Introduction to Probability and Bayesian Inference

Vineeth B. S.

Department of Avionics,
Indian Institute of Space Science and Technology,
Thiruvananthapuram.

Workshop on Data Assimilation in Weather and Climate Models

6th May 2024

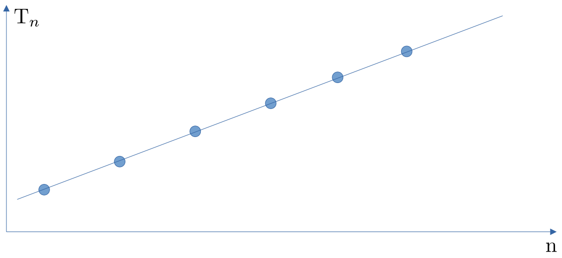
Motivation

A "data assimilation problem"

- Consider a simple problem where we consider the temperature at some point at some instants of time (say in days)
- Let the temperature at the n^{th} instant be T_n
- Suppose we think the temperature evolves in the following way

$$T_{n+1} = T_n + I_n.$$

- This is a model which we use to think about the temperature, maybe to even predict the temperature
- We could think of $I_n = I$ a constant parameter
- What is I in Bangalore?
- Where do we get I from?



A “data assimilation problem”

- Temperature model

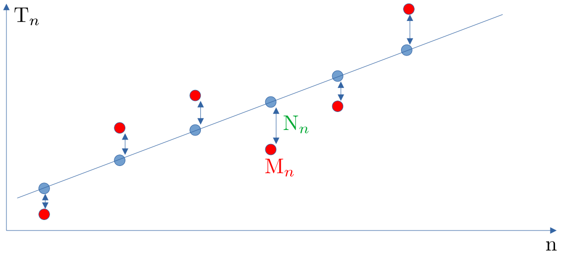
$$T_{n+1} = T_n + I.$$

- We have temperature measurements
- Do you think it would like?

$$M_n = T_n$$

- Or

$$M_n = T_n + N_n$$



A “data assimilation problem”

- How do we think about or model N_n?
- How do we find out I from data?
- We need the framework of probability and inference for this!

Introduction to probability

Sample Space

Definition

- The set of all possible outcomes
- Mutually exclusive
- Exhaustive with as much granularity as required
- The sample space is usually denoted by Ω
- Individual outcomes are represented by ω

Sample Space

Definition

- The set of all possible outcomes
- Mutually exclusive
- Exhaustive with as much granularity as required
- The sample space is usually denoted by Ω
- Individual outcomes are represented by ω

Examples

- For a coin toss: $\{\text{Head}, \text{Tail}\}$
- For a die roll: $\{1, 2, 3, 4, 5, 6\}$
- Position of a sensor: $[0, 1] \times [0, 1]$

Events

Definition

- A subset of the sample space
- The set of all such subsets is denoted as $\mathcal{F}$

Events

Definition

- A subset of the sample space
- The set of all such subsets is denoted as $\mathcal{F}$

Examples

- The number on the rolled dice is even
- The sensor lies within a distance of 0.25 meters from a relay

Probability

Definition

- Is a function that maps events to real numbers
- The function value can be interpreted as the long term fraction of time an event occurs
- The function value can also be interpreted as an amount of belief in the occurrence in the event
- The probability of an event E is denoted as $\Pr(E)$

Probability

Definition

- Is a function that maps events to real numbers
- The function value can be interpreted as the long term fraction of time an event occurs
- The function value can also be interpreted as an amount of belief in the occurrence in the event
- The probability of an event E is denoted as $\Pr(E)$

- $\Pr(\Omega) = 1$
- $0 \leq \Pr(E) \leq 1$
- $(A_1, A_2, \ldots, A_n, \ldots)$ are disjoint; $\sum_{i=1}^{\infty} \Pr(A_i) = \Pr(\bigcup_{i=1}^{\infty} A_i)$

Probability

Definition

- Is a function that maps events to real numbers
- The function value can be interpreted as the long term fraction of time an event occurs
- The function value can also be interpreted as an amount of belief in the occurrence in the event
- The probability of an event E is denoted as $\Pr(E)$

- $\Pr(\Omega) = 1$
- $0 \leq \Pr(E) \leq 1$
- $(A_1, A_2, \ldots, A_n, \ldots)$ are disjoint; $\sum_{i=1}^{\infty} \Pr(A_i) = \Pr(\bigcup_{i=1}^{\infty} A_i)$

Examples

- Die roll: $\Pr(\{f\}) = \frac{1}{6}$
- Probability of sensor in an area A inside $[0,1] \times [0,1]$ is $\Pr(A) = A$

Conditional probability

Definition

- A and B are two events
- Probability of A given that B has occurred; denoted by $\Pr(A|B)$
- Universe is now B
- If $\Pr(B) > 0$, then $\text{Pr}\{A|B\} = \frac{\text{Pr}\{A \cap B\}}{\text{Pr}\{B\}}$
- If $\Pr(B) = 0$, then $\text{Pr}\{A|B\}$ is undefined

Conditional probability

Definition

- A and B are two events
- Probability of A given that B has occurred; denoted by $\Pr(A|B)$
- Universe is now B
- If $\Pr(B) > 0$, then $\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}$
- If $\Pr(B) = 0$, then $\Pr\{A|B\}$ is undefined

Example

- Suppose you roll a fair six sided die
- What is the probability that the face is two given that the face is even?

Total probability theorem

Definition

- Suppose $B \subseteq \Omega$
- Suppose $(A_1, A_2, \ldots, A_n)$ are disjoint and $\bigcup_{i=1}^{n} A_i = \Omega$
- The total probability theorem states that

$$\Pr\{B\} = \sum_{i=1}^{n} \Pr\{A_i\} \Pr\{B|A_i\}$$

Total probability theorem

Definition

- Suppose $B \subseteq \Omega$
- Suppose $(A_1, A_2, \ldots, A_n)$ are disjoint and $\bigcup_{i=1}^{n} A_i = \Omega$
- The total probability theorem states that

$$\Pr\{B\} = \sum_{i=1}^{n} \Pr\{A_i\} \Pr\{B|A_i\}$$

Question

How to derive the above theorem?

Independent events

Definition

- Events A and B are independent if

$$
\begin{aligned}
\Pr\{A \cap B\} &= \Pr\{A\} \times \Pr\{B\} \\
\Pr\{A|B\} &= \Pr\{A\}
\end{aligned}
$$

Independent events

Definition

- Events A and B are independent if

$$\begin{aligned} \Pr\{A \cap B\} &= \Pr\{A\} \times \Pr\{B\} \\ \Pr\{A|B\} &= \Pr\{A\} \end{aligned}$$

Question

- Assume A and B are independent
- Now suppose an event C has occurred
- Are A and B independent given that C has occurred?

Discrete Random Variable

Definition

- $X : \Omega \to \mathbb{R}$
- X could be discrete or continuous valued
- We consider the case where X is discrete first

Discrete Random Variable

Definition

- $X : \Omega \to \mathbb{R}$
- X could be discrete or continuous valued
- We consider the case where X is discrete first

Examples

- X is the number of heads in 10 tosses of a coin with bias p
- X is the number of tosses until the first head

Probability Mass Function

Definition

- The probability mass function $p_X(x) = \Pr\{X = x\}$
- $p_X(x) = \Pr\{\omega : X(\omega) = x\}$
- $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$

Probability Mass Function

Definition

- The probability mass function $p_X(x) = \Pr\{X = x\}$
- $p_X(x) = \Pr\{\omega : X(\omega) = x\}$
- $p_X(x) \geq 0$ and $\sum_x p_X(x) = 1$

Examples

- X is the number of heads in N tosses of a coin with bias p. Then X is Binomial(N, p)
- X is the number of tosses until the first head. Then X is a Geometric(p) random variable

Probability Distributions

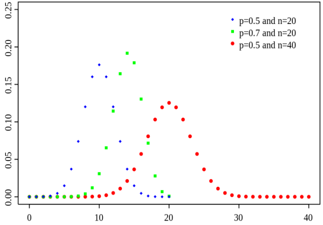
- The cumulative distribution function $F_X(x) = \Pr\{X \leq x\}$
- The complementary cumulative distribution function $F_X^c(x) = \Pr\{X > x\}$

Probability Distributions

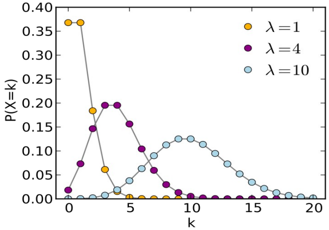
- The cumulative distribution function $F_X(x) = \mathsf{Pr}\{X \leq x\}$
- The complementary cumulative distribution function $F_X^c(x) = \mathsf{Pr}\{X > x\}$
- $F_X(-\infty) = 0, F_X(\infty) = 1$
- $F_X(x) = 1 - F_X^c(x)$

Some standard distributions

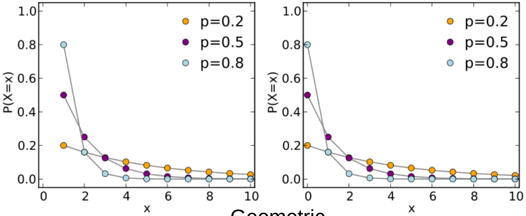
- Binomial random variable
- Geometric random variable
- Poisson random variable



Binomial



Poisson



Geometric

Expectation of a discrete random variable

Definition

- X is a non-negative discrete random variable
- The expectation of X is defined as $\sum_x p_X(x)x$
- The expectation is denoted as $\mathbb{E}X$

Expectation of a discrete random variable

Definition

- X is a non-negative discrete random variable
- The expectation of X is defined as $\sum_x p_X(x)x$
- The expectation is denoted as $\mathbb{E}X$

Question

- Suppose X is Uniform on $\{1, 2, 3, \ldots, 10\}$. What is $\mathbb{E}X$?
- If X is not restricted to be non-negative, how do you think $\mathbb{E}X$ will be defined?

