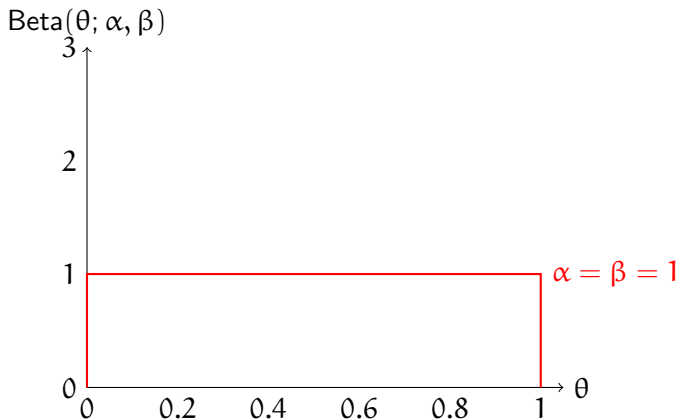
where $B(\alpha, \beta)$ is a normalizing constant, known as the beta function, given by

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} = \frac{(\alpha - 1)!(\beta - 1)!}{(\alpha + \beta - 1)!}$$

(We can use the factorial notation if the quantity is a nonnegative integer.) We toss the coin n times, and let $X = k$ be the number of heads observed. Find the MAP estimate.

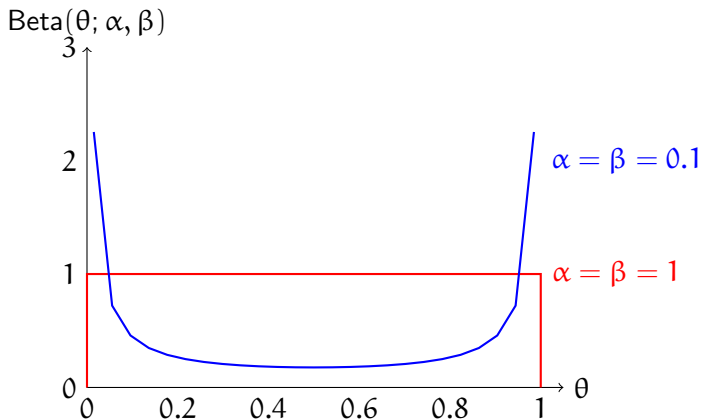
Beta distribution

The beta distribution takes on different shapes for various values of α and β , which are not necessarily integers. Specifically, when $\alpha = \beta = 1$, we get the uniform distribution.



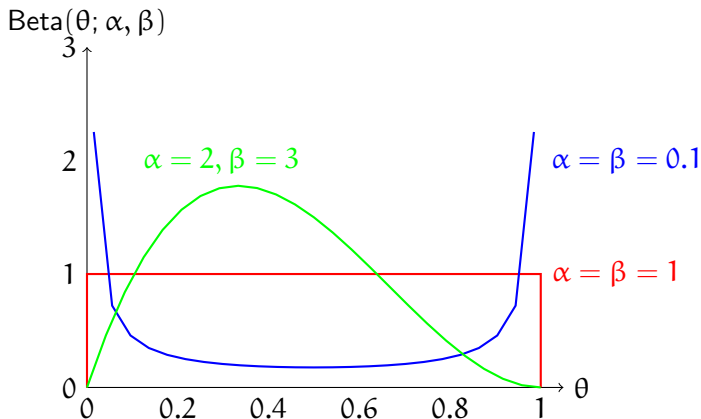
Beta distribution

The beta distribution takes on different shapes for various values of α and β , which are not necessarily integers. Specifically, when $\alpha = \beta = 1$, we get the uniform distribution.



