

Parameter Estimation

Dr. Edmund Lam

Department of Electrical and Electronic Engineering
The University of Hong Kong

ELEC2844: Probabilistic Systems Analysis
(Second Semester, 2018–19)

<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 find p is perhaps to experiment:

- What is your estimate of p if we flip 1000 times and you get 500 heads?
- What is your estimate of p if we flip 999 times and you get 500 heads?
- What is your estimate of p if we flip once and you get head?
- What is your estimate of p 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 make estimate on probabilistic quantities based on observations. We begin with the basic *language* and *context* of this class of problems.

1 Estimator and estimation error

2 Maximum likelihood (ML)

3 Linear regression

- ML with Gaussian noise (2D)
- ML with Gaussian noise (higher dimensions)
- ML with Laplacian noise (higher dimensions)

4 Information theory and ML

5 Maximum a posteriori (MAP)

- Incorporating a prior
- Conjugate prior
- Sequential learning

Definition

The “classical” method of (statistical) parameter estimation goes as follow.

We want to estimate an unknown constant θ ,
based on observations $X = (X_1, \dots, X_n)$ that depend on this parameter.

Definition

The “classical” method of (statistical) parameter estimation goes as follow.

We want to **estimate an unknown constant θ** ,
based on **observations $X = (X_1, \dots, X_n)$** that depend on this parameter.

$$\text{Estimator: } \hat{\Theta} = g(X) \quad (1)$$

Definition

The “classical” method of (statistical) parameter estimation goes as follow.

We want to **estimate an unknown constant θ** , based on **observations $X = (X_1, \dots, X_n)$** that depend on this parameter.

$$\text{Estimator: } \hat{\Theta} = g(X) \quad (1)$$

- X depends on θ , so $\hat{\Theta}$ depends on θ
- X is a random variable, so $\hat{\Theta}$ is also a random variable
 - We can ask for $\mathbb{E}(\hat{\Theta})$ and $\text{var}(\hat{\Theta})$, for example
- The actual realized value of $\hat{\Theta}$ is the **estimate**

Definition

Often, the number of individual observations n has a significant influence on the “goodness” of the estimator, so we emphasize that in the notation:

$$\text{Estimator: } \hat{\Theta}_n = g(X) \quad (2)$$

Definition

Often, the number of individual observations n has a significant influence on the “goodness” of the estimator, so we emphasize that in the notation:

$$\text{Estimator: } \hat{\Theta}_n = g(X) \quad (2)$$

For different values of n :

- We can view $\hat{\Theta}_n$ as *a sequence of estimators*
- We can also have a sequence of $\mathbb{E}(\hat{\Theta}_n)$ and $\text{var}(\hat{\Theta}_n)$

Definition

Estimates are not 100% accurate:

$$\text{Estimation error: } \tilde{\Theta}_n = \hat{\Theta}_n - \theta \quad (3)$$

The estimation error depends on θ and observations X .

Definition

Estimates are not 100% accurate:

$$\text{Estimation error: } \tilde{\Theta}_n = \hat{\Theta}_n - \theta \quad (3)$$

The estimation error depends on θ and observations X .

Bias is the expected value of the estimation error:

$$b(\tilde{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \theta \quad (4)$$

Expected value, variance, and bias depend on θ .

Definition

Estimates are not 100% accurate:

$$\text{Estimation error: } \tilde{\Theta}_n = \hat{\Theta}_n - \theta \quad (3)$$

The estimation error depends on θ and observations X .

Bias is the expected value of the estimation error:

$$b(\tilde{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \theta \quad (4)$$

Expected value, variance, and bias depend on θ .

We have an unbiased estimator if $b(\tilde{\Theta}_n) = 0$ for all possible values of θ .

Definition

The general question:

*How do we **design the function** $g(\cdot)$, such that the estimator $\hat{\Theta}_n = g(X)$ is “good” in some sense? (e.g. small estimation error, unbiased)*

The classical approach is something called **maximum likelihood**.

- 1 Estimator and estimation error
- 2 Maximum likelihood (ML)**
- 3 Linear regression
 - ML with Gaussian noise (2D)
 - ML with Gaussian noise (higher dimensions)
 - ML with Laplacian noise (higher dimensions)
- 4 Information theory and ML
- 5 Maximum a posteriori (MAP)
 - Incorporating a prior
 - Conjugate prior
 - Sequential learning

Definition

Depending on an unknown parameter θ , the observations $X = (X_1, \dots, X_n)$ can be described by a joint PMF / PDF:

$$\begin{array}{ll} \text{discrete:} & p_X(x_1, \dots, x_n; \theta) \\ \text{continuous:} & f_X(x_1, \dots, x_n; \theta) \end{array}$$

We call these the **likelihood function**.

Interpretation: $p_X(x; \theta)$ or $f_X(x; \theta)$ is the probability of the observed value x when the parameter is θ , but **not** the probability that the unknown parameter is θ .

Main idea

*What is the value of θ such that the observations $\mathbf{x} = x_1, \dots, x_n$ are **most likely** to appear?*

Main idea

*What is the value of θ such that the observations $x = x_1, \dots, x_n$ are **most likely** to appear?*

ANS: We find the θ that maximizes the likelihood function, hence the method is known as **maximum likelihood (ML)**.

$$\text{discrete:} \quad \hat{\theta}_n = \arg \max_{\theta} p_X(x_1, \dots, x_n; \theta) \quad (5)$$

$$\text{continuous:} \quad \hat{\theta}_n = \arg \max_{\theta} f_X(x_1, \dots, x_n; \theta) \quad (6)$$

Note: The notation $\arg \max$ means we want to maximize the quantity, but we don't quite care the value of that expression, but are interested in the argument that leads to such maximum.