Expectation of a discrete random variable

Properties

- The expectation is linear. Suppose X and Y are two random variables,g then $\mathbb{E}\left[\alpha X + \beta Y\right] = \alpha \mathbb{E} X + \beta \mathbb{E} Y$
- Suppose $Y = g(X)$, then $\mathbb{E} Y = \sum_x g(x) p_X(x)$

Higher moments and Variance

Definition

- The n^{th} moment of a random variable X is $\mathbb{E}X^n$
- The variance of a random variable X is $\mathbb{E}\left(X - \mathbb{E}X\right)^2$

Higher moments and Variance

Definition

- The n^{th} moment of a random variable X is $\mathbb{E}X^n$
- The variance of a random variable X is $\mathbb{E}\left(X - \mathbb{E}X\right)^2$

A question

- Find an expression for variance of X in terms of the mean and second moment of X

Conditional expectation

- Conditional probability: $p_{X|A}(x)$
- Conditional expectation: $\mathbb{E}\left[X|A\right]$
- Total expectation theorem: Suppose $(A_1, A_2, \ldots, A_n)$ are disjoint. Then
  $\mathbb{X} = \mathsf{Pr}\left\{A_1\right\}\mathbb{E}\left[X|A_1\right] + \cdots + \mathsf{Pr}\left\{A_n\right\}\mathbb{E}\left[X|A_n\right]$

Multiple discrete random variables

- X and Y are two discrete random variables
- The joint probability mass function $\Pr\{X = x, Y = y\}$ is denoted as $p_{X,Y}(x, y)$
- The marginal probability mass function $p_X(x) = \sum_y p_{X,Y}(x, y)$
- The conditional probability mass function $p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

Independent random variables

- Suppose X and Y are discrete random variables with probability mass functions $p_X(x)$ and $p_Y(y)$
- X and Y are independent if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all x and y
- X and Y are independent if $p_{X|Y=y}(x) = p_X(x)$ for all x and y
- $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$

Continuous Random Variable

Definition

- $X : \Omega \to \mathbb{R}$
- X is described by a probability density function f_X
- $\mathsf{Pr}\{a \leq X < b\} = \int_a^b f_X(x)dx$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $\mathbb{E}X = \int_0^{\infty} x f_X(x)dx$ for non-negative X
- Similar definitions for CDF and CCDF

Continuous Random Variable

Definition

- $X : \Omega \to \mathbb{R}$
- X is described by a probability density function f_X
- Pr $\{a \leq X < b\} = \int_a^b f_X(x)dx$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $\mathbb{E}X = \int_0^{\infty} x f_X(x)dx$ for non-negative X
- Similar definitions for CDF and CCDF

Example

- X is Uniform$[a, b]$. $f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
- X is Normal with mean μ and variance σ^2. $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Multiple continuous random variables

Definitions

- X and Y are two continuous random variables
- The joint distribution is $f_{X,Y}(x, y)$
- Marginal distribution of X is $f_X(x)$
- Conditional distribution is $f_{X|Y=y}(x)$ defined as $\frac{f_{X,Y}(x,y)}{f_Y(y)}$
- X and Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x and y

Multiple continuous random variables

Definitions

- X and Y are two continuous random variables
- The joint distribution is $f_{X,Y}(x, y)$
- Marginal distribution of X is $f_X(x)$
- Conditional distribution is $f_{X|Y=y}(x)$ defined as $\frac{f_{X,Y}(x,y)}{f_Y(y)}$
- X and Y are independent if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all x and y

Question

- Suppose you have a stick of length l
- You break it once at a position uniformly distributed in $[0, l]$ and then again break the left portion at a uniformly distributed position
- What is the joint distribution of the two "left" portions?

Bayes' rule

Definition

- X and Y are two random variables
- $p_{X|Y=y}(x) = \frac{p_X(x) p_{Y|X=x}(y)}{p_Y(y)}$
- $p_{X|Y=y}(x) = \frac{p_X(x) f_{Y|X=x}(y)}{f_Y(y)}$
- $f_{X|Y=y}(x) = \frac{f_X(x) p_{Y|X=x}(y)}{p_Y(y)}$
- $f_{X|Y=y}(x) = \frac{f_X(x) f_{Y|X=x}(y)}{f_Y(y)}$

Bayes' rule

Definition

- X and Y are two random variables
- $p_{X|Y=y}(x) = \frac{p_X(x)p_{Y|X=x}(y)}{p_Y(y)}$
- $p_{X|Y=y}(x) = \frac{p_X(x)f_{Y|X=x}(y)}{f_Y(y)}$
- $f_{X|Y=y}(x) = \frac{f_X(x)p_{Y|X=x}(y)}{p_Y(y)}$
- $f_{X|Y=y}(x) = \frac{f_X(x)f_{Y|X=x}(y)}{f_Y(y)}$

A question

- Suppose X takes values -1 and 1 with probability p and $1-p$
- Y is normally distributed with mean X and variance of 1
- Suppose you have observed $Y = -0.5$
- What is $p_{X|Y=-0.5}(1)$?

Bayesian Inference

Our example problem

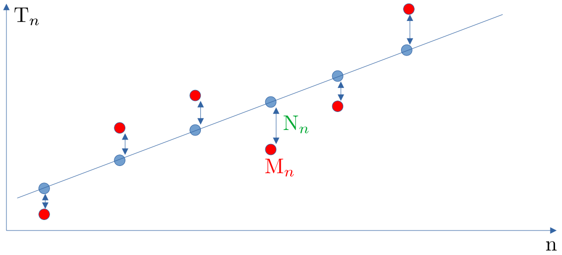
- Temperature model

$$T_{n+1} = T_n + I.$$

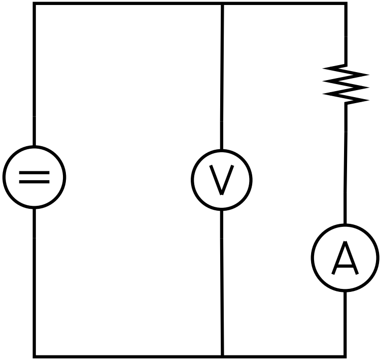
- We have temperature measurements

$$M_n = T_n + N_n$$

- N_n is measurement noise - modelled as a Normal$(0, \sigma^2)$ random variable - independent across n.
- How do we find I?

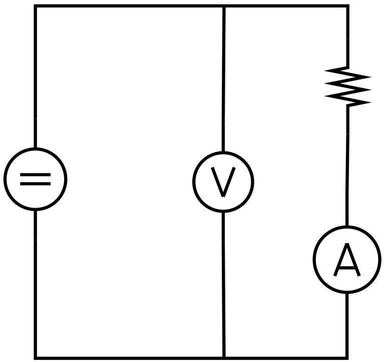


Another example

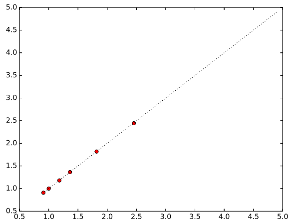


No.	Vr(V)	Ir(A)
0	0.909091	0.909091
1	1.000091	0.999092
2	1.182033	1.179673
3	1.363636	1.363636
4	1.818182	1.818182
5	2.455656	2.443439
6	2.728752	
7	1.455465	
8	1.092377	
9	4.555556	

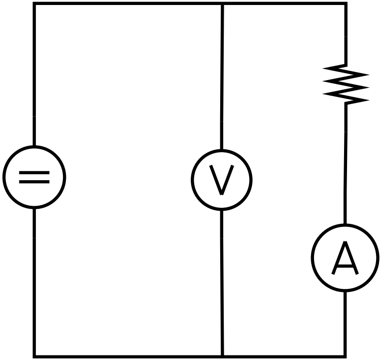
Another example



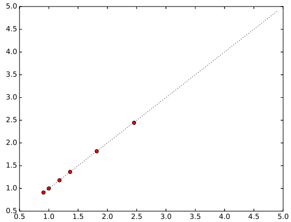
No.	Vr(V)	Ir(A)
0	0.909091	0.909091
1	1.000091	0.999092
2	1.182033	1.179673
3	1.363636	1.363636
4	1.818182	1.818182
5	2.455656	2.443439
6	2.728752	
7	1.455465	
8	1.092377	
9	4.555556	



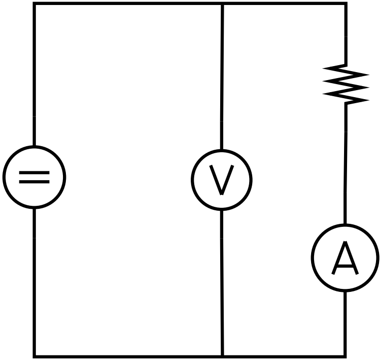
Another example



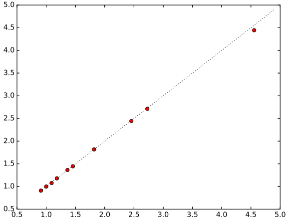
No.	Vr(V)	Ir(A)
0	0.909091	0.909091
1	1.000091	0.999092
2	1.182033	1.179673
3	1.363636	1.363636
4	1.818182	1.818182
5	2.455656	2.443439
6	2.728752	2.728752
7	1.455465	1.455465
8	1.092377	1.092377
9	4.555556	4.555556



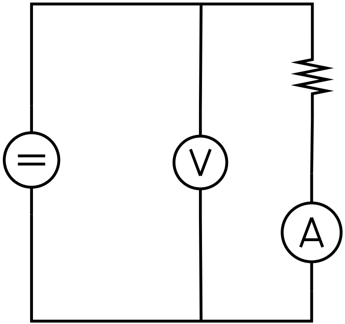
Another example



No.	Vr(V)	Ir(A)
0	0.909091	0.909091
1	1.000091	0.999092
2	1.182033	1.179673
3	1.363636	1.363636
4	1.818182	1.818182
5	2.455656	2.443439
6	2.728752	2.712477
7	1.455465	1.445348
8	1.092377	1.076233
9	4.555556	4.444444

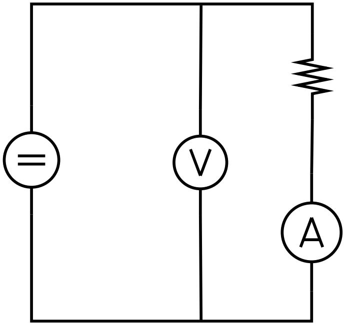


An example



No.	Vr(V)	Ir(A)	T(C)
0	0.909091	0.909091	25
1	1.000091	0.999092	26
2	1.182033	1.179673	27
3	1.363636	1.363636	25
4	1.818182	1.818182	25
5	2.455656	2.443439	30
6	2.728752		31
7	1.455465		32
8	1.092377		40
9	4.555556		50

