Information Theory and Probabilistic Programming

Samuel Cheng

School of ECE
University of Oklahoma

September 13, 2023

Lecture 1: Overview and review of probabilities

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?
2. Statistical inference
   - Bayesian and Monte Carlo techniques

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?
2. Statistical inference
   - Bayesian and Monte Carlo techniques
3. Introduction of probabilistic programming
   - Solve inference problems with programming

About this course

1. Learn some basic information theory (what is it? how is it useful?)
   - Understand basic terminology: what is entropy all about?
2. Statistical inference
   - Bayesian and Monte Carlo techniques
3. Introduction of probabilistic programming
   - Solve inference problems with programming
4. Get better understanding of probability

What is information theory?

- Study of "information" using probability

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability

What is information theory?

- Study of “information” using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science

What is information theory?

- Study of “information” using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science
  - The theoretical basis of the entire telecom industry is built on top of that

What is information theory?

- Study of "information" using probability
- Can be treated as a subfield of applied probability
- But it has a huge impact to communications and information science
  - The theoretical basis of the entire telecom industry is built on top of that
  - Study of extreme cases. What is possible and what is not?

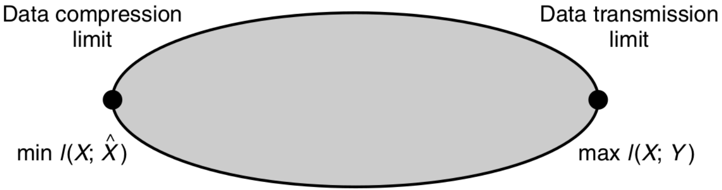


FIGURE 1.2. Information theory as the extreme points of communication theory.

(From Cover and Thomas)

Connection to other fields

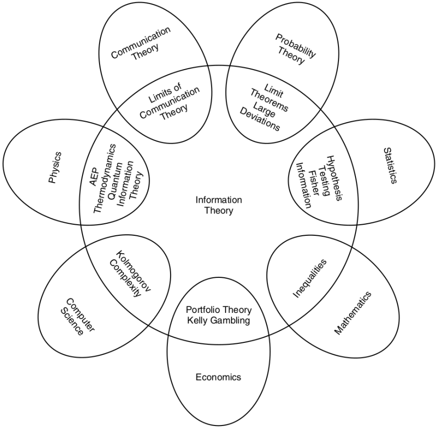


FIGURE 1.1. Relationship of information theory to other fields.

(From Cover and Thomas)

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
  - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
  - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity
    - Give the capacity of Gaussian channel as an example

Shannon's paper

- A majority of the ideas are nicely included in Claude Shannon's seminal paper "A Mathematical Theory of Communication" in the Bell System Technical Journal in July and October 1948
  - There he introduced the idea of an information measure (entropy) to quantify the amount of "information" in a source
    - Introduced the term "bit" (concept of bits dated way back to at least to the 1800 century though) as a unit for the measure
    - As a consequence, it is impossible to compress a source to a size smaller than its entropy and yet recover it losslessly
  - Argue that there is a (capacity) limit of lossless communication under a noisy channel and theoretically we can have lossless communications as long as smaller than the capacity
    - Give the capacity of Gaussian channel as an example
- Some similar ideas were explored earlier in Bell Labs by Harry Nyquist and Ralph Hartley. But those results are limited to events with equal probability

What is “information” in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome

What is “information” in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the “knowledge” gained you have knowing that piece of information

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
    - More information if the outcomes of the event are less predictable

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
    - More information if the outcomes of the event are less predictable
    - Entropy is a measure of **uncertainty**

What is “information” in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the “knowledge” gained you have knowing that piece of information
  - More information if the outcomes of the event are less predictable
  - Entropy is a measure of **uncertainty**
- A Preview:

$$H(X) = \sum_x p(x) \underbrace{H(X=x)}_{\text{info revealed when } X=x}$$

What is "information" in information theory?