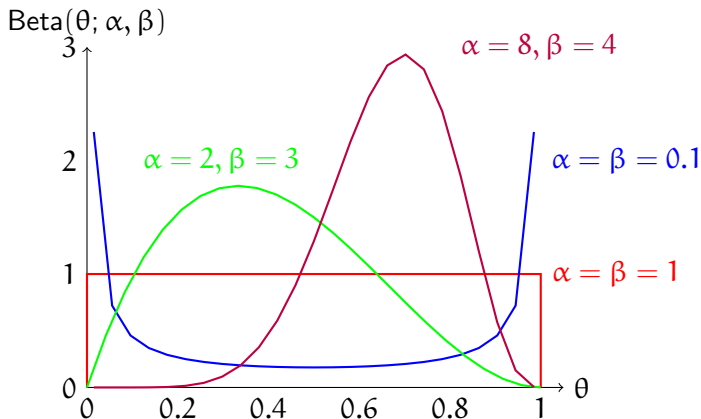
Beta distribution

The beta distribution takes on different shapes for various values of α and β , which are not necessarily integers. Specifically, when $\alpha = \beta = 1$, we get the uniform distribution.



Beta distribution

The beta distribution takes on different shapes for various values of α and β , which are not necessarily integers. Specifically, when $\alpha = \beta = 1$, we get the uniform distribution.



Example

ANS: From Bayes' rule, the posterior PDF of Θ is

$$\begin{aligned}
 f_{\Theta|X}(\theta|k) &\propto f_{\Theta}(\theta) p_{X|\Theta}(k|\theta) \\
 &= \left(\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \right) \theta^k (1-\theta)^{n-k} \\
 &= c \theta^{k+\alpha-1} (1-\theta)^{n-k+\beta-1}
 \end{aligned}$$

Example

ANS: From Bayes' rule, the posterior PDF of Θ is

$$\begin{aligned} f_{\Theta|X}(\theta|k) &\propto f_{\Theta}(\theta) p_{X|\Theta}(k|\theta) \\ &= \left(\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \right) \theta^k (1-\theta)^{n-k} \\ &= c \theta^{k+\alpha-1} (1-\theta)^{n-k+\beta-1} \end{aligned}$$

Taking the derivative w.r.t. θ and setting it to zero,

$$0 = (k+\alpha-1)(1-\theta)^{n-k+\beta-1} \theta^{k+\alpha-2} - (n-k+\beta-1)(1-\theta)^{n-k+\beta-2} \theta^{k+\alpha-1}$$

which yields the MAP estimator $\hat{\Theta} = \frac{k + (\alpha - 1)}{n + (\alpha + \beta - 2)}$.

Designing the prior

The beta distribution is called a **conjugate prior** of the Bernoulli, because they have the same form

$$\theta^{\gamma_1} (1 - \theta)^{\gamma_2}$$

for some γ_1 and γ_2 .

Designing the prior

The beta distribution is called a **conjugate prior** of the Bernoulli, because they have the same form

$$\theta^{\gamma_1} (1 - \theta)^{\gamma_2}$$

for some γ_1 and γ_2 .

The parameters of the prior are called **hyper-parameters**. We can set them to encode our prior beliefs.

Designing the prior

Example: Suppose we want $\mathbb{E}(\theta) = 0.7$ and $\text{var}(\theta) = 0.04$.

Designing the prior

Example: Suppose we want $\mathbb{E}(\theta) = 0.7$ and $\text{var}(\theta) = 0.04$.

With a bit of mathematical derivations, we can show that the beta distribution has the following expressions for mean and variance:

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$\text{var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Plugging in $\frac{\alpha}{\alpha + \beta} = 0.7$, we find $\alpha = \frac{7}{3}\beta$.

We then use the expression for the variance, and after some calculations, we have $\beta = \frac{51}{40} = 1.275$. Hence, $\alpha = 2.975$.

Multinomial and conjugate prior

We have seen that the discrete **binomial random variable** or **likelihood**

$$X \sim \text{Bin}(n, \theta) = p_{X|\Theta}(k|\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad k = 0, 1, \dots, n$$

has a conjugate prior given by the continuous **beta distribution**

$$f_{\Theta}(\theta) \sim \text{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} \quad 0 < \theta < 1$$

which can encode a range of possibilities of prior information, and make it relatively easy to calculate the posterior. Together, we call this the **beta-binomial model**. We can generalize to cases with multiple outcomes (i.e., from a coin to a die).