Main idea

Often, the observations X_i are assumed to be independent, and perhaps even identically distributed. The likelihood function is then

$$\text{discrete:} \quad p_X(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p_{X_i}(x_i; \theta)$$

$$\text{continuous:} \quad f_X(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$$

Main idea

Often, the observations X_i are assumed to be independent, and perhaps even identically distributed. The likelihood function is then

$$\text{discrete:} \quad p_X(x_1, \dots, x_n; \theta) = \prod_{i=1}^n p_{X_i}(x_i; \theta)$$

$$\text{continuous:} \quad f_X(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta)$$

Computationally, for ML, we often maximize the **log-likelihood**:

$$\text{discrete:} \quad \log p_X(x; \theta) = \log \prod_{i=1}^n p_{X_i}(x_i; \theta) = \sum_{i=1}^n \log p_{X_i}(x_i; \theta)$$

$$\text{continuous:} \quad \log f_X(x; \theta) = \log \prod_{i=1}^n f_{X_i}(x_i; \theta) = \sum_{i=1}^n \log f_{X_i}(x_i; \theta)$$

Example

Determine the ML estimator of the parameter of a Poisson random variable based on independent identically distributed observations $X_1, \dots, X_n$. Is the estimator unbiased?

Example

Determine the ML estimator of the parameter of a Poisson random variable based on independent identically distributed observations $X_1, \dots, X_n$. Is the estimator unbiased?

ANS: Let us proceed step-by-step.

- 1 Let the unknown Poisson parameter be θ .

Example

Determine the ML estimator of the parameter of a Poisson random variable based on independent identically distributed observations $X_1, \dots, X_n$. Is the estimator unbiased?

ANS: Let us proceed step-by-step.

① Let the unknown Poisson parameter be θ .

② The PMF of X_i is $p_{X_i}(x_i) = \frac{e^{-\theta} \theta^{x_i}}{x_i!} \quad x_i = 0, 1, \dots$

Example

Determine the ML estimator of the parameter of a Poisson random variable based on independent identically distributed observations $X_1, \dots, X_n$. Is the estimator unbiased?

ANS: Let us proceed step-by-step.

- 1 Let the unknown Poisson parameter be θ .
- 2 The PMF of X_i is $p_{X_i}(x_i) = \frac{e^{-\theta} \theta^{x_i}}{x_i!} \quad x_i = 0, 1, \dots$
- 3 The log-likelihood function is

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

Example

- ④ We maximize it by setting the derivative (w.r.t. θ) to 0:

$$0 = -n + \frac{1}{\theta} \sum_{i=1}^n x_i$$

so the estimator is $\hat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$

Example

- 4 We maximize it by setting the derivative (w.r.t. θ) to 0:

$$0 = -n + \frac{1}{\theta} \sum_{i=1}^n x_i$$

so the estimator is $\hat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$

- 5 Since $\mathbb{E}(X_i) = \theta$, we have

$$\mathbb{E}(\hat{\Theta}_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \theta$$

so the estimator is unbiased.

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, being uniform over $[0, \theta]$, where θ is unknown. Find the ML estimator of θ .

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, being uniform over $[0, \theta]$, where θ is unknown. Find the ML estimator of θ .

ANS:

① The PDF of X_i is $f_{X_i}(x_i) = \begin{cases} 1/\theta & \text{if } 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, being uniform over $[0, \theta]$, where θ is unknown. Find the ML estimator of θ .

ANS:

① The PDF of X_i is $f_{X_i}(x_i) = \begin{cases} 1/\theta & \text{if } 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$

② The likelihood function is

$$\begin{aligned} f_X(x_1, \dots, x_n; \theta) &= f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta) \\ &= \begin{cases} 1/\theta^n & \text{if } 0 \leq \max_{i=1, \dots, n} x_i \leq \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, being uniform over $[0, \theta]$, where θ is unknown. Find the ML estimator of θ .

ANS:

① The PDF of X_i is $f_{X_i}(x_i) = \begin{cases} 1/\theta & \text{if } 0 \leq x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$

② The likelihood function is

$$\begin{aligned} f_X(x_1, \dots, x_n; \theta) &= f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta) \\ &= \begin{cases} 1/\theta^n & \text{if } 0 \leq \max_{i=1, \dots, n} x_i \leq \theta \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

③ We maximize the likelihood function by noting that we need θ to be as small as possible, hence $\hat{\Theta}_n = \max_{i=1, \dots, n} X_i$

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, following the (mean-shifted) Laplacian PDF $f_{X_i}(x_i) = \frac{1}{2}e^{-|x_i - \theta|}$, where θ is unknown. Find the ML estimator of θ .

Example

Observations $X_1, \dots, X_n$ are independent identically distributed, following the (mean-shifted) Laplacian PDF $f_{X_i}(x_i) = \frac{1}{2}e^{-|x_i - \theta|}$, where θ is unknown. Find the ML estimator of θ .

ANS: From the PDF, we can find the log-likelihood function

$$\begin{aligned} \log f_X(x; \theta) &= \sum_{i=1}^n \log \frac{1}{2} e^{-|x_i - \theta|} \\ &= n \log \frac{1}{2} - \sum_{i=1}^n |x_i - \theta| \end{aligned}$$

This is maximized when the second term is minimized, which occurs when

$$\hat{\Theta}_n = \text{median}(X_1, \dots, X_n)$$

Example

A coin has an unknown probability θ to result in head.