- Consider a probabilistic event with uncertain outcomes. Information is the knowledge of the final outcome
- The amount of information can be considered as the "knowledge" gained you have knowing that piece of information
  - More information if the outcomes of the event are less predictable
  - Entropy is a measure of **uncertainty**
- A Preview:

$$H(X) = \sum_x p(x) \underbrace{H(X=x)}_{\text{info revealed when } X=x}$$

A good guess for $H(X=x)$: $\log \frac{1}{p(x)}$

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it
- Nice philosophically but doesn't go much anywhere

Computer scientists' treatment

- Kolmogorov complexity (algorithm information theory): quantify a piece of information as the size of smallest program describing it
- Nice philosophically but doesn't go much anywhere
- We will take the probabilistic view (electrical/communication engineers treatment here) to quantify information theory who usually study with Bayesian models

Neumann-Shannon Anecdote

When Shannon discovered this function he was faced with the need to name it, for it occurred quite often in the theory of communication he was developing. He considered naming it “information” but felt that this word had unfortunate popular interpretations that would interfere with his intended uses of it in the new theory. He was inclined towards naming it “uncertainty” and discussed the matter with the late John Von Neumann. Von Neumann suggested that the function ought to be called “entropy” since it was already in use in some treatises on statistical thermodynamics (e.g. ref. 12). Von Neumann, Shannon reports, suggested that there were two good reasons for calling the function “entropy”. ”It is already in use under that name,” he is reported to have said, ”and besides, it will give you a great edge in debates because nobody really knows what entropy is anyway.” Shannon called the function “entropy” and used it as a measure of “uncertainty,” interchanging the two words in his writings without discrimination.
–From wikipedia

Probability model

- A probability model is used to model uncertain event that can have non-deterministic outcomes
- A probability model can have finite or infinite number of outcomes and even continuous outcomes
- We call the "undetermine" random variable, short for r.v.
- The probability of an **outcome** is the relative chance of getting that outcome
  - For outcome a, we may denote as $Pr(X = a)$ or $p_X(a)$ or even $p(a)$ when it is understood that we are considering variable X
  - $0 \leq p(a) \leq 1$
- We often denote a r.v. using upper case (such as X) and its realization (what was actually observed) using lower case (such as x)

Some probability properties

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$

Some probability properties

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$

Some probability properties

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$

Some probability properties

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x, y) = p(y)$
- Conditional probability (Bayes' rule): $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$

Some probability properties

- Probability mass function (pmf) for discrete random variable (r.v.) X
  - $p(x) \geq 0$
  - $p(x) \leq 1$
  - $\sum_x p(x) = 1$
- Probability density function (pdf) for continuous r.v. X
  - $p(x) \geq 0$
  - $p(x)$ can be larger than 1
  - $Pr(a \leq X \leq b) = \int_a^b p(x)$ (Area between $p(x)$ and x-axis)
  - $\int_x p(x) = 1$
- Marginalization: $\sum_x p(x,y) = p(y)$
- Conditional probability (Bayes' rule): $p(x|y) = \frac{p(x,y)}{p(y)}$
  - N.B. $\sum_x p(x|y) = 1$ but $\sum_y p(x|y) \neq 1$
- Chain rule: $p(x,y,z) = p(x)p(y|x)p(z|x,y)$
  $RHS = p(x)p(y|x)p(z|x,y) = p(x)\frac{p(x,y)}{p(x)}\frac{p(x,y,z)}{p(x,y)} = p(x,y,z) = LHS$

Probabilities and counting

- Six students A, B, C, D, E, F randomly lined up in a row, what is the probability that the order is exactly ABCDEF?
- Six students randomly assigned into two teams (black and white), what is the probability that A,B,C assigned to Team Black and the rest assigned to Team White?