Multinomial distribution

Generalizing the binomial distribution, for each event with K possible outcomes, we let

$$\mathbf{x} = (x_1, \dots, x_K) \quad ; \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$$

where x_j denotes the number of times event j has occurred, with probability θ_j .

Multinomial distribution

Generalizing the binomial distribution, for each event with K possible outcomes, we let

$$\mathbf{x} = (x_1, \dots, x_K) \quad ; \quad \boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$$

where x_j denotes the number of times event j has occurred, with probability θ_j .

The *random vector* $\mathbf{X}$ follows a **multinomial distribution**

$$\mathbf{X} \sim \text{Mult}(n, \boldsymbol{\theta}) = \binom{n}{x_1, \dots, x_K} \prod_{j=1}^K \theta_j^{x_j} \quad (5)$$

where $\sum_{k=1}^K x_k = n$, and $\binom{n}{x_1, \dots, x_K} = \frac{n!}{x_1! \cdots x_K!}$ are partitions.

Multinomial distribution

When $n = 1$, binomial becomes Bernoulli.

Similarly, when $n = 1$, multinomial becomes a “categorical” distribution: random vector $\mathbf{x}$ denotes a *one-hot encoding* of event j that has occurred, with

$$\mathbf{x} = (0, \dots, \underbrace{1}_{j\text{th}}, \dots, 0)$$

This is very useful for classification problems.

Dirichlet distribution

Generalizing the beta distribution, we have the **Dirichlet distribution**, defined over the *probability simplex*

$$S_K = \left\{ \theta : 0 \leq \theta_k \leq 1, \sum_{k=1}^K \theta_k = 1 \right\}$$

such that

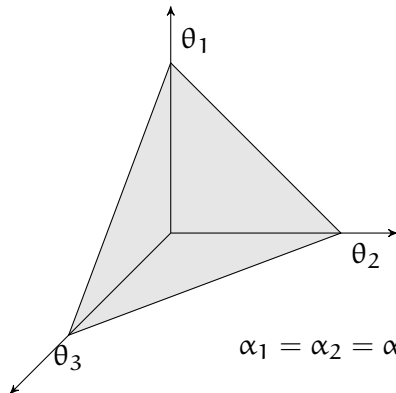
$$\text{Dir}(\theta; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_k^{\alpha_k - 1} \quad (6)$$

with $\alpha_0 = \sum_{k=1}^K \alpha_k$, and $B(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\alpha_0)}$.

Dirichlet distribution

Remember for $K = 2$ (beta distribution), $\alpha_1 = \alpha_2 = 1$ gives a uniform distribution.

For $K = 3$, The probability simplex is the surface:



$\alpha_1 = \alpha_2 = \alpha_3 = 1$ is a uniform distribution

Dirichlet-multinomial model

Using the Dirichlet as prior, and multinomial as the likelihood, the posterior is given by

$$\begin{aligned}
 f_{\Theta|X}(\theta | x_1, \dots, x_K) &\propto f_{\Theta}(\theta) p_{X|\Theta}(x_1, \dots, x_K | \theta) \\
 &\propto \left(\prod_{k=1}^K \theta_k^{\alpha_k - 1} \right) \left(\prod_{j=1}^K \theta_j^{x_j} \right) \\
 &= \prod_{k=1}^K \theta_k^{\alpha_k + x_k - 1} \\
 &\propto \text{Dir}(\theta; \alpha_1 + x_1, \dots, \alpha_K + x_K)
 \end{aligned}$$

Dirichlet-multinomial model

We compute the MAP estimate by maximizing the (log) posterior.

$$\begin{aligned} &\text{maximize} && \log \prod_{k=1}^K \theta_k^{\alpha_k + x_k - 1} \\ &\text{subject to} && \sum_{k=1}^K \theta_k = 1 \end{aligned}$$

We solve by using $\lambda =$ Lagrange multiplier,

$$\begin{aligned} L &= \sum_{k=1}^K (\alpha_k + x_k - 1) \log \theta_k + \lambda \sum_{k=1}^K \theta_k \\ \frac{\partial L}{\partial \theta_k} &= 0 \quad \Rightarrow \quad \theta_k = -\frac{\alpha_k + x_k - 1}{\lambda} \end{aligned}$$