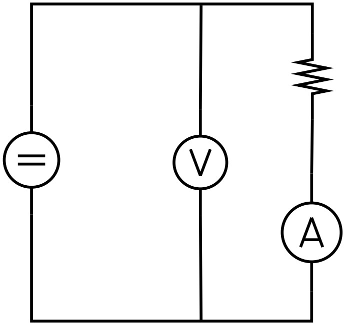
An example



No.	Vr(V)	Ir(A)	T(C)
0	0.909091	0.909091	25
1	1.000091	0.999092	26
2	1.182033	1.179673	27
3	1.363636	1.363636	25
4	1.818182	1.818182	25
5	2.455656	2.443439	30
6	2.728752		31
7	1.455465		32
8	1.092377		40
9	4.555556		50

$$\text{Ir} = \text{Vr}/(1 + 0.01 * (T - 25))$$

An example



No.	Vr(V)	Ir(A)	T(C)
0	0.909091	0.909091	25
1	1.000091	0.999092	26
2	1.182033	1.179673	27
3	1.363636	1.363636	25
4	1.818182	1.818182	25
5	2.455656	2.443439	30
6	2.728752	2.712477	31
7	1.455465	1.445348	32
8	1.092377	1.076233	40
9	4.555556	4.444444	50

$$\mathsf{Ir} = \mathsf{Vr}/(1 + 0.01 * (T - 25))$$

Problem ingredients

- Data

Problem ingredients

- Data with features

Problem ingredients

- Data with features
- A model

Problem ingredients

- Data with features
- A model with parameters

Problem ingredients

- Data with features
- A model with parameters
- A method of choosing parameters for the model

Problem ingredients

- Data with features
- A model with parameters
- A method of choosing parameters for the model from the data

Problem ingredients

- Data with features
- A model with parameters
- A method of choosing parameters for the model from the data (Inference)

Problem ingredients

- Data with features
- A model with parameters
- A method of choosing parameters for the model from the data (Inference)
- A method to predict using the model

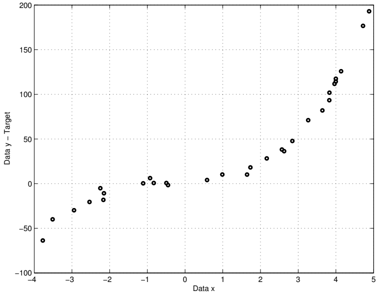
Problem ingredients

- Data with features
- A model with parameters
- A method of choosing parameters for the model from the data (Inference)
- A method to predict using the model (Prediction)

Problem ingredients

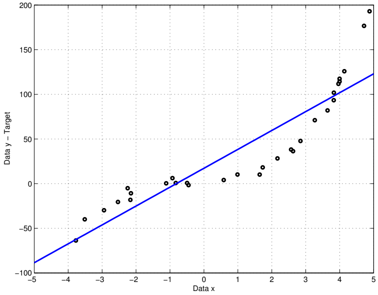
- Data with features
- A model with parameters
- A method of choosing parameters for the model from the data (Inference)
- A method to predict using the model (Prediction)
- Iterate, evaluate and select models (Model selection)

Example - Regression problem



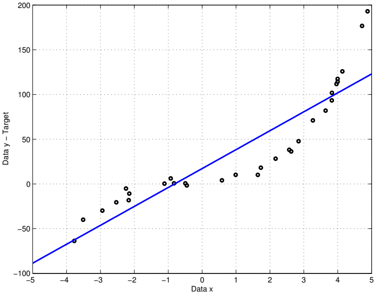
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable

Example - Regression problem



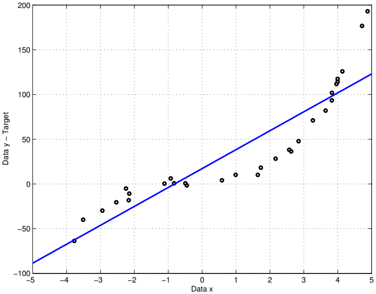
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable
- A model - a linear model for y in x is assumed. Parameters are slope and intercept.

Example - Regression problem



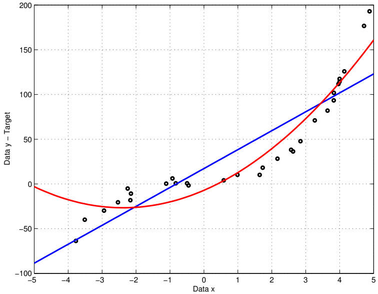
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable
- A model - a linear model for y in x is assumed. Parameters are slope and intercept.
- We use a squared error minimization technique to find out the parameters.

Example - Regression problem



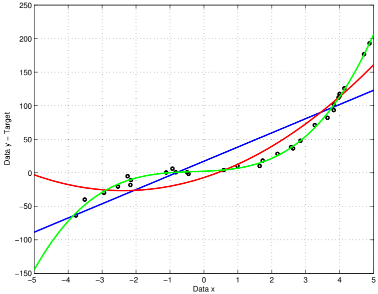
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable
- A model - a linear model for y in x is assumed. Parameters are slope and intercept.
- We use a squared error minimization technique to find out the parameters.
- We can use the model for prediction for new values of x.

Example - Regression problem



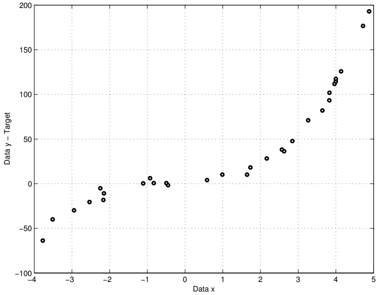
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable
- A model - a linear model for y in x is assumed. Parameters are slope and intercept.
- We use a squared error minimization technique to find out the parameters.
- We can use the model for prediction for new values of x.
- Model comparisons can be done, other polynomials in x.

Example - Regression problem



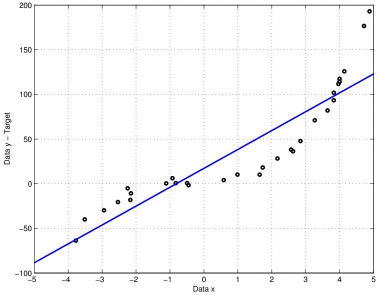
- Data (with two features or 2 dimensions) is given.
- One feature y is designated as a target variable
- A model - a linear model for y in x is assumed. Parameters are slope and intercept.
- We use a squared error minimization technique to find out the parameters.
- We can use the model for prediction for new values of x.
- Model comparisons can be done, other polynomials in x.

A framework for doing Inference, Prediction, Model Selection



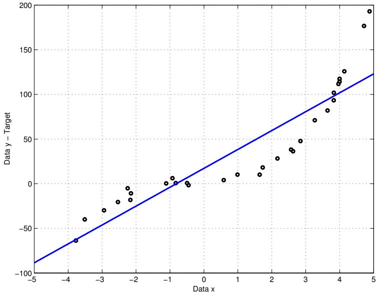
- We need a framework for doing the three basic problems that arise.

A framework for doing Inference, Prediction, Model Selection



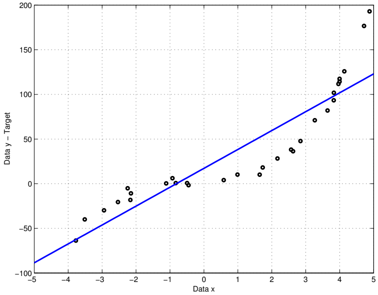
- We need a framework for doing the three basic problems that arise.
- It seems that we need to find "best" parameters for the model that we have assumed
- An approach which is widely used is to form a loss function or objective function that measures how good the model predicts on training data and use optimization techniques

A framework for doing Inference, Prediction, Model Selection



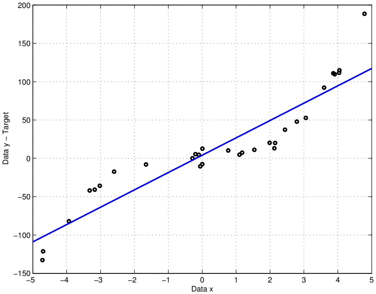
- We need a framework for doing the three basic problems that arise.
- It seems that we need to find “best” parameters for the model that we have assumed
- An approach which is widely used is to form a loss function or objective function that measures how good the model predicts on training data and use optimization techniques
- In other problems, similar optimization approaches can be used.

A framework for doing Inference, Prediction, Model Selection



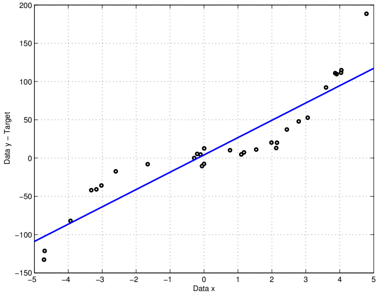
- We need a framework for doing the three basic problems that arise.
- It seems that we need to find "best" parameters for the model that we have assumed
- An approach which is widely used is to form a loss function or objective function that measures how good the model predicts on training data and use optimization techniques
- In other problems, similar optimization approaches can be used.
- Another approach is the Bayesian approach.

A question - interpretation of results

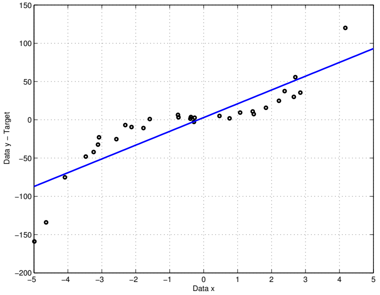


$$y = 22.6232x + 4.1200$$

A question - interpretation of results

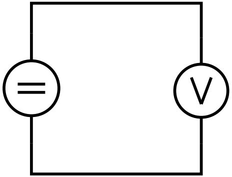


$$y = 22.6232x + 4.1200$$



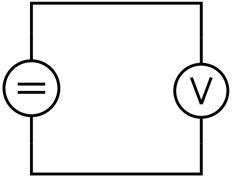
$$y = 18.0137x + 2.9210$$

Measuring a voltage source



- We measure the value of a DC voltage source.

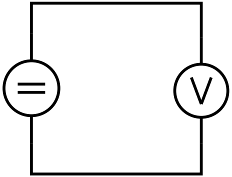
Measuring a voltage source



- We measure the value of a DC voltage source.

No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

Measuring a voltage source

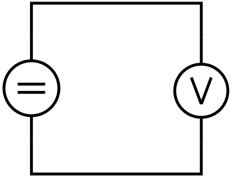


- We measure the value of a DC voltage source.