Example: Two jars

- Both Jars A and B have 4 balls
  - Jar A has 1 white and 3 black
  - Jar B has 2 white and 2 black
- Let's draw balls from the jars multiple times. **And put the drawn ball back after each draw.** Can you answer the following?
  - What is the probability of get a white ball from Jar A?
  - What is the probability of getting 3 whites after 6 drawings?
  - If someone randomly pick a jar to draw from and get 3 whites after 6 drawing, what is the probability that he drew from Jar A?

Bayes rule

- Both Jars A and B have 4 balls
  - Jar A has 1 white and 3 black
  - Jar B has 2 white and 2 black
- Say probability of picking Jar A, $Pr(Jar = A) = 0.5$
  - What is the probability of picking from Jar A and getting a white ball $Pr(Jar = A, Ball = white)$?
  - What is $Pr(Ball = white|Jar = A)$?
  - What is $Pr(Jar = A|Ball = white)$?

Expectation

- Recall that $p(x)$ as the distribution of a r.v. X
- The expected value of X is $E[X] \triangleq \sum_x x \cdot p(x)$
- In general, the expected value of a function $f(\cdot)$ of X is $E[f(X)] \triangleq \sum_x f(x) \cdot p(x)$
- Examples
  - $E[X]$ is just the mean of X, often denote as $\overline{X}$
  - The variance of X is $E[(X - \overline{X})^2]$

Independence and conditional independence

- Independence: $p(x,y)=p(x)p(y)$, $X \perp\!\!\!\perp Y$
  - By chain rule, $p(x,y)=p(x)p(y|x)$. Therefore the condition implies that $p(y|x)=p(y)$. In other words, no matter what value X takes, the probability of Y given X is not going to change. So reasonably, they are independent

Independence and conditional independence

- Independence: $p(x,y)=p(x)p(y)$, $X \perp\!\!\!\perp Y$
  - By chain rule, $p(x,y)=p(x)p(y|x)$. Therefore the condition implies that $p(y|x)=p(y)$. In other words, no matter what value X takes, the probability of Y given X is not going to change. So reasonably, they are independent
- Markov property and conditional independence: $p(x,y|z)=p(x|z)p(y|z)$, $X \perp\!\!\!\perp Y|Z, X \leftrightarrow Z \leftrightarrow Y$
  - Similar to independence, by chain rule, we have $p(x,y|z)=p(x|z)p(y|x,z)$. Along with the above condition, $p(y|x,z)=p(y|z)$. Thus given Z, it does not matter what X supposed to be, the probability of given both variables will not depend on X. Hence, X and Y are conditionally independent given Z
- Caveat: independence and conditional independence are two "independent concepts", we can have both satisfied, none of them satisfied, or one of them satisfied. A common **mistake** is to think that independence leads to conditional independence or vice versa. But that is WRONG

Independence but not conditional independence

Consider flipping two coins with outcomes store as X and Y, say 1 represents a head and 0 represents a tail

- In general the two outcomes should be independent (maybe unless if you are some professional/magical gambler), so we have $X \perp\!\!\!\perp Y$
- Now, let $Z = X \oplus Y$, where $\oplus$ is the exclusive or operation $(1 \oplus 0 = 0 \oplus 1 = 1$ and $1 \oplus 1 = 0 \oplus 0 = 0)$
  - Even though $X \perp\!\!\!\perp Y$, $X \not\perp\!\!\!\perp Y|Z$
  - Actually given Z, X "depends" very much on Y since from $X = Y \oplus Z$, we can find out X precisely given Y
  - We can also check the condition $X \perp\!\!\!\perp Y|Z$ by comparing the probability $p(x|z,y)$ with $p(x|z)$
    - For example, $p_{X|Z}(0|0) = 0.5 \neq 1 = p_{X|Z,Y}(0|0,0)$. Thus $X \perp\!\!\!\perp Y|Z$ cannot be true