- (a) We toss the coin n times, where random variable K denotes the resulting number of heads. Find the ML estimator of θ based on K .
- (b) We fix k , and let N be a random variable denoting the number of tosses until the k th head occurs. Find the ML estimator of θ based on N .

Example

ANS: (a) The PMF of K is binomial:

$$p_K(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad k = 0, 1, 2, \dots, n$$

This is also the likelihood function (to get our observation $K = k$).

Example

ANS: (a) The PMF of K is binomial:

$$p_K(k; \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \quad k = 0, 1, 2, \dots, n$$

This is also the likelihood function (to get our observation $K = k$).

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

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

which yields the ML estimator $\hat{\Theta} = \frac{K}{n}$.

This is the fraction of heads in the observed trials.

Example

ANS: (b) We need to first find the PMF of N . The k th head occurs at $N = n$ toss means that exactly $(k - 1)$ heads have occurred among the first $(n - 1)$ tosses, and that the n th toss is a head. Therefore,

$$p_N(n; \theta) = \binom{n-1}{k-1} \theta^{k-1} (1-\theta)^{n-k} \cdot \theta = \binom{n-1}{k-1} \theta^k (1-\theta)^{n-k}$$

for $n = k, k + 1, \dots$

Example

ANS: (b) We need to first find the PMF of N . The k th head occurs at $N = n$ toss means that exactly $(k - 1)$ heads have occurred among the first $(n - 1)$ tosses, and that the n th toss is a head. Therefore,

$$p_N(n; \theta) = \binom{n-1}{k-1} \theta^{k-1} (1-\theta)^{n-k} \cdot \theta = \binom{n-1}{k-1} \theta^k (1-\theta)^{n-k}$$

for $n = k, k + 1, \dots$

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

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

which yields the ML estimator $\hat{\Theta} = \frac{k}{N}$.

This is again the fraction of heads in the observed trials.

1 Estimator and estimation error

2 Maximum likelihood (ML)

3 Linear regression

- ML with Gaussian noise (2D)
- ML with Gaussian noise (higher dimensions)
- ML with Laplacian noise (higher dimensions)

4 Information theory and ML

5 Maximum a posteriori (MAP)

- Incorporating a prior
- Conjugate prior
- Sequential learning

Setting

We consider the following scenario:

- We have two variables of interest, x and y , with a collection of data points, (x_i, y_i) for $i = 1, \dots, n$.
- We *assume* a linear relationship between them,

$$y \approx \theta_0 + \theta_1 x$$

where θ_0 and θ_1 are *unknown parameters to be estimated*.

- We approach the problem by assuming that each y_i is a realization of a random variable Y_i , generated according to

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

where ϵ_i 's are independent, identically distributed Gaussian random variables with mean zero and variance σ^2 .

- We compute the ML estimates of θ_0 and θ_1 .

Computing the ML estimates

- 1 Given θ_0 and θ_1 , we have $Y_i = \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$

Computing the ML estimates

- 1 Given θ_0 and θ_1 , we have $Y_i = \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$
- 2 Likelihood function is therefore

$$f_Y(y; \theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right\}$$

Computing the ML estimates

- ① Given θ_0 and θ_1 , we have $Y_i = \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$
- ② Likelihood function is therefore

$$f_Y(y; \theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right\}$$

- ③ Log-likelihood function is

$$\log f_Y(y; \theta_0, \theta_1) = \sum_{i=1}^n \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right)$$

Computing the ML estimates

- 1 Given θ_0 and θ_1 , we have $Y_i = \mathcal{N}(\theta_0 + \theta_1 x_i, \sigma^2)$
- 2 Likelihood function is therefore

$$f_Y(y; \theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right\}$$

- 3 Log-likelihood function is

$$\log f_Y(y; \theta_0, \theta_1) = \sum_{i=1}^n \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2} \right)$$

- 4 Maximizing the log-likelihood is equivalent to

$$\text{minimize} \quad \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2 \quad (7)$$

Computing the ML estimates

5 Let $S = \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_i)^2$.

$$\frac{\partial S}{\partial \theta_0} = \sum_{i=1}^n 2(y_i - \theta_0 - \theta_1 x_i)(-1) = 0$$

$$\sum_{i=1}^n y_i - n\theta_0 - \theta_1 \sum_{i=1}^n x_i = 0$$

$$\theta_0 = \frac{1}{n} \sum_{i=1}^n y_i - \theta_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i$$

By defining $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$,

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \quad (8)$$

Computing the ML estimates

$$\begin{aligned}\frac{\partial S}{\partial \theta_1} &= \sum_{i=1}^n 2(y_i - \theta_0 - \theta_1 x_i)(-x_i) = 0 \\ \sum_{i=1}^n x_i y_i - \theta_0 \sum_{i=1}^n x_i - \theta_1 \sum_{i=1}^n x_i^2 &= 0 \\ \sum_{i=1}^n x_i y_i - (\bar{y} - \theta_1 \bar{x}) \sum_{i=1}^n x_i - \theta_1 \sum_{i=1}^n x_i^2 &= 0 \\ \theta_1 &= \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}\end{aligned}$$

A more “aesthetic” formula is

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (9)$$

1 Estimator and estimation error

2 Maximum likelihood (ML)

3 Linear regression

- ML with Gaussian noise (2D)
- ML with Gaussian noise (higher dimensions)
- ML with Laplacian noise (higher dimensions)

4 Information theory and ML

5 Maximum a posteriori (MAP)

- Incorporating a prior
- Conjugate prior
- Sequential learning

Higher dimensions

We can generalize the above to a linear regression of the form

$$y \approx w_0 x_0 + w_1 x_1 + \dots + w_D x_D$$

where we can assume $x_0 = 1$ to take care of the constant, and $x_1, \dots, x_D$ are D “features”. Sometimes, we simply ignore the $w_0 x_0$ term in the derivations.