No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

- We report the average 5.2587.

Measuring a voltage source

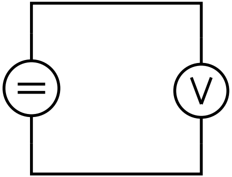


- We measure the value of a DC voltage source.

No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

- We report the average 5.2587.
- When we report the average we are fitting a constant to the data using minimum squared error.

Measuring a voltage source

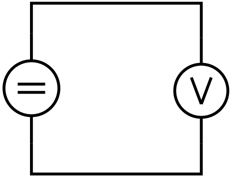


- We measure the value of a DC voltage source.

No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

- We report the average 5.2587.
- When we report the average we are fitting a constant to the data using minimum squared error.
- We again measure!

Measuring a voltage source



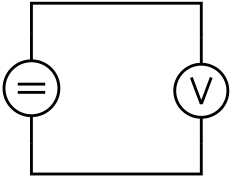
- We measure the value of a DC voltage source.

No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

- We report the average 5.2587.
- When we report the average we are fitting a constant to the data using minimum squared error.
- We again measure!

No.	Vm(V)
1	3.6923
2	4.5664
3	5.3426
4	8.5784
5	7.7694

Measuring a voltage source



- We measure the value of a DC voltage source.

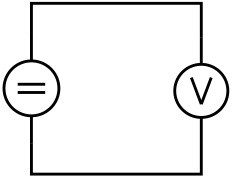
No.	Vm(V)
1	5.5377
2	6.8339
3	2.7412
4	5.8622
5	5.3188

- We report the average 5.2587.
- When we report the average we are fitting a constant to the data using minimum squared error.
- We again measure!

No.	Vm(V)
1	3.6923
2	4.5664
3	5.3426
4	8.5784
5	7.7694

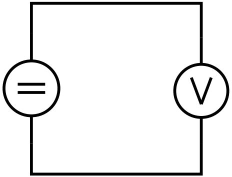
- The new average is 5.9898.

Measuring a voltage source



- We measure the value of a DC voltage source.

Measuring a voltage source



- We measure the value of a DC voltage source.
- For each measurement there is variation from a constant DC value because of some noise or some source of randomness
- So we will say that the i^{th} measurement is $5 + X_i$; X_i is Gaussian(0, 1).
- What interval shall we report?

Example: Coin bias

- You have a coin, which you use for deciding who gets to bat first
- You want to know(infer) whether the coin is fair or not
- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

Example: Coin bias

- You have a coin, which you use for deciding who gets to bat first
- You want to know(infer) whether the coin is fair or not
- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- So is the coin biased? What is the bias?

Example: Coin bias

- You have a coin, which you use for deciding who gets to bat first
- You want to know(infer) whether the coin is fair or not
- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- So is the coin biased? What is the bias?
- What if you know that the coin is not from a government mint?

What have we seen?

- A plethora of seemingly unconnected procedures for doing inference
- A not so easily understood way of reporting results
- An inability to incorporate prior information or domain information

Bayesian approach

- Bayesian belief: A quantification of how much we believe a particular statement is true.

Bayesian approach

- Bayesian belief: A quantification of how much we believe a particular statement is true.
  - The value of the dc source is 5 volts (100 % belief)
  - The value of the dc source is between 4.5 and 5.3 volts.
- We start with a prior belief about the statement

Bayesian approach

- Bayesian belief: A quantification of how much we believe a particular statement is true.
    - The value of the dc source is 5 volts (100 % belief)
    - The value of the dc source is between 4.5 and 5.3 volts.
- We start with a prior belief about the statement
- Then we observe data which depends on the statement

Bayesian approach

- Bayesian belief: A quantification of how much we believe a particular statement is true.
  - The value of the dc source is 5 volts (100 % belief)
  - The value of the dc source is between 4.5 and 5.3 volts.
- We start with a prior belief about the statement
- Then we observe data which depends on the statement
- We update our belief on the basis of our data (inference)

Bayesian approach

- Bayesian belief: A quantification of how much we believe a particular statement is true.
  - The value of the dc source is 5 volts (100 % belief)
  - The value of the dc source is between 4.5 and 5.3 volts.
- We start with a prior belief about the statement
- Then we observe data which depends on the statement
- We update our belief on the basis of our data (inference)
- We use the updated belief to make predictions and model selection.

Connections to probability

- Do we need to develop a new system of thinking based on belief quantities, a new arithmetic for beliefs?

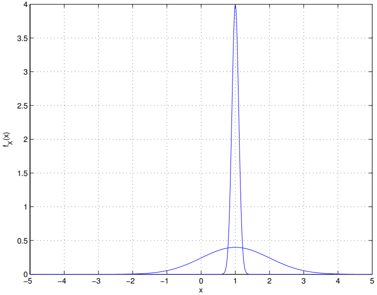
Connections to probability

- Do we need to develop a new system of thinking based on belief quantities, a new arithmetic for beliefs?
- Probability comes to the rescue
  - Any coherent belief quantification system should satisfy the rules of probability
  - Cox's axiomatic approach
  - Dutch book theorem

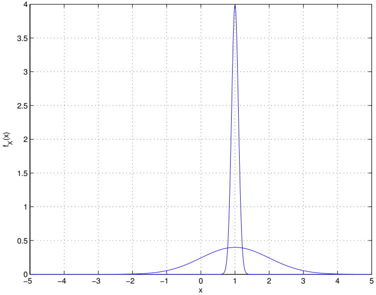
Connections to probability

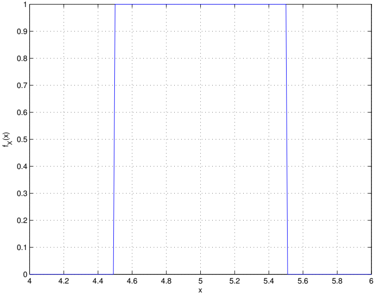
- Do we need to develop a new system of thinking based on belief quantities, a new arithmetic for beliefs?
- Probability comes to the rescue
  - Any coherent belief quantification system should satisfy the rules of probability
  - Cox's axiomatic approach
  - Dutch book theorem
- A statement can have a set of possible values; each value has an associated belief
- A random variable can have a set of possible values; the belief is given by the probability distribution

Priors beliefs or prior probabilities

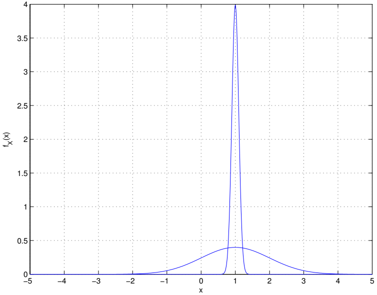


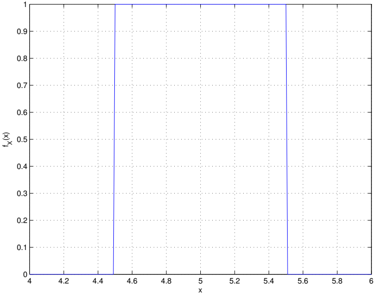
Priors beliefs or prior probabilities

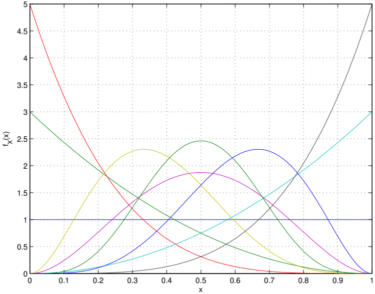




Priors beliefs or prior probabilities







Digression: Bayes' Rule

- X and Y are two discrete random variables
- $\Pr\{X = x|Y = y\} = \frac{\Pr\{X=x, Y=y\}}{\Pr\{Y=y\}}$
- Bayes' rule: $\Pr\{X = x|Y = y\} = \frac{\Pr\{X=x\}\Pr\{Y=y|X=x\}}{\Pr\{Y=y\}}$
- Another form: $\Pr\{X = x|Y = y\} = \frac{\Pr\{X=x\}\Pr\{Y=y|X=x\}}{\sum_{x'}\Pr\{X=x'\}\Pr\{Y|X=x'\}}$

Digression: An example for Bayes rule

	Y		
X	1	2	3
1	0.05	0.10	0.05
2	0.10	0.02	0.20
3	0.10	0.28	0.10

Table: Joint probability distribution Pr $\{X = x, Y = y\}$

Digression: An example for Bayes rule

	Y		
X	1	2	3
1	0.05	0.10	0.05
2	0.10	0.02	0.20
3	0.10	0.28	0.10

Table: Joint probability distribution $\Pr\{X = x, Y = y\}$

- What is $\Pr\{X = 1 | Y = 3\}$?

Digression: An example for Bayes rule

	Y		
X	1	2	3
1	0.05	0.10	0.05
2	0.10	0.02	0.20
3	0.10	0.28	0.10

Table: Joint probability distribution $\Pr\{X = x, Y = y\}$

- What is $\Pr\{X = 1 | Y = 3\}$?
- $\Pr\{X = 1, Y = 3\} = 0.05$

Digression: An example for Bayes rule

	Y		
X	1	2	3
1	0.05	0.10	0.05
2	0.10	0.02	0.20
3	0.10	0.28	0.10

Table: Joint probability distribution Pr $\{X = x, Y = y\}$

- What is Pr $\{X = 1 | Y = 3\}$?
- Pr $\{X = 1, Y = 3\} = 0.05$
- Pr $\{Y = 3\} = 0.05 + 0.20 + 0.10$

Digression: An example for Bayes rule

	Y		
X	1	2	3
1	0.05	0.10	0.05
2	0.10	0.02	0.20
3	0.10	0.28	0.10

Table: Joint probability distribution $\Pr\{X = x, Y = y\}$

- What is $\Pr\{X = 1 | Y = 3\}$?
- $\Pr\{X = 1, Y = 3\} = 0.05$
- $\Pr\{Y = 3\} = 0.05 + 0.20 + 0.10$
- $\Pr\{X = 1 | Y = 3\} = \frac{\Pr\{X=1, Y=3\}}{\Pr\{Y=3\}} = \frac{1}{7}$

Digression: An example for Bayes rule

X	1	2	3
1	0.20	0.32	0.48

Table: Marginal probability distribution Pr $\{X = x\}$

	Y		
X	1	2	3
1	0.2500	0.5000	0.2500
2	0.3125	0.0625	0.6250
3	0.2083	0.5833	0.2083

Table: Conditional probability distribution Pr $\{Y = y|X = x\}$

Digression: An example for Bayes rule

X	1	2	3
1	0.20	0.32	0.48

Table: Marginal probability distribution Pr $\{X = x\}$

	Y		
X	1	2	3
1	0.2500	0.5000	0.2500
2	0.3125	0.0625	0.6250
3	0.2083	0.5833	0.2083

Table: Conditional probability distribution Pr $\{Y = y|X = x\}$

- What is Pr $\{X = 1|Y = 3\}$?