Dirichlet-multinomial model

We next solve for λ by summing over k :

$$1 = \sum_{k=1}^K \theta_k = -\frac{1}{\lambda} \sum_{k=1}^K (\alpha_k + x_k - 1) = -\frac{1}{\lambda} (\alpha_0 + n - K)$$

Thus, $\lambda = -(\alpha_0 + n - K)$. Hence, the MAP estimate is

$$\hat{\Theta}_k = \frac{\alpha_k + x_k - 1}{\alpha_0 + n - K}$$

Dirichlet-multinomial model

We next solve for λ by summing over k :

$$1 = \sum_{k=1}^K \theta_k = -\frac{1}{\lambda} \sum_{k=1}^K (\alpha_k + x_k - 1) = -\frac{1}{\lambda} (\alpha_0 + n - K)$$

Thus, $\lambda = -(\alpha_0 + n - K)$. Hence, the MAP estimate is

$$\hat{\Theta}_k = \frac{\alpha_k + x_k - 1}{\alpha_0 + n - K}$$

In case of uninformative prior (uniform distribution), $\alpha_k = 1$,

$$\hat{\Theta}_k = \frac{x_k}{n}$$

i.e., the empirical fraction of time event k occurs.

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior
- 6 Sequential learning with MAP**
- 7 Hypothesis testing

Recursive learning

The Bayesian estimation technique can be adapted to a successive refinement procedure *as data come in*. This is known as **recursive learning** or **online learning**.

Example: Suppose we believe our data come from a uniform distribution

$$f_X(x) \sim U(0, \theta)$$

but θ is something you need to estimate, based on successive appearance of new data. Before seeing any data, we can only assume that $0 \leq \theta \leq 10$. The data that we subsequently observe are, in order, $\{4, 7, 2, 8\}$. Show how our understanding of θ changes.

Recursive learning

ANS: Let the data be labeled $x_1 = 4$, $x_2 = 7$, $x_3 = 2$, $x_4 = 8$. We make use of Bayes' rule (and ignore normalization throughout):

$$f_{\Theta|X_1}(\theta|x_1) \propto f_{X_1|\Theta}(x_1|\theta) f_{\Theta}(\theta)$$

Recursive learning

ANS: Let the data be labeled $x_1 = 4$, $x_2 = 7$, $x_3 = 2$, $x_4 = 8$. We make use of Bayes' rule (and ignore normalization throughout):

$$f_{\Theta|X_1}(\theta|x_1) \propto f_{X_1|\Theta}(x_1|\theta) f_{\Theta}(\theta)$$

- Since $0 \leq \theta \leq 10$, without further information, we assume it is a uniform distribution. Hence,

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{10} & 0 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Recursive learning

ANS: Let the data be labeled $x_1 = 4$, $x_2 = 7$, $x_3 = 2$, $x_4 = 8$. We make use of Bayes' rule (and ignore normalization throughout):

$$f_{\Theta|X_1}(\theta|x_1) \propto f_{X_1|\Theta}(x_1|\theta) f_{\Theta}(\theta)$$

- Since $0 \leq \theta \leq 10$, without further information, we assume it is a uniform distribution. Hence,

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{10} & 0 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

- $f_{X_1|\Theta}(x_1|\theta) = 1/\theta$, *provided* $x_1 < \theta$.