This is extremely useful when we want to estimate the weights w_i for each feature to match our observation y , so that in the future, when we have a new set of features, we can generate a new y easily.

Linear regression in higher dimensions

For linear regression, we approach the problem by assuming that each y_i is a realization of a random variable Y_i , generated according to

$$y_i = \theta^T x_i + \epsilon_i \quad i = 1, \dots, n$$

where ϵ_i 's are independent, identically distributed Gaussian random variables with mean zero and variance σ^2 . Here, θ is a length-D vector of parameters to estimate; x_i is a length-D vector of features.

We therefore aim to compute the ML estimates of θ .

Linear regression in higher dimensions

- 1 Given the vector θ , we have $Y_i = \mathcal{N}(\theta^T x_i, \sigma^2)$

Linear regression in higher dimensions

- ① Given the vector θ , we have $Y_i = \mathcal{N}(\theta^T \mathbf{x}_i, \sigma^2)$
- ② Likelihood function is therefore

$$f_Y(\mathbf{y}; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_i - \theta^T \mathbf{x}_i)^2}{2\sigma^2} \right\}$$

Linear regression in higher dimensions

- ① Given the vector θ , we have $Y_i = \mathcal{N}(\theta^T x_i, \sigma^2)$
- ② Likelihood function is therefore

$$f_Y(y; \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right\}$$

- ③ Log-likelihood function is

$$\begin{aligned} \log f_Y(y; \theta) &= \sum_{i=1}^n \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - \theta^T x_i)^2}{2\sigma^2} \right) \\ &= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} (X\theta - y)^T (X\theta - y) \end{aligned}$$

where y is a length- n vector containing all y_i , and X is a $n \times D$ matrix containing all x_i .

Linear regression in higher dimensions

Maximizing the log-likelihood is equivalent to

$$\text{minimize } (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

which has an analytic solution

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (10)$$

This is also known as the **minimum mean square error (MMSE)** or ℓ_2 -norm solution.

1 Estimator and estimation error

2 Maximum likelihood (ML)

3 **Linear regression**

- ML with Gaussian noise (2D)
- ML with Gaussian noise (higher dimensions)
- **ML with Laplacian noise (higher dimensions)**

4 Information theory and ML

5 Maximum a posteriori (MAP)

- Incorporating a prior
- Conjugate prior
- Sequential learning

Linear regression in higher dimensions

We again approach the problem by assuming that each y_i is a realization of a random variable Y_i , generated according to

$$y_i = \theta^T x_i + \epsilon_i \quad i = 1, \dots, n$$

but this time, ϵ_i 's are independent, identically distributed Laplacian random variables with mean zero and parameter λ . We use a Laplacian distribution to be robust against outliers.

Once again, we aim to compute the ML estimates of θ .

Linear regression in higher dimensions

- 1 Given the vector θ , we have $Y_i = \text{Lap}(\theta^T x_i, \lambda)$

Linear regression in higher dimensions

- 1 Given the vector θ , we have $Y_i = \text{Lap}(\theta^T \mathbf{x}_i, \lambda)$
- 2 Likelihood function is therefore

$$f_Y(\mathbf{y}; \theta) = \prod_{i=1}^n \frac{\lambda}{2} \exp \{-\lambda |y_i - \theta^T \mathbf{x}_i|\}$$

Linear regression in higher dimensions

- ① Given the vector θ , we have $Y_i = \text{Lap}(\theta^T \mathbf{x}_i, \lambda)$
- ② Likelihood function is therefore

$$f_Y(\mathbf{y}; \theta) = \prod_{i=1}^n \frac{\lambda}{2} \exp \{-\lambda |y_i - \theta^T \mathbf{x}_i|\}$$

- ③ Log-likelihood function is

$$\begin{aligned} \log f_Y(\mathbf{y}; \theta) &= \sum_{i=1}^n \left(\log \frac{\lambda}{2} - \lambda |y_i - \theta^T \mathbf{x}_i| \right) \\ &= n \log \frac{\lambda}{2} - \lambda \sum_{i=1}^n |y_i - \theta^T \mathbf{x}_i| \end{aligned}$$

Linear regression in higher dimensions

Maximizing the log-likelihood is equivalent to

$$\text{minimize} \quad \sum_{i=1}^n |y_i - \theta^T x_i| \quad (11)$$

This is also known as the **minimum absolute difference (MAD)** or **minimum absolute error (MAE)**, or ℓ_1 -norm solution.

Generalized Gaussian distribution

It is possible to define a **generalized Gaussian distribution**

$$f_X(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp \left\{ - \left(\frac{|x - \mu|}{\alpha} \right)^\beta \right\} \quad (12)$$

where $\Gamma(\cdot)$ is the gamma function.

Generalized Gaussian distribution

It is possible to define a **generalized Gaussian distribution**

$$f_X(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp \left\{ - \left(\frac{|x - \mu|}{\alpha} \right)^\beta \right\} \quad (12)$$

where $\Gamma(\cdot)$ is the gamma function.

- When $\beta = 2$: Gaussian with mean μ and variance $\alpha^2/2$
- When $\beta = 1$: Laplacian with mean μ and parameter $\lambda = 1/\alpha$
- As $\beta \rightarrow \infty$: approaches $U(\mu - \alpha, \mu + \alpha)$

We can show that the ML estimation relates to the ℓ_p norm where p is given by the shape parameter β .