Digression: An example for Bayes rule

X	1	2	3
1	0.20	0.32	0.48

Table: Marginal probability distribution Pr $\{X = x\}$

	Y		
X	1	2	3
1	0.2500	0.5000	0.2500
2	0.3125	0.0625	0.6250
3	0.2083	0.5833	0.2083

Table: Conditional probability distribution Pr $\{Y = y|X = x\}$

- What is Pr $\{X = 1|Y = 3\}$?
- Pr $\{X = 1\} \times$ Pr $\{Y = 3|x = 1\} = 0.20 \times 0.25 = 0.05$

Digression: An example for Bayes rule

X	1	2	3
1	0.20	0.32	0.48

Table: Marginal probability distribution Pr $\{X = x\}$

	Y		
X	1	2	3
1	0.2500	0.5000	0.2500
2	0.3125	0.0625	0.6250
3	0.2083	0.5833	0.2083

Table: Conditional probability distribution Pr $\{Y = y|X = x\}$

- What is Pr $\{X = 1|Y = 3\}$?
- Pr $\{X = 1\} \times$ Pr $\{Y = 3|x = 1\} = 0.20 \times 0.25 = 0.05$
- Pr $\{X = 1\} \times$ Pr $\{Y = 3|x = 1\} +$ Pr $\{X = 2\} \times$ Pr $\{Y = 3|x = 2\} +$ Pr $\{X = 3\} \times$ Pr $\{Y = 3|x = 3\} =$ $0.05 + 0.2 + 0.1 = 0.35$

Digression: An example for Bayes rule

X	1	2	3
1	0.20	0.32	0.48

Table: Marginal probability distribution Pr $\{X = x\}$

	Y		
X	1	2	3
1	0.2500	0.5000	0.2500
2	0.3125	0.0625	0.6250
3	0.2083	0.5833	0.2083

Table: Conditional probability distribution Pr $\{Y = y|X = x\}$

- What is Pr $\{X = 1|Y = 3\}$?
- Pr $\{X = 1\} \times$ Pr $\{Y = 3|x = 1\} = 0.20 \times 0.25 = 0.05$
- Pr $\{X = 1\} \times$ Pr $\{Y = 3|x = 1\} +$ Pr $\{X = 2\} \times$ Pr $\{Y = 3|x = 2\} +$ Pr $\{X = 3\} \times$ Pr $\{Y = 3|x = 3\} =$ $0.05 + 0.2 + 0.1 = 0.35$
- Pr $\{Y = 3|X = 1\} = \frac{\text{Pr}\{X=1\}\times\text{Pr}\{Y=3|x=1\}}{\text{Pr}\{X=1\}\times\text{Pr}\{Y=3|x=1\}+\text{Pr}\{X=2\}\times\text{Pr}\{Y=3|x=2\}+\text{Pr}\{X=3\}\times\text{Pr}\{Y=3|x=3\}} = \frac{1}{7}$

Example: Bayesian inference of bias of a coin

- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

Example: Bayesian inference of bias of a coin

- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- $\Pr\{Y = y|\Theta\} = \Theta^7(1-\Theta)^3$

Example: Bayesian inference of bias of a coin

- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- Pr $\{Y = y|\Theta\} = \Theta^7(1-\Theta)^3$
- Pr $\{\Theta = 0.4\} = 0.2$, Pr $\{\Theta = 0.5\} = 0.6$, Pr $\{\Theta = 0.6\} = 0.2$

Example: Bayesian inference of bias of a coin

- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- $\Pr\{Y = y|\Theta\} = \Theta^7(1-\Theta)^3$
- $\Pr\{\Theta = 0.4\} = 0.2, \Pr\{\Theta = 0.5\} = 0.6, \Pr\{\Theta = 0.6\} = 0.2$
- $\Pr\{Y = y\} = 0.2 \times (0.4)^7(0.6)^3 + 0.6 \times (0.5)^{10} + 0.2 \times (0.6)^7(0.4)^3 = 0.001$

Example: Bayesian inference of bias of a coin

- We observe the following sequence as the result of coin tosses

	1	2	3	4	5	6	7	8	9	10
Y	H	H	L	L	L	H	H	H	H	H

Table: 10 coin tosses

- $\Pr\{Y = y|\Theta\} = \Theta^7(1-\Theta)^3$
- $\Pr\{\Theta = 0.4\} = 0.2, \Pr\{\Theta = 0.5\} = 0.6, \Pr\{\Theta = 0.6\} = 0.2$
- $\Pr\{Y = y\} = 0.2 \times (0.4)^7(0.6)^3 + 0.6 \times (0.5)^{10} + 0.2 \times (0.6)^7(0.4)^3 = 0.001$
- $\Pr\{\Theta = 0.4|Y = y\} = \frac{0.2\times(0.4)^7(0.6)^3}{0.001} = 0.07$
- $\Pr\{\Theta = 0.5|Y = y\} = \frac{0.6\times(0.5)^{10}}{0.001} = 0.58$
- $\Pr\{\Theta = 0.6|Y = y\} = \frac{0.2\times(0.6)^7(0.4)^3}{0.001} = 0.35$

The Bayesian Procedure

$$\boxed{\mathsf{Pr}\left\{\Theta|Y=y\right\}} \quad \boxed{=}$$

The Bayesian Procedure

$\boxed{\mathsf{Pr}\left\{\Theta|Y=y\right\}}$ $\boxed{=}$ $\boxed{\mathsf{Pr}\left\{\Theta\right\}}$ $\boxed{\times}$

The Bayesian Procedure

$$\mathsf{Pr}\left\{\Theta | Y = y\right\} \quad = \quad \mathsf{Pr}\left\{\Theta\right\} \times \mathsf{Pr}\left\{Y = y | \Theta\right\}$$

The Bayesian Procedure

$$\Pr\{\Theta|Y=y\} = \frac{\Pr\{\Theta\} \times \Pr\{Y=y|\Theta\}}{\quad}$$

The Bayesian Procedure

$$\boxed{\text{Pr}\{\Theta | Y = y\}} \; \boxed{=} \; \frac{\boxed{\text{Pr}\{\Theta\}} \; \boxed{\times} \; \boxed{\text{Pr}\{Y = y | \Theta\}}}{\boxed{\text{Pr}\{Y = y\}}}$$

The Bayesian Procedure

$$\boxed{\text{Posterior}} \; \boxed{=} \; \frac{\boxed{\Pr\{\Theta\}} \; \boxed{\times} \; \boxed{\Pr\{Y = y|\Theta\}}}{\boxed{\Pr\{Y = y\}}}$$

The Bayesian Procedure

$$\text{Posterior} = \frac{\text{Prior} \times \Pr\{Y = y|\Theta\}}{\Pr\{Y = y\}}$$

The Bayesian Procedure

$$\boxed{\text{Posterior}} \; \boxed{=} \; \frac{\boxed{\text{Prior}} \; \boxed{\times} \; \boxed{\text{Likelihood}}}{\boxed{\Pr\{Y = y\}}}$$

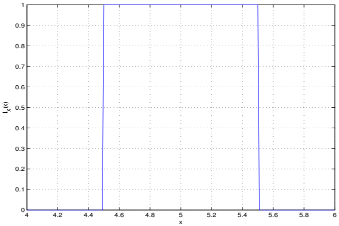
The Bayesian Procedure

$$\boxed{\text{Posterior}} \boxed{=} \frac{\boxed{\text{Prior}} \boxed{\times} \boxed{\text{Likelihood}}}{\boxed{\text{Evidence}}}$$

Bayesian voltage measurement

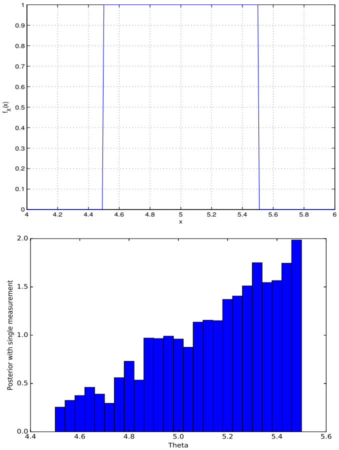
- Let Θ be the unknown voltage

Bayesian voltage measurement



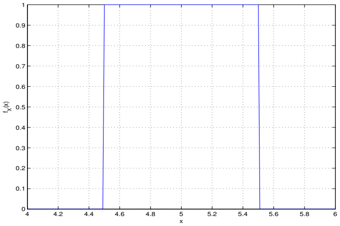
- Let Θ be the unknown voltage
- We believe that Θ could take values equally likely between 4.5 and 5.5 (the prior)

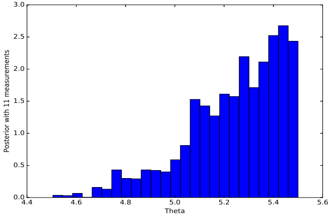
Bayesian voltage measurement

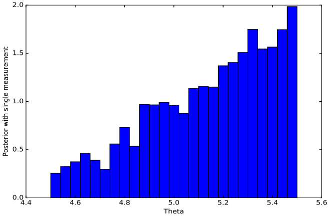


- Let Θ be the unknown voltage
- We believe that Θ could take values equally likely between 4.5 and 5.5 (the prior)
- We measure a value 6.5532. Using Bayes rule we obtain a posterior distribution from the prior