Recursive learning

ANS: Let the data be labeled $x_1 = 4$, $x_2 = 7$, $x_3 = 2$, $x_4 = 8$. We make use of Bayes' rule (and ignore normalization throughout):

$$f_{\Theta|X_1}(\theta|x_1) \propto f_{X_1|\Theta}(x_1|\theta) f_{\Theta}(\theta)$$

- Since $0 \leq \theta \leq 10$, without further information, we assume it is a uniform distribution. Hence,

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{10} & 0 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

- $f_{X_1|\Theta}(x_1|\theta) = 1/\theta$, *provided* $x_1 < \theta$.
- Putting the two together, with $x_1 = 4$,

$$f_{\Theta|X_1}(\theta|x_1) \propto \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Recursive learning

We apply Bayes' rule again, noting that X_1 and X_2 are conditionally independent given Θ :

$$\begin{aligned} f_{\Theta|X_1,X_2}(\theta|x_1,x_2) &\propto f_{X_1,X_2|\Theta}(x_1,x_2|\theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2|\theta) f_{X_1|\Theta}(x_1|\theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2|\theta) f_{\Theta|X_1}(\theta|x_1) \end{aligned}$$

Recursive learning

We apply Bayes' rule again, noting that X_1 and X_2 are conditionally independent given Θ :

$$\begin{aligned} f_{\Theta|X_1, X_2}(\theta | x_1, x_2) &\propto f_{X_1, X_2|\Theta}(x_1, x_2 | \theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2 | \theta) f_{X_1|\Theta}(x_1 | \theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2 | \theta) f_{\Theta|X_1}(\theta | x_1) \end{aligned}$$

- $f_{X_2|\Theta}(x_2 | \theta) = 1/\theta$, *provided* $x_2 < \theta$.

Recursive learning

We apply Bayes' rule again, noting that X_1 and X_2 are conditionally independent given Θ :

$$\begin{aligned} f_{\Theta|X_1, X_2}(\theta | x_1, x_2) &\propto f_{X_1, X_2|\Theta}(x_1, x_2 | \theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2 | \theta) f_{X_1|\Theta}(x_1 | \theta) f_{\Theta}(\theta) \\ &= f_{X_2|\Theta}(x_2 | \theta) f_{\Theta|X_1}(\theta | x_1) \end{aligned}$$

- $f_{X_2|\Theta}(x_2 | \theta) = 1/\theta$, *provided* $x_2 < \theta$.
- Putting it together with our previous result, with $x_2 = 7$,

$$f_{\Theta|X_1, X_2}(\theta | x_1, x_2) \propto \begin{cases} 1/\theta^2 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Recursive learning

We apply Bayes' rule once again,

$$f_{\Theta|X_1, X_2, X_3}(\theta|x_1, x_2, x_3) \propto f_{X_3|\Theta}(x_3|\theta) f_{\Theta|X_1, X_2}(\theta|x_1, x_2)$$

Recursive learning

We apply Bayes' rule once again,

$$f_{\Theta|X_1, X_2, X_3}(\theta | x_1, x_2, x_3) \propto f_{X_3|\Theta}(x_3 | \theta) f_{\Theta|X_1, X_2}(\theta | x_1, x_2)$$

With $x_3 = 2$, we do not shrink the range of θ , and therefore

$$f_{\Theta|X_1, X_2, X_3}(\theta | x_1, x_2, x_3) \propto \begin{cases} 1/\theta^3 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Recursive learning

We apply Bayes' rule once again,

$$f_{\Theta|X_1, X_2, X_3}(\theta | x_1, x_2, x_3) \propto f_{X_3|\Theta}(x_3 | \theta) f_{\Theta|X_1, X_2}(\theta | x_1, x_2)$$

With $x_3 = 2$, we do not shrink the range of θ , and therefore

$$f_{\Theta|X_1, X_2, X_3}(\theta | x_1, x_2, x_3) \propto \begin{cases} 1/\theta^3 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

We apply Bayes' rule for one more time,

$$f_{\Theta|X_1, X_2, X_3, X_4}(\theta | x_1, x_2, x_3, x_4) \propto f_{X_4|\Theta}(x_4 | \theta) f_{\Theta|X_1, X_2, X_3}(\theta | x_1, x_2, x_3)$$

Recursive learning

We apply Bayes' rule once again,

$$f_{\Theta | x_1, x_2, x_3}(\theta | x_1, x_2, x_3) \propto f_{x_3 | \Theta}(x_3 | \theta) f_{\Theta | x_1, x_2}(\theta | x_1, x_2)$$