- 1 Estimator and estimation error
- 2 Maximum likelihood (ML)
- 3 Linear regression
 - ML with Gaussian noise (2D)
 - ML with Gaussian noise (higher dimensions)
 - ML with Laplacian noise (higher dimensions)
- 4 Information theory and ML**
- 5 Maximum a posteriori (MAP)
 - Incorporating a prior
 - Conjugate prior
 - Sequential learning

Information theory

Our goal in this part is to introduce a few key concepts in information theory, and relate in particular to the ML estimation for further insight.

- Entropy
- Joint entropy
- Conditional entropy
- Relative entropy, also known as the Kullback-Leibler divergence
- Mutual information

Entropy

Entropy is a measure of *uncertainty of a random variable*. Let X be a random variable with PMF $p_X(x)$. Then the entropy of X is defined as

$$H(X) = - \sum_x p_X(x) \log_2 p_X(x) = \mathbb{E} \left(\log \frac{1}{p_X(x)} \right) \quad (13)$$

The entropy is a measure of uncertainty of a random variable. The logarithm is base 2, and entropy is measured in bits.

- By convention, $0 \log_2 0 = 0$, which is justified because $\lim_{x \rightarrow 0} x \log x = 0$.
- Interpretation: Adding terms with zero probability should not change the entropy.
- $H(X) \geq 0$ because $-\log_2 p_X(x) \geq 0$, since $0 \leq p_X(x) \leq 1$.

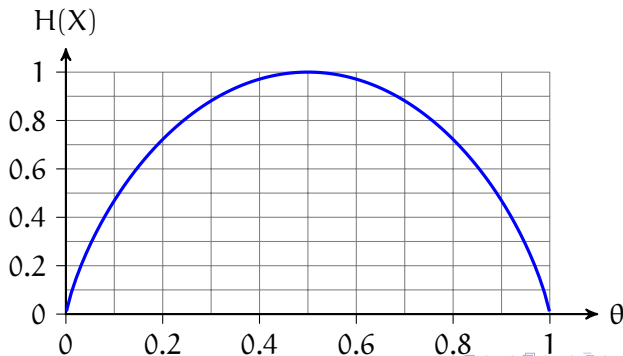
Entropy

For a Bernoulli random variable with parameter θ ,

$$p_X(x) = \theta^x (1 - \theta)^{1-x} \quad x = 0, 1$$

Therefore, the entropy is

$$H(X) = -\theta \log_2(\theta) - (1 - \theta) \log_2(1 - \theta)$$



Example

Entropy of flipping coins:

- Unbiased coin: $H(X) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ bit.
- Biased coin: $H(X) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.469$ bit.

Example

Entropy of flipping coins:

- Unbiased coin: $H(X) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ bit.
- Biased coin: $H(X) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.469$ bit.

Another example: $X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & \text{with probability } \frac{1}{4} \\ c & \text{with probability } \frac{1}{8} \\ d & \text{with probability } \frac{1}{8} \end{cases}$ Then,

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = \frac{7}{4} \text{ bit.}$$

Example

Entropy of flipping coins:

- Unbiased coin: $H(X) = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$ bit.
- Biased coin: $H(X) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 = 0.469$ bit.

Another example: $X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & \text{with probability } \frac{1}{4} \\ c & \text{with probability } \frac{1}{8} \\ d & \text{with probability } \frac{1}{8} \end{cases}$ Then,

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = \frac{7}{4} \text{ bit.}$$

$H(X)$ measures the number of bits needed to represent the information.
The more uneven the distribution, the better! (smaller $H(X)$)

Joint entropy

The **joint entropy** is a measure of *uncertainty of a pair of random variables*. Let (X, Y) be a pair of random variables with PMF $p_{X,Y}(x, y)$. Then the joint entropy of (X, Y) is defined as

$$H(X, Y) = - \sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X,Y}(x, y) = \mathbb{E} \left(\log \frac{1}{p_{X,Y}(x, y)} \right)$$

(14)

Conditional entropy

The **conditional entropy** is defined as

$$\begin{aligned}
 H(Y|X) &= \sum_x p_X(x) H(Y|X=x) \\
 &= - \sum_x p_X(x) \left(\sum_y p_{Y|X}(y|x) \log p_{Y|X}(y|x) \right) \\
 &= - \sum_x \sum_y p_{X,Y}(x,y) \log p_{Y|X}(y|x) \\
 &= -\mathbb{E}(\log p_{Y|X}(y|x))
 \end{aligned}$$

which leads to the chain rule

$$H(X, Y) = H(X) + H(Y|X) \quad (15)$$

since $\log p_{X,Y}(x,y) = \log p_X(x) + \log p_{Y|X}(y|x)$.

Relative entropy

To simplify notation, we use

$p(x)$ for one PMF over different values of $\{X = x\}$

$q(x)$ for another PMF over the same set of $\{X = x\}$

The **relative entropy**, a.k.a. the **Kullback-Leibler (or KL) divergence**, between two distributions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) = \mathbb{E}_p \left[\log_2 \left(\frac{p(x)}{q(x)} \right) \right] \quad (16)$$

Relative entropy

To simplify notation, we use

$p(x)$ for one PMF over different values of $\{X = x\}$

$q(x)$ for another PMF over the same set of $\{X = x\}$

The **relative entropy**, a.k.a. the **Kullback-Leibler (or KL) divergence**, between two distributions $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{q(x)} \right) = \mathbb{E}_p \left[\log_2 \left(\frac{p(x)}{q(x)} \right) \right] \quad (16)$$

Note: $D(p\|q) \geq 0$ and equals to zero if and only if $p(x) = q(x)$.