Bayesian voltage measurement

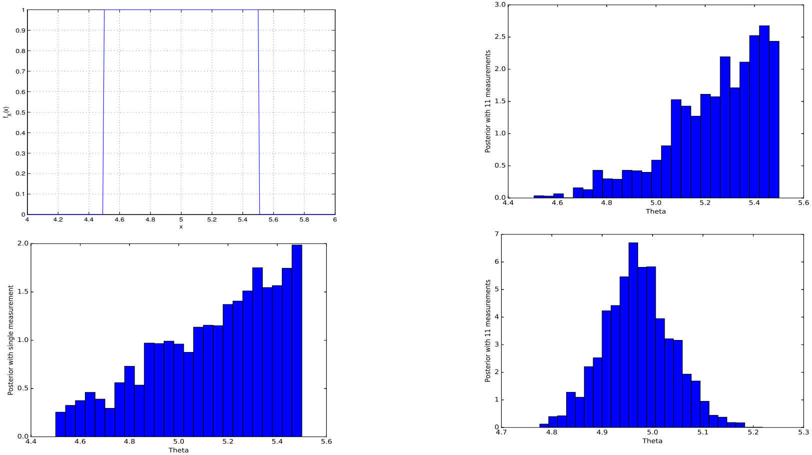






- Let Θ be the unknown voltage
- We believe that Θ could take values equally likely between 4.5 and 5.5 (the prior)
- We measure a value 6.5532. Using Bayes rule we obtain a posterior distribution from the prior
- Suppose new measurements are obtained. Then the current posterior is our current belief and plays the role of the prior for the next set of Bayesian updates

Bayesian voltage measurement

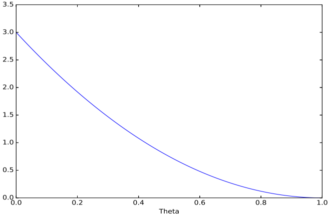


- Let Θ be the unknown voltage
- We believe that Θ could take values equally likely between 4.5 and 5.5 (the prior)
- We measure a value 6.5532. Using Bayes rule we obtain a posterior distribution from the prior
- Suppose new measurements are obtained. Then the current posterior is our current belief and plays the role of the prior for the next set of Bayesian updates

Bayesian coin bias

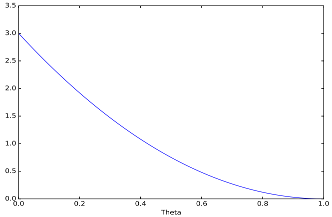
- Let Θ be the unknown bias of the coin

Bayesian coin bias



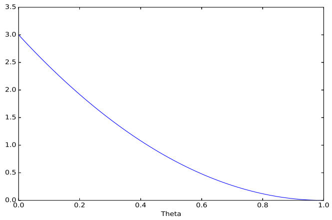
- Let Θ be the unknown bias of the coin
- We believe that the coin is unfair - the prior represents our belief

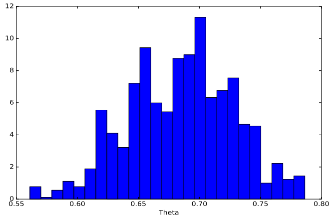
Bayesian coin bias



- Let Θ be the unknown bias of the coin
- We believe that the coin is unfair - the prior represents our belief
- We count 70 heads happening in 100 tosses of the coin

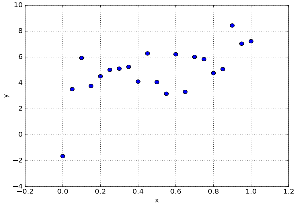
Bayesian coin bias





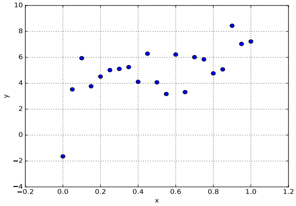
- Let Θ be the unknown bias of the coin
- We believe that the coin is unfair - the prior represents our belief
- We count 70 heads happening in 100 tosses of the coin
- Our posterior belief is as shown.

Bayesian regression



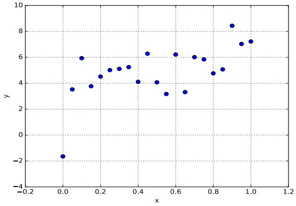
- Suppose we observe data as shown.

Bayesian regression



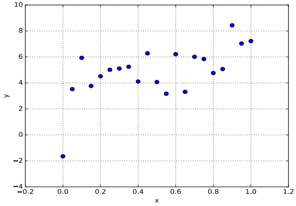
- Suppose we observe data as shown.
- We believe that the data can be explained as $y = ax + b + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$.

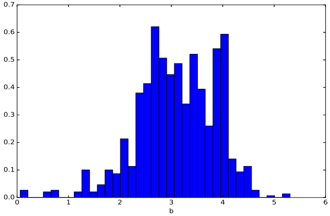
Bayesian regression

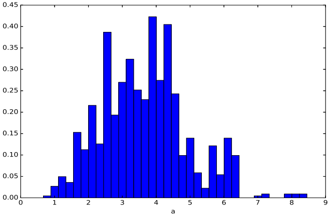


- Suppose we observe data as shown.
- We believe that the data can be explained as $y = ax + b + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$.
- We believe that $a \sim \mathcal{N}(0, 100), b \sim \mathcal{N}(0, 100), \tau \sim Gamma(0.1, 0.1)$.

Bayesian regression

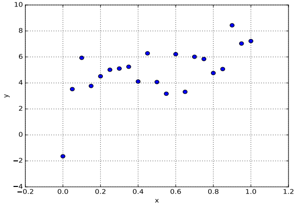


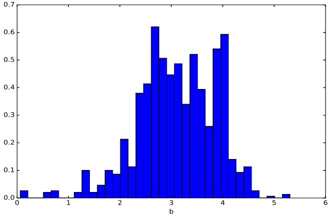


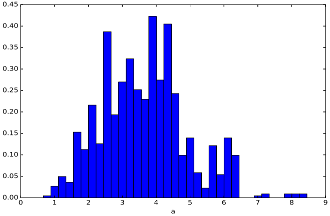


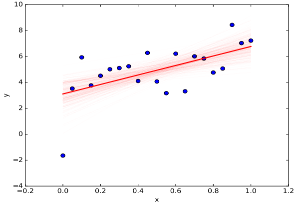
- Suppose we observe data as shown.
- We believe that the data can be explained as $y = ax + b + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$.
- We believe that $a \sim \mathcal{N}(0, 100), b \sim \mathcal{N}(0, 100), \tau \sim Gamma(0.1, 0.1)$.
- Suppose we use the data to obtain the posteriors on a, b, and τ. These posteriors are shown.

Bayesian regression









- Suppose we observe data as shown.
- We believe that the data can be explained as $y = ax + b + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \tau)$.
- We believe that $a \sim \mathcal{N}(0, 100), b \sim \mathcal{N}(0, 100), \tau \sim Gamma(0.1, 0.1)$.
- Suppose we use the data to obtain the posteriors on a, b, and τ. These posteriors are shown.

What have we seen?

- Introduction to Bayesian approach
- Probability for Bayesian inference - Bayes rule
- Examples of Bayesian inference

Computing the posterior

$$\mathsf{Pr}\left\{\Theta|Y=y\right\} = \frac{\mathsf{Pr}\left\{\Theta\right\}\mathsf{Pr}\left\{Y=y|\Theta\right\}}{\mathsf{Pr}\left\{Y=y\right\}}$$

- Computing the term in the denominator is hard!

Approximating the prior and likelihood

- The prior and likelihood functions are chosen such that the posterior distribution is known in closed form.
- Furthermore, the posterior distribution is "similar" to the prior distribution
- This is useful for updating the posterior for a sequence of observations

Approximating the prior and likelihood

- The prior and likelihood functions are chosen such that the posterior distribution is known in closed form.
- Furthermore, the posterior distribution is "similar" to the prior distribution
- This is useful for updating the posterior for a sequence of observations
- For example, earlier we had the prior $\Pr\{\Theta = 0.4\} = 0.2, \Pr\{\Theta = 0.5\} = 0.6, \Pr\{\Theta = 0.6\} = 0.2$
- The likelihood $\Pr\{Y = y|\Theta\} = \Theta^y (1-\Theta)^{10-y}$
- Suppose the posterior $\Pr\{\Theta|Y = y\}$ needs to have the same mathematical form as the likelihood
- Choose $\Pr\{\Theta\} \propto \Theta^a (1-\Theta)^b$
- In fact, choose $\Pr\{\Theta\} = \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{B(a,b)}$ (actually, this is the PDF $f_\Theta(.)$ of a Beta distribution)

Approximating the prior and likelihood

- The prior and likelihood functions are chosen such that the posterior distribution is known in closed form.
- Furthermore, the posterior distribution is "similar" to the prior distribution
- This is useful for updating the posterior for a sequence of observations
- For example, earlier we had the prior $\Pr\{\Theta = 0.4\} = 0.2, \Pr\{\Theta = 0.5\} = 0.6, \Pr\{\Theta = 0.6\} = 0.2$
- The likelihood $\Pr\{Y = y|\Theta\} = \Theta^y(1-\Theta)^{10-y}$
- Suppose the posterior $\Pr\{\Theta|Y = y\}$ needs to have the same mathematical form as the likelihood
- Choose $\Pr\{\Theta\} \propto \Theta^a(1-\Theta)^b$
- In fact, choose $\Pr\{\Theta\} = \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{B(a,b)}$ (actually, this is the PDF $f_\Theta(.)$ of a Beta distribution)
- $\Pr\{\Theta|Y = y\} \propto \Theta^{a-1+y}(1-\Theta)^{b+9-y}$

Approximating the prior and likelihood

- The prior and likelihood functions are chosen such that the posterior distribution is known in closed form.
- Furthermore, the posterior distribution is "similar" to the prior distribution
- This is useful for updating the posterior for a sequence of observations
- For example, earlier we had the prior $\Pr\{\Theta = 0.4\} = 0.2, \Pr\{\Theta = 0.5\} = 0.6, \Pr\{\Theta = 0.6\} = 0.2$
- The likelihood $\Pr\{Y = y|\Theta\} = \Theta^y(1-\Theta)^{10-y}$
- Suppose the posterior $\Pr\{\Theta|Y = y\}$ needs to have the same mathematical form as the likelihood
- Choose $\Pr\{\Theta\} \propto \Theta^a(1-\Theta)^b$
- In fact, choose $\Pr\{\Theta\} = \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{B(a,b)}$ (actually, this is the PDF $f_\Theta(.)$ of a Beta distribution)
- $\Pr\{\Theta|Y = y\} \propto \Theta^{a-1+y}(1-\Theta)^{b+9-y}$
- Then $\Pr\{\Theta|Y = y\} = \frac{\Theta^{a-1+y}(1-\Theta)^{b+9-y}}{B(a-1+y,b+9-y)}$

Samples from the posterior distribution

$$\mathsf{Pr}\{\Theta|Y=y\} = \frac{\mathsf{Pr}\{\Theta\}\,\mathsf{Pr}\{Y=y|\Theta\}}{\mathsf{Pr}\{Y=y\}}$$

$$\mathsf{Pr}\{\Theta|Y=y\} \propto \mathsf{Pr}\{\Theta\}\,\mathsf{Pr}\{Y=y|\Theta\}$$

- We do not know what the normalized posterior distribution is!
- Suppose we have a mechanism for generating samples using the unnormalized posterior distribution
- The empirical distribution of the samples closely approximates the normalized posterior distribution.
- Markov chain Monte Carlo is a technique for obtaining samples from the unnormalized posterior distribution.