With $x_3 = 2$, we do not shrink the range of θ , and therefore

$$f_{\Theta | x_1, x_2, x_3}(\theta | x_1, x_2, x_3) \propto \begin{cases} 1/\theta^3 & 7 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

We apply Bayes' rule for one more time,

$$f_{\Theta | x_1, x_2, x_3, x_4}(\theta | x_1, x_2, x_3, x_4) \propto f_{x_4 | \Theta}(x_4 | \theta) f_{\Theta | x_1, x_2, x_3}(\theta | x_1, x_2, x_3)$$

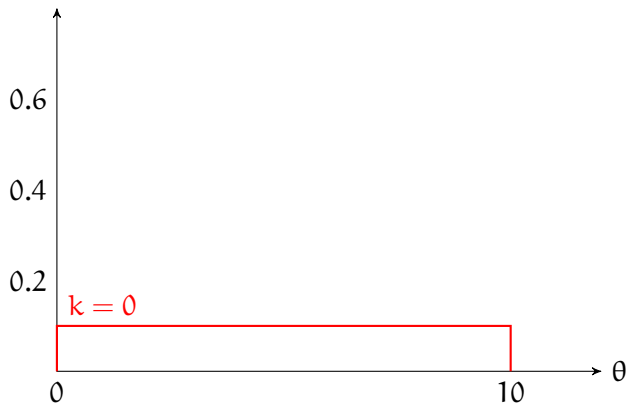
With $x_4 = 8$, we have

$$f_{\Theta | x_1, x_2, x_3, x_4}(\theta | x_1, x_2, x_3, x_4) \propto \begin{cases} 1/\theta^4 & 8 \leq \theta \leq 10 \\ 0 & \text{otherwise.} \end{cases}$$

Recursive learning

Visualizing the results in plots:

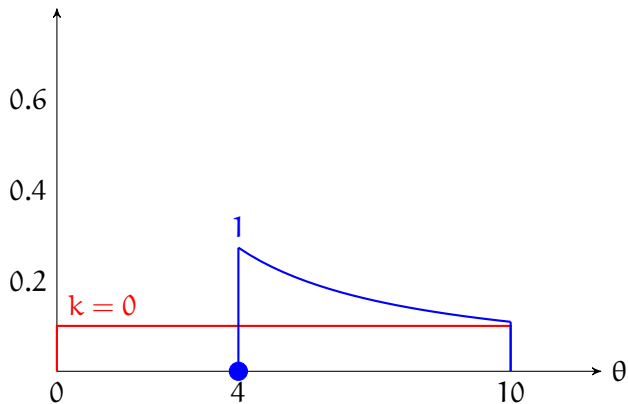
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

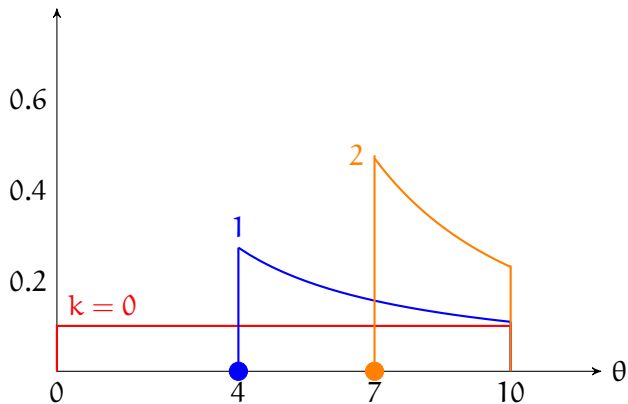
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

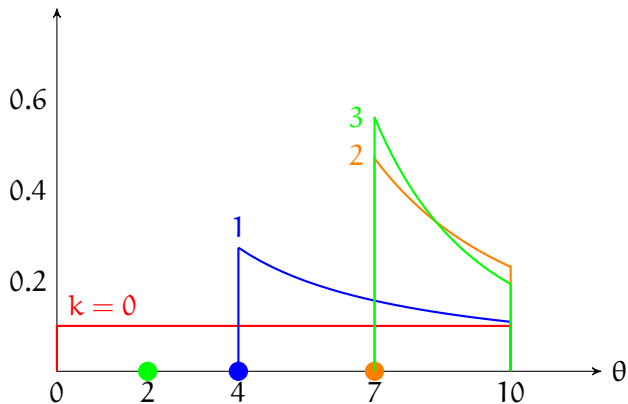
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