Relative entropy

The KL divergence can be thought of as measuring the “distance” between two distributions, but mathematically does not qualify as a distance metric.

Example: Suppose we have two Bernoulli distributions $p \sim \text{Ber}(r)$ and $q \sim \text{Ber}(s)$.

$$D(p\|q) = (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s}$$

$$D(q\|p) = (1-s) \log \frac{1-s}{1-r} + s \log \frac{s}{r}$$

If $r = s$, we can see $D(p\|q) = D(q\|p) = 0$.

If $r = 0.5$ and $s = 0.25$, we have $D(p\|q) = 0.2075$ and $D(q\|p) = 0.1887$. Note that $D(p\|q) \neq D(q\|p)$ in general.

Relative entropy and mutual information

The relative entropy leads us to the definition of **mutual information**, which is the relative entropy between the joint distribution and the product distribution:

$$I(X; Y) = \sum_{x, y} p_{X, Y}(x, y) \log \frac{p_{X, Y}(x, y)}{p_X(x) p_Y(y)} = D(p_{X, Y}(x, y) \parallel p_X(x) p_Y(y))$$

(17)

Relative entropy and mutual information

The relative entropy leads us to the definition of **mutual information**, which is the relative entropy between the joint distribution and the product distribution:

$$I(X; Y) = \sum_{x, y} p_{X, Y}(x, y) \log \frac{p_{X, Y}(x, y)}{p_X(x) p_Y(y)} = D(p_{X, Y}(x, y) \| p_X(x) p_Y(y))$$
(17)

We can further derive that

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
(18)

Relative entropy and mutual information

The relative entropy leads us to the definition of **mutual information**, which is the relative entropy between the joint distribution and the product distribution:

$$I(X; Y) = \sum_{x, y} p_{X, Y}(x, y) \log \frac{p_{X, Y}(x, y)}{p_X(x) p_Y(y)} = D(p_{X, Y}(x, y) \| p_X(x) p_Y(y)) \quad (17)$$

We can further derive that

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (18)$$

Mutual information $I(X; Y)$ is the reduction in uncertainty of X due to knowledge of Y .

Relative entropy and mutual information

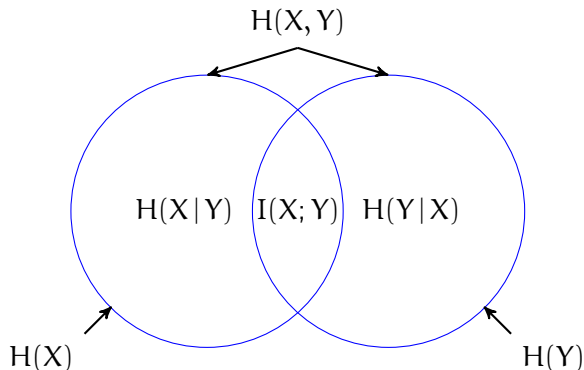
We can also substitute $H(X, Y) = H(X) + H(Y|X)$ to get

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (19)$$

Relative entropy and mutual information

We can also substitute $H(X, Y) = H(X) + H(Y|X)$ to get

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (19)$$



Relative entropy and ML estimation

Relationship between relative entropy and ML estimation:

Problem setting:

- We get data $x_1, \dots, x_N$ from some unknown distribution $p(x)$ (not necessarily of any particular parametric form).
- We wish to fit a distribution $q(x; \theta)$ that has some parametric form (e.g. Gaussian) that best explains the data.
- We derive the ML estimate for the parameters of distributions (θ) of a certain parametric form.

Relative entropy and ML estimation

The ML estimate is given by:

$$\begin{aligned}
 \hat{\Theta} &= \arg \max_{\theta} \prod_{i=1}^N q(x_i | \theta) \\
 &= \arg \max_{\theta} \sum_{i=1}^N \log q(x_i | \theta) \\
 &= \arg \max_{\theta} \frac{1}{N} \sum_{i=1}^N \log q(x_i | \theta) - \frac{1}{N} \sum_{i=1}^N \log p(x_i) \\
 &= \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N \log \left(\frac{p(x_i)}{q(x_i | \theta)} \right)
 \end{aligned}$$

Note that the highlighted term does not affect the arg max.

Relative entropy and ML estimation

In the limit as $N \rightarrow \infty$, we replace the *average* by the *expectation*,

$$\begin{aligned}\hat{\Theta} &\approx \arg \min_{\theta} \mathbb{E} \left[\log \left(\frac{p(\mathbf{x})}{q(\mathbf{x} | \theta)} \right) \right] \\ &= \arg \min_{\theta} D(p(\mathbf{x}) \| q(\mathbf{x} | \theta))\end{aligned}$$

Relative entropy and ML estimation

In the limit as $N \rightarrow \infty$, we replace the *average* by the *expectation*,

$$\begin{aligned}\hat{\Theta} &\approx \arg \min_{\theta} \mathbb{E} \left[\log \left(\frac{p(\mathbf{x})}{q(\mathbf{x}|\theta)} \right) \right] \\ &= \arg \min_{\theta} D(p(\mathbf{x}) \| q(\mathbf{x}|\theta))\end{aligned}$$

Interpretation: The ML estimate can be viewed as finding parameters (from some family of distributions) that minimizes the relative entropy between the true distribution generating the data and the modeled distribution from this family.