Sampling from a distribution

- You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

Sampling from a distribution

- You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

- rand() gives you a number uniformly distributed between 0 and 1

Sampling from a distribution

- You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

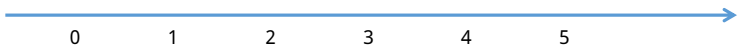
- rand() gives you a number uniformly distributed between 0 and 1

```python
import numpy as np

CDF = [0,0.1,0.3,0.5,0.6,1]

def sampleFromCDF():
    unifSample = np.random.rand()
    for i in [0,1,2,3,4]:
        if (unifSample > CDF[i] and
            unifSample <= CDF[i + 1]):
            return i + 1
```

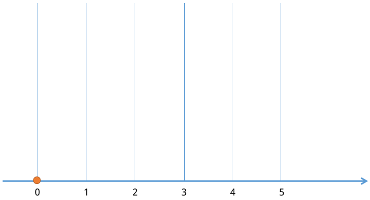
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

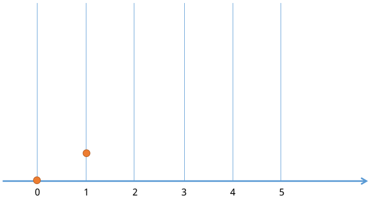
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

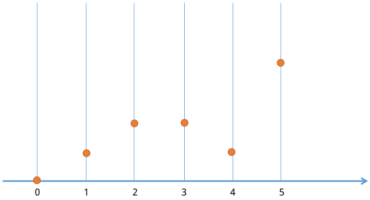
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

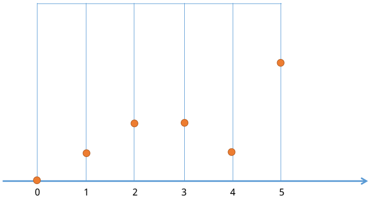
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

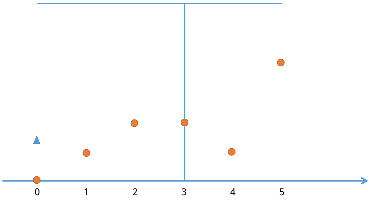
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

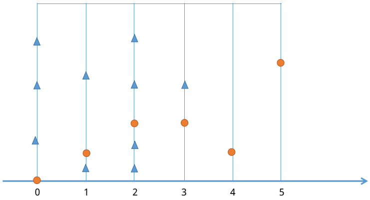
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

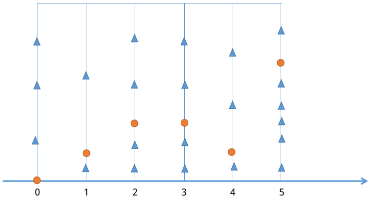
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

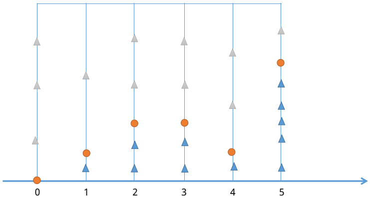
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

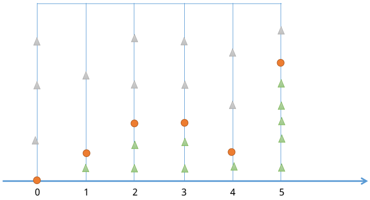
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

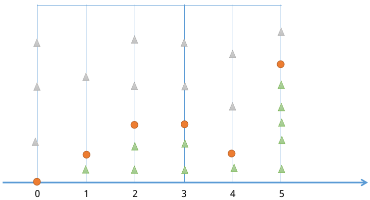
Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

Acceptance-Rejection sampling



You want to generate samples from the given distribution

x	0	1	2	3	4	5
PMF	0	0.1	0.2	0.2	0.1	0.4
CDF	0	0.1	0.3	0.5	0.6	1

- Let the number of accepted points be $N(x)$ and total points be N
- Then $N(x) \approx \frac{1}{6}N \times \Pr\{X = x\}$ for every large N
- Empirical probability or $\frac{N(x)}{N}$ is $\Pr\{X = x\}$
- Note that even if $\Pr\{X = x\}$ is not normalized, this would work! How?

Acceptance-Rejection sampling

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- Assume that $P_X(x)$ is known only upto a normalizing constant

Acceptance-Rejection sampling

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- Assume that $P_X(x)$ is known only upto a normalizing constant
- Suppose there is a distribution $Q_X(x)$ such that there is a $M > 0$ for which

$$P_X(x) \leq MQ_X(x)$$

Acceptance-Rejection sampling

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- Assume that $P_X(x)$ is known only upto a normalizing constant
- Suppose there is a distribution $Q_X(x)$ such that there is a $M > 0$ for which

$$P_X(x) \leq MQ_X(x)$$

- Sample $x \sim Q_X(.)$ and $u \sim$ Uniform$[0, 1]$
- If $u < \frac{P_X(x)}{MQ_X(x)}$ accept x, else reject x

Acceptance-Rejection sampling

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- Assume that $P_X(x)$ is known only upto a normalizing constant
- Suppose there is a distribution $Q_X(x)$ such that there is a $M > 0$ for which

$$P_X(x) \leq MQ_X(x)$$

- Sample $x \sim Q_X(.)$ and $u \sim$ Uniform$[0, 1]$
- If $u < \frac{P_X(x)}{MQ_X(x)}$ accept x, else reject x
- Acceptance rate is $\frac{1}{M}$
- Low acceptance rate, inefficient especially in high dimensions

Metropolis-Hastings sampler

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- But $P_X(x)$ is known only upto a normalizing constant

Metropolis-Hastings sampler

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- But $P_X(x)$ is known only upto a normalizing constant
- Suppose we pick an initial point x_0 in the domain of P_X
- We sample a x_1 according to a proposal distribution $Q_{X|X'}(x|x')$
- We accept x_1 with the probability $\min\left(1, \frac{P_X(x_1)Q_{X|X'}(x_0|x_1)}{P_X(x_0)Q_{X|X'}(x_1|x_0)}\right)$.

Metropolis-Hastings sampler

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- But $P_X(x)$ is known only upto a normalizing constant
- Suppose we pick an initial point x_0 in the domain of P_X
- We sample a x_1 according to a proposal distribution $Q_{X|X'}(x|x')$
- We accept x_1 with the probability $\min\left(1, \frac{P_X(x_1)Q_{X|X'}(x_0|x_1)}{P_X(x_0)Q_{X|X'}(x_1|x_0)}\right)$.
- We sample a x_2 according to a proposal distribution $Q_{X|X'}(x|x')$
- We accept x_2 with the probability $\min\left(1, \frac{P_X(x_2)Q_{X|X'}(x_1|x_0)}{P_X(x_1)Q_{X|X'}(x_2|x_1)}\right)$.

Metropolis-Hastings sampler

- Suppose there is a distribution $P_X(x)$ that you want to draw samples from
- But $P_X(x)$ is known only upto a normalizing constant
- Suppose we pick an initial point x_0 in the domain of P_X
- We sample a x_1 according to a proposal distribution $Q_{X|X'}(x|x')$
- We accept x_1 with the probability $\min\left(1, \frac{P_X(x_1)Q_{X|X'}(x_0|x_1)}{P_X(x_0)Q_{X|X'}(x_1|x_0)}\right)$.
- We sample a x_2 according to a proposal distribution $Q_{X|X'}(x|x')$
- We accept x_2 with the probability $\min\left(1, \frac{P_X(x_2)Q_{X|X'}(x_1|x_0)}{P_X(x_1)Q_{X|X'}(x_2|x_1)}\right)$.
- $\ldots$

How to do Bayesian voltage measurement?

```python
import pymc
import numpy as np
import matplotlib.pyplot as plt

number_measurements = 200;
y = np.random.normal(5, 1, number_measurements)

# Prior
mu = pymc.Uniform('mu', lower = 4.5, upper = 5.5)
# Likelihood
y_obs = pymc.Normal('y_obs', mu = mu, tau = 1, value = y, observed = True)
# Inference
m = pymc.Model([mu, y])
mc = pymc.MCMC(m)
mc.sample(iter=15000, burn=10000)

# Posterior
plt.hist(mu.trace(), 25, normed=True, label='post');
```

How to find the bias of a coin?

```
n = 100
h = 70
alpha = 1
beta = 3

p = pymc.Beta('p', alpha=alpha, beta=beta)
y = pymc.Binomial('y', n=n, p=p, value=h, observed=True)
m = pymc.Model([p, y])

mc = pymc.MCMC(m, )
mc.sample(iter=11000, burn=10000)
plt.hist(p.trace(), 25, normed=True, label='post');
plt.xlabel('Theta')
plt.show()
```

How to do Bayesian regression?

```
# Generating sampled observed data
n = 21
a = 6
b = 2
sigma = 2
x = np.linspace(0, 1, n)
y_obs = a*x + b + np.random.normal(0, sigma, n)
data = pd.DataFrame(np.array([x, y_obs]).T, columns=['x', 'y'])
data.plot(x = 'x', y = 'y', kind = 'scatter', s = 50);
plt.grid()

# Define priors
a = pymc.Normal('slope', mu=0, tau=1.0/10**2)
b = pymc.Normal('intercept', mu=0, tau=1.0/10**2)
tau = pymc.Gamma("tau", alpha=0.1, beta=0.1)

# Define likelihood
@pymc.deterministic
def mu(a=a, b=b, x=x):
    return a*x + b

y = pymc.Normal('y', mu=mu, tau=tau, value=y_obs, observed=True)

# Inference
m = pymc.Model([a, b, tau, x, y])
mc = pymc.MCMC(m)
mc.sample(iter=11000, burn=10000)

# Posterior
abar = a.stats()['mean']
bbar = b.stats()['mean']
data.plot(x='x', y='y', kind='scatter', s=50);
xp = np.array([x.min(), x.max()])
plt.plot(a.trace()*xp[:, None] + b.trace(), c='red', alpha=0.01)
plt.plot(xp, abar*xp + bbar, linewidth=2, c='red');
plt.show()
```