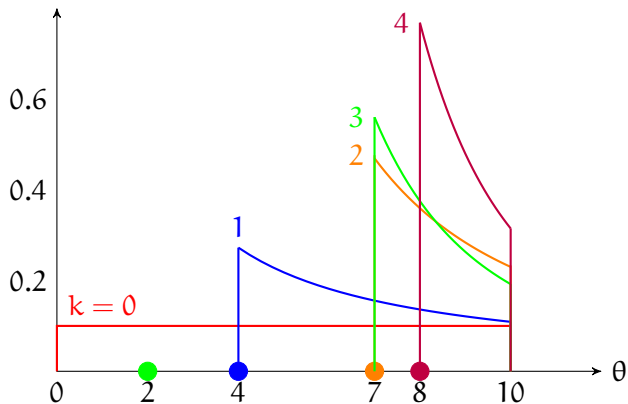
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

MAP rule is to maximize $f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$.

- With only x_1 , the maximum is at $\theta = 4$.
- With x_1 and x_2 , the maximum is at $\theta = 7$.
- With x_1 to x_3 , the maximum is at $\theta = 7$.
- With x_1 to x_4 , the maximum is at $\theta = 8$.
- Every additional data point changes our belief about the random variable Θ .

- 1 Maximum a posteriori (MAP)
- 2 MAP estimation examples
- 3 Comparing MAP and ML estimations
- 4 Least mean square (LMS) estimate
- 5 Conjugate prior
- 6 Sequential learning with MAP
- 7 Hypothesis testing**

Hypothesis testing as an estimation problem

Binary hypothesis testing involves deciding between two “hypotheses”, e.g., $\mathbb{P}(H) = 0.5$ versus $\mathbb{P}(H) = 0.1$ for a particular coin, based on observations of outcomes.

We can look at this as a special case of parameter estimation, where θ takes on two possible values. Both ML and MAP can be used. Nevertheless, we'll introduce the specific terminologies relating to hypothesis testing.

Hypothesis testing terminology

We use H_0 and H_1 to denote the two hypotheses.

H_0 : null hypothesis

H_1 : alternative hypothesis

$X = (X_1, \dots, X_n)$ is a vector of random variables whose distribution depends on the hypothesis. Certain set (or **acceptance region**) of X will lead us to decide on H_0 ; outside that set (or **rejection region**), we will conclude on H_1 .

We are interested to find a rule to determine where the regions are.

Hypothesis testing terminology

We can have two types of errors:

Type I (false rejection): reject H_0 even though it is true

Type II (false acceptance): accept H_0 even though it is false

Type I error means X lies in the rejection region R , even when the reality should follow H_0 . This occurs with probability

$$\alpha(R) = \mathbb{P}(X \in R; H_0) \quad (7)$$

Type II error means X does not lie in the rejection region R , even when the reality should follow H_1 . This occurs with probability

$$\beta(R) = \mathbb{P}(X \notin R; H_1) \quad (8)$$

We often have to tradeoff between the two types of errors.

Hypothesis testing and ML

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)} \quad \text{or} \quad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} \quad (9)$$

Hypothesis testing and ML

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_X(x; H_1)}{p_X(x; H_0)} \quad \text{or} \quad L(x) = \frac{f_X(x; H_1)}{f_X(x; H_0)} \quad (9)$$

ML says we pick H_1 if $L(x) > 1$, and H_0 otherwise.