- 1 Estimator and estimation error
- 2 Maximum likelihood (ML)
- 3 Linear regression
 - ML with Gaussian noise (2D)
 - ML with Gaussian noise (higher dimensions)
 - ML with Laplacian noise (higher dimensions)
- 4 Information theory and ML
- 5 Maximum a posteriori (MAP)
 - Incorporating a prior
 - Conjugate prior
 - Sequential learning

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:

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (20)$$

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:

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (20)$$

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

The interpretation is:

$$\text{posterior} = \text{prior} \times \text{likelihood} \quad (21)$$

Bayesian

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

$$f_{\Theta|X}(\theta|x) \propto f_{\Theta}(\theta) p_{X|\Theta}(x|\theta) \quad (20)$$

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

The interpretation is:

$$\text{posterior} = \text{prior} \times \text{likelihood} \quad (21)$$

A further interpretation:

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}) \quad (22)$$

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

Example

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

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

We obtain the MAP estimate by maximizing

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

Example

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

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

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

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

(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

(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

(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

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

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 Estimator and estimation error
- 2 Maximum likelihood (ML)
- 3 Linear regression
 - ML with Gaussian noise (2D)
 - ML with Gaussian noise (higher dimensions)
 - ML with Laplacian noise (higher dimensions)
- 4 Information theory and ML
- 5 Maximum a posteriori (MAP)
 - Incorporating a prior
 - **Conjugate prior**
 - Sequential learning

Example

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

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

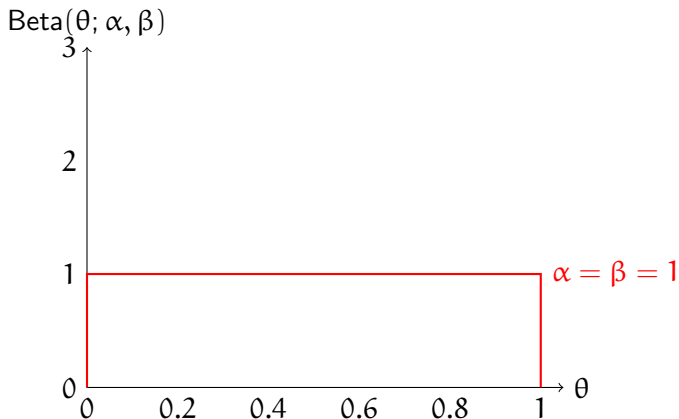
where $B(\alpha, \beta)$ is a normalizing constant, known as the beta function, given by

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

(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.

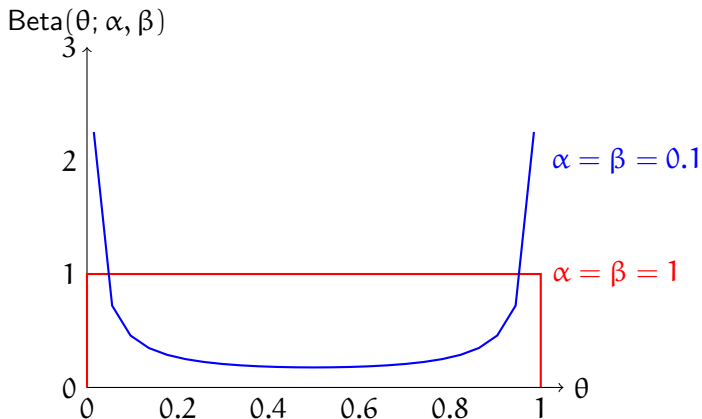
Beta distribution

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.



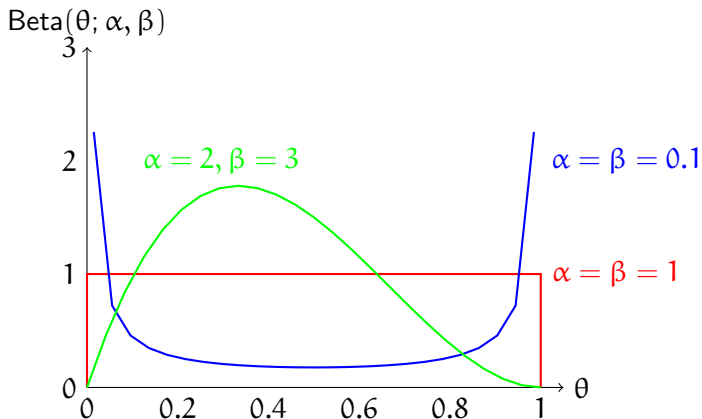
Beta distribution

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.



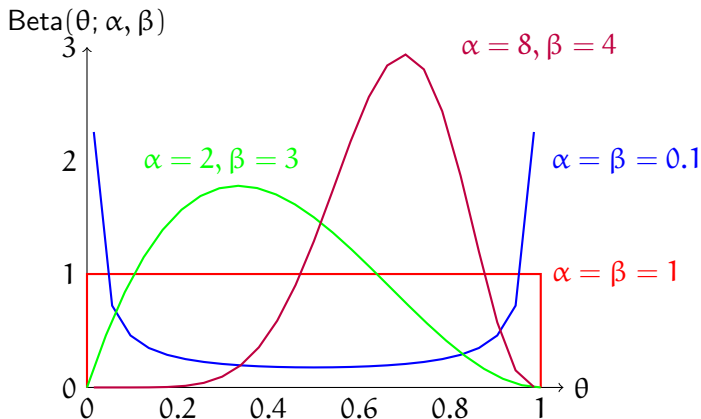
Beta distribution

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**

$$\mathbf{X} \sim \text{Mult}(n, \boldsymbol{\theta}) = \binom{n}{x_1, \dots, x_K} \prod_{j=1}^K \theta_j^{x_j} \quad (23)$$

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

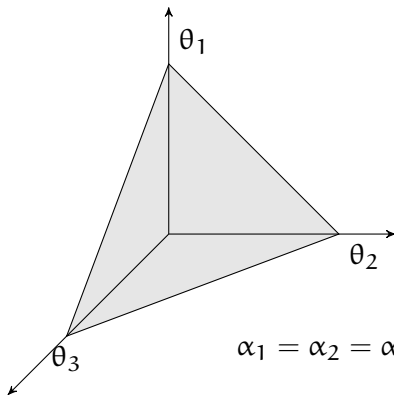
$$\text{Dir}(\theta; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_k^{\alpha_k - 1} \quad (24)$$

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.