Revisiting conditional independence

- Conditional independence for events
  - A, B, C are three events, which are subsets of Ω and elements of $\mathcal{F}$
  - The events A and B are said to be independent if

$$\Pr(AB) = \Pr(A)\Pr(B)$$

  - The events A and B are said to be conditionally independent given the event C if

$$\Pr(AB|C) = \Pr(A|C)\Pr(B|C)$$

- Conditional independence for discrete random variables
  - Let us consider discrete random variables X, Y, and Z.
  - X and Y are conditionally independent given Z if

$$\Pr(X = x, Y = y|Z = z) = \Pr(X = x|Z = z)\Pr(Y = y|Z = z)$$

    for every x, y, z
  - The joint conditional probability distribution factors into the product of the individual conditional probability distributions

Discrete time Markov Chains

- We are modelling a system evolution in discrete time
- The state space is assumed to be discrete: $\mathcal{S} = \{0, 1, 2, 3, \ldots, s\}$
- The system evolution

  $$(X_0, X_1, X_2, X_3, \ldots, X_n, \ldots)$$

  is a discrete time Markov chain (DTMC) iff

  $$\Pr\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots\} = \Pr\{X_{n+1} = j | X_n = i\}$$

- Note that the LHS contains three parameters - n, i, j and is denoted by $p_{i,j}(n)$
- The probability $p_{i,j}(n)$ is the transition probability of the Markov chain from state i to state j at time n.
- The above conditional independence property is called the Markov property.
- The Markov property says that given the present the future probabilistic evolution of the random process is independent of the past

Exercise - I

- Suppose I take a coin and toss it continuously. Each toss is independent of any other toss.
- Whenever I see a heads on a coin toss I get Re. 1
- Let the probability of getting a heads be p. Suppose p does not change with the tosses.
- Let the amount that I earn on the n^{th} coin toss be X_n
- Then $\Pr\{X_n = 1\} = p$ and $\Pr\{X_n = 1\} = 1 - p$
- Consider

$$(X_1, X_2, X_3, \ldots, X_n, \ldots)$$

- Is the above random process a Markov chain?

Exercise - I

- Suppose I take a coin and toss it continuously. Each toss is independent of any other toss.
- Whenever I see a heads on a coin toss I get Re. 1
- Let the probability of getting a heads be p. Suppose p does not change with the tosses.
- Let the amount that I earn on the n^{th} coin toss be X_n
- Then $\Pr\{X_n = 1\} = p$ and $\Pr\{X_n = 1\} = 1 - p$
- Consider
$$(X_1, X_2, X_3, \ldots, X_n, \ldots)$$
- Is the above random process a Markov chain?
- Yes!
- $\Pr\{X_n = 1 | X_{n-1} = i, X_{n-2} = i_{n-2}, \ldots\} = p.$
- Any IID process is Markov!

Exercise - II

- Let us continue with the coin tossing experiment in the previous slide
- Consider
$$(X_1, X_2, X_3, \ldots, X_n, \ldots)$$
  as before
- Now let $Y_n = \sum_{k=1}^{n} X_k$.
- Consider
$$(Y_1, Y_2, Y_3, \ldots, Y_n, \ldots)$$
- Is the above process Markov?

Exercise - II

- Let us continue with the coin tossing experiment in the previous slide
- Consider
$$(X_1, X_2, X_3, \ldots, X_n, \ldots)$$
  as before
- Now let $Y_n = \sum_{k=1}^{n} X_k$.
- Consider
$$(Y_1, Y_2, Y_3, \ldots, Y_n, \ldots)$$
- Is the above process Markov?
- Yes!
- What all values can Y_n take?
- $\Pr\{Y_n = j | Y_{n-1} = i, Y_{n-2} = i_{n-2}, \ldots\} = \Pr\{X_n = j - i | Y_{n-1} = i\}$
- The probability on the RHS is non-zero only for $j = i$ or $j = i + 1$

Simulating DTMCs

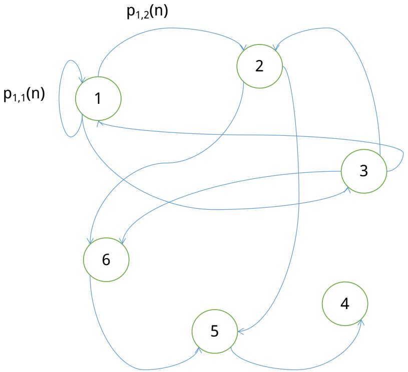
```
transition_probability_matrix = [0.1, 0.1, 0.8;
    0.5, 0.3, 0.2;
    0.4, 0.1, 0.5];
initial_state = 1;
current_state = initial_state;
for i = 1: number_of_simulated_steps
    transition_probability = transition_probability_matrix( current_state , : );
    next_state = sample_from_pmf (transition_probability);
    current_state = next_state;
end
```

Specification of a DTMC model

- Specification of the state space $\mathcal{S}$
- Specification of the transistion probability $p_{i,j}(n)$
- Starting state*

P(n)

$$\begin{pmatrix} p_{1,1}(n) & p_{1,2}(n) & p_{1,s}(n) \\ p_{2,1}(n) & & \\ p_{3,1}(n) & \dots & \\ & & p_{s,s}(n) \end{pmatrix}$$



Homogeneous DTMCs

- A DTMC is said to be (time) homogeneous iff the transition probabilities $p_{i,j}(n)$ do not depend on time, i.e.,

$$p_{i,j}(n) = p_{i,j}$$

- A homogeneous DTMC is then fully represented by its transition probability matrix P, where $[P]_{i,j} = p_{i,j}$
- For a homogeneous DTMC we can talk about n step transition probabilities

$$p_{i,j}^{(n)} = \Pr\{X_n = j | X_0 = i\}$$

  i.e., the probability that the Markov chain will move from i to j in $n \geq 1$ steps.
- We can also talk about an n step transition probability matrix $P^{(n)}$, where $[P^{(n)}]_{i,j} = p_{i,j}^{(n)}$.

State after n steps

- Suppose $X_0 = i$
- We let the DTMC evolve and we are interested in the state after n steps
- Is this a random variable?

State after n steps

- Suppose $X_0 = i$
- We let the DTMC evolve and we are interested in the state after n steps
- Is this a random variable? This is the random variable X_n

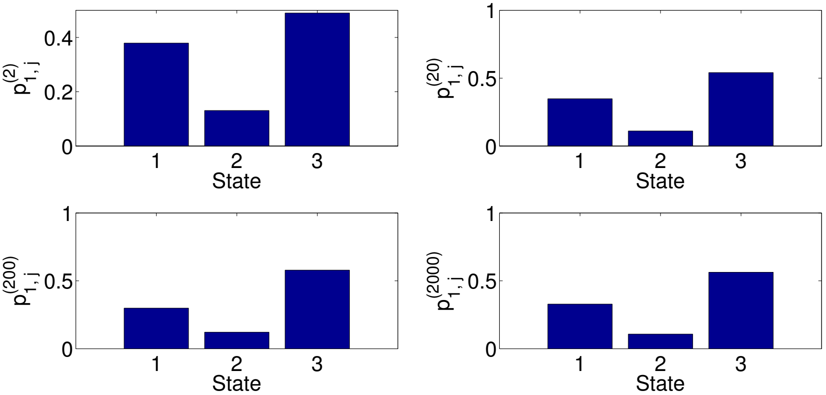
State after n steps

- Suppose $X_0 = i$
- We let the DTMC evolve and we are interested in the state after n steps
- Is this a random variable? This is the random variable X_n
- What is the distribution of X_n?

State after n steps

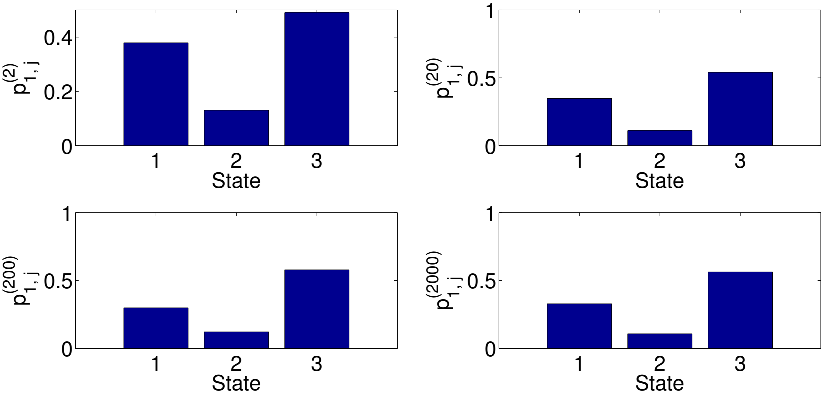
- Suppose $X_0 = i$
- We let the DTMC evolve and we are interested in the state after n steps
- Is this a random variable? This is the random variable X_n
- What is the distribution of X_n? This is the n-step probabilities $p_{i,j}^{(n)}$

Stationary distribution



$$P = \begin{bmatrix} 0.1 & 0.1 & 0.8 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}$$

Stationary distribution



$$P = \begin{bmatrix} 0.1 & 0.1 & 0.8 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}$$

- $\lim_{n\to\infty} p^{(n)}_{i,j} = \pi_j$
- A solution to $\pi = \pi P$ is $\pi = [0.3173, 0.1250, 0.5577]$

Markov Chain Monte Carlo Samplers

- A Markov Chain Monte Carlo (MCMC) sampler is basically a Markov chain with the stationary distribution being the posterior that we want to sample from.

References

- MIT Opencourseware - 6.041 (notes, exercises, video lectures)
- Introduction to Probability 2nd Ed. (Athena Scientific) - Bertsekas and Tsitsiklis
- An exploration of random processes for scientists and engineers - Bruce Hajek
- John K. Kruschke - Doing Bayesian Data Analysis - Academic Press (2010)
- Andrieu et al. - An Introduction to MCMC for Machine Learning, Machine Learning (50), 2003

Thank you!

Contact: vineethbs@gmail.com