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)} \quad \text{or} \quad L(x) = \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_0)} \quad (10)$$

Hypothesis testing and MAP

Let H_0 means $\Theta = \theta_0$ with prior probability $p_{\Theta}(\theta_0)$, and H_1 means $\Theta = \theta_1$ with prior probability $p_{\Theta}(\theta_1)$. The MAP rule for $\Theta = \theta_1$ is

$$p_{\Theta}(\theta_0) p_{X|\Theta}(x|\theta_0) < p_{\Theta}(\theta_1) p_{X|\Theta}(x|\theta_1)$$

We define the **likelihood ratio** $L(x)$ for both discrete and continuous r.v.:

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)} \quad \text{or} \quad L(x) = \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_0)} \quad (10)$$

MAP says we pick H_1 if $L(x) > \zeta$, and H_0 otherwise, with

$$\zeta = \frac{p_{\Theta}(\theta_0)}{p_{\Theta}(\theta_1)} \quad (11)$$

Example

We have a die and have two hypotheses about its fairness:

$$H_0(\text{fair die}) : \quad p_X(x; H_0) = \frac{1}{6} \quad x = 1, \dots, 6$$

$$H_1(\text{loaded die}) : \quad p_X(x; H_1) = \begin{cases} \frac{1}{4} & x = 1, 2 \\ \frac{1}{8} & x = 3, 4, 5, 6 \end{cases}$$

Example

We have a die and have two hypotheses about its fairness:

$$H_0(\text{fair die}) : \quad p_X(x; H_0) = \frac{1}{6} \quad x = 1, \dots, 6$$

$$H_1(\text{loaded die}) : \quad p_X(x; H_1) = \begin{cases} \frac{1}{4} & x = 1, 2 \\ \frac{1}{8} & x = 3, 4, 5, 6 \end{cases}$$

$$\textcircled{1} \text{ For } x = 1, 2: L(x) = \frac{1/4}{1/6} = \frac{3}{2}$$

$$\textcircled{2} \text{ For } x = 3, 4, 5, 6: L(x) = \frac{1/8}{1/6} = \frac{3}{4}$$

If we use ML, observing 1 or 2 will lead us to H_1 , and observing 3 to 6 will lead us to H_0 .

Example

If we use “MAP”, we are more flexible to set ζ .

- If we set $\zeta < \frac{3}{4}$: we will reject H_0 for all x .

$$\alpha(\zeta) = 1 \quad \beta(\zeta) = 0$$

- If we set $\zeta > \frac{3}{2}$: we will accept H_0 for all x .

$$\alpha(\zeta) = 0 \quad \beta(\zeta) = 1$$

- If we set $\frac{3}{4} < \zeta < \frac{3}{2}$: we will accept H_0 for $x = 3, 4, 5, 6$ and reject it for $x = 1, 2$.

$$\alpha(\zeta) = \mathbb{P}(X = 1, 2; H_0) = \frac{1}{3}$$

$$\beta(\zeta) = \mathbb{P}(X = 3, 4, 5, 6; H_1) = \frac{1}{2}$$

Tradeoff

Choosing ζ represents a tradeoff in Type I and Type II errors. In fact, using the likelihood ratio as a means of deciding on the acceptance and rejection regions, reducing one type of error leads to an increase in another, given by the **Neyman-Pearson Lemma**.

Consider using the likelihood ratio $L(X)$ for testing, with error probabilities

$$\mathbb{P}(L(X) > \zeta; H_0) = \alpha \quad \text{and} \quad \mathbb{P}(L(X) \leq \zeta; H_1) = \beta$$

Suppose some other test results in a rejection region R achieving a smaller (or equal) false rejection probability $\mathbb{P}(X \in R; H_0) \leq \alpha$, then we must have

$$\mathbb{P}(X \notin R; H_1) \geq \beta$$

with strict inequality $\mathbb{P}(X \notin R; H_1) > \beta$ when $\mathbb{P}(X \in R; H_0) < \alpha$.

Conclusions

Parameter estimation based on observations has increasingly widespread use, especially because we are in a *data-driven* era.

Conclusions

Pierre-Simon Laplace (1749–1827) once said:

“Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.”

Conclusions

Pierre-Simon Laplace (1749–1827) once said:

“Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.”

Pierre-Simon Laplace



The most important questions of life are indeed, for the most part, really only problems of probability.

Maybe this is true on 3 June 2020?