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

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 Estimator and estimation error
- 2 Maximum likelihood (ML)
- 3 Linear regression
 - ML with Gaussian noise (2D)
 - ML with Gaussian noise (higher dimensions)
 - ML with Laplacian noise (higher dimensions)
- 4 Information theory and ML
- 5 Maximum a posteriori (MAP)
 - Incorporating a prior
 - Conjugate prior
 - Sequential learning

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,

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

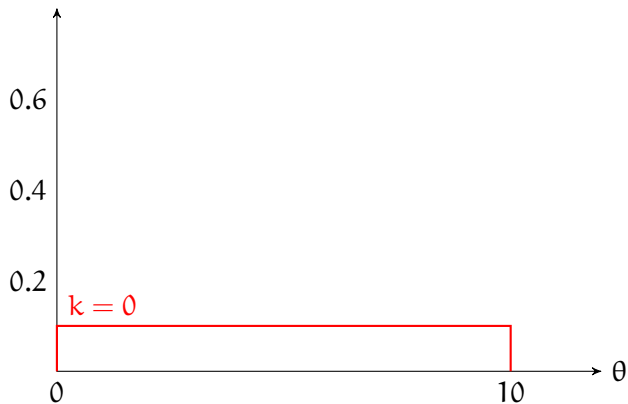
With $x_4 = 8$, we have

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

Recursive learning

Visualizing the results in plots:

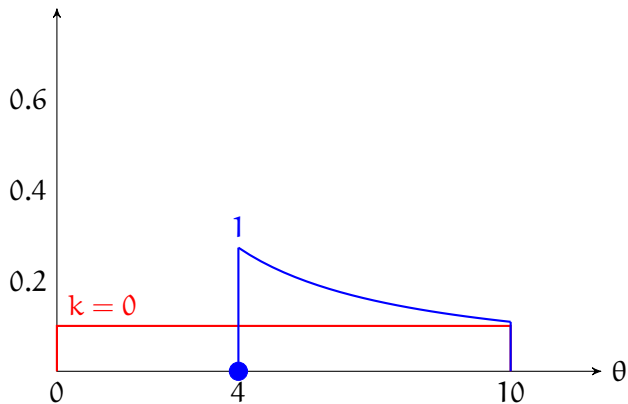
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

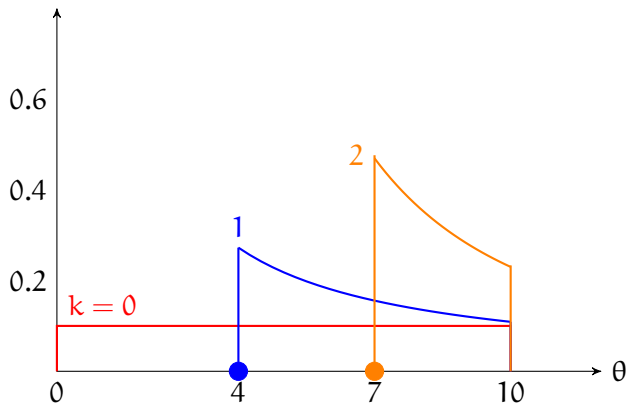
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

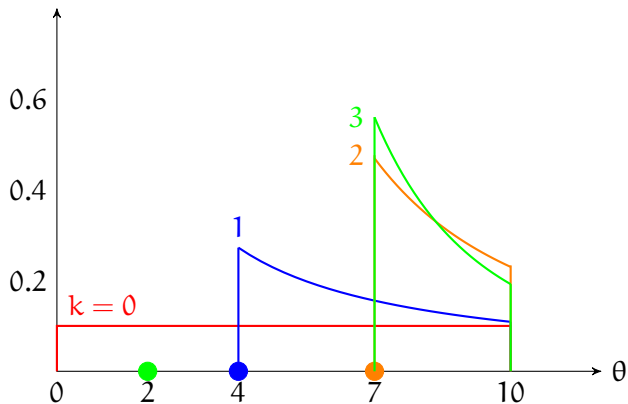
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



Recursive learning

Visualizing the results in plots:

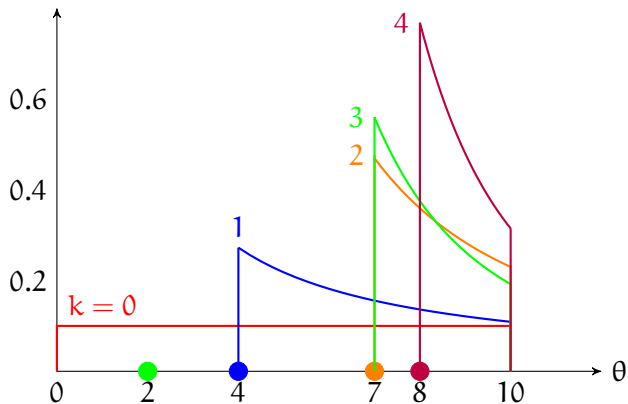
$$f_{\Theta | x_1 \dots x_k}(\theta | x_1, \dots, x_k)$$



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 .

Conclusions

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