A digression: Naive Bayes Algorithm

- Naive Bayes is a simple machine learning algorithm to classify an object with its features
- Basically, we are simply assuming the features are conditionally independent given the object class
- Say if O is the object that $c(O)$ is the corresponding class (can be $c_1, c_2, \cdots$). And say $f_1(O), f_2(O), \cdots, f_K(O)$ are K features of the object
  - For simplicity, let's rewrite $c(O)$ as C and $f_i(O)$ as F_i. But it is important to realize that the "randomness" of $c(O), f_i(O)$ is originated from O

$$
\begin{aligned}
p(c|f_1,\cdots,f_K) &= \frac{p(c,f_1,\cdots,f_K)}{p(f_1,\cdots,f_K)} = \frac{p(c)p(f_1,\cdots,f_K|c)}{p(f_1,\cdots,f_K)} && \text{Bayes' rule} \\
&= \frac{p(c)p(f_1|c)\cdots p(f_K|c)}{p(f_1,\cdots,f_K)} && \text{Assume } F_i \perp\!\!\!\perp F_j|C \\
&= \frac{p(c)p(f_1|c)\cdots p(f_K|c)}{p(f_1)\cdots p(f_K)} && \text{If also assume } F_i \perp\!\!\!\perp F_j \\
&= p(c)\frac{p(f_1|c)}{p(f_1)}\cdots\frac{p(f_K|c)}{p(f_K)}
\end{aligned}
$$

A digression: Naive Bayes Algorithm

- In most classification problem, we are interested to compute the most likely class. So we really will go through all possible $c_1, c_2, \cdots$ for $p(c|f_1, \cdots, f_K)$
- Rather than assuming both $F_i \perp\!\!\!\perp F_j|C$ and $F_i \perp\!\!\!\perp F_j$, the latter really is not necessary as we can write

$$p(c|f_1, \cdots, f_K) = \frac{p(c)p(f_1|c) \cdots p(f_K|c)}{\sum_i p(c_i)p(f_1|c_i) \cdots p(f_K|c_i)}$$

  Actually if we only care about which is the most likely class, we can even skip computing the denominator as it is a constant w.r.t. c
- You can find a numerical example here
  - N.B. the author assumes independence of the features in his explanation but the condition is not necessary as noted above

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed
2. Identify distributions and conditions (independence, conditional independence, variable relationship)

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed
2. Identify distributions and conditions (independence, conditional independence, variable relationship)
3. Identify (conditional) probability to address the question

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed
2. Identify distributions and conditions (independence, conditional independence, variable relationship)
3. Identify (conditional) probability to address the question
4. Insert dummy variables to probability to leverage conditional independence by marginalization

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed
2. Identify distributions and conditions (independence, conditional independence, variable relationship)
3. Identify (conditional) probability to address the question
4. Insert dummy variables to probability to leverage conditional independence by marginalization
5. Expand probabilities into (conditional) probabilities and evaluate them

Epilogue: an engineer (dummy) approach to solve probability problems

1. Introduce helper variables if needed
2. Identify distributions and conditions (independence, conditional independence, variable relationship)
3. Identify (conditional) probability to address the question
4. Insert dummy variables to probability to leverage conditional independence by marginalization
5. Expand probabilities into (conditional) probabilities and evaluate them
6. Compute sum/integral

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
    - Let's denote O as the other door both guest and host did not pick

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
    - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
    - $P \perp\!\!\!\perp G, O = \{1, 2, 3\} \setminus \{G, H\}, p(G) = p(H) = \frac{1}{3}$, etc.

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization
   - $p(O_1|P_1) = \sum_{i,j} p(O_1, G_i, H_j|P_1)$

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization
   - $p(O_1|P_1) = \sum_{i,j} p(O_1, G_i, H_j|P_1)$
5. Expand probabilities into (conditional) probabilities and evaluate them

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization
   - $p(O_1|P_1) = \sum_{i,j} p(O_1, G_i, H_j|P_1)$
5. Expand probabilities into (conditional) probabilities and evaluate them
   - $p(O_1|P_1) = \sum_{i,j} p(G_i|P_1)p(H_j|P_1, G_i)p(O_1|G_i, H_j, P_1)$
     $= p(G_1)(p(H_1|G_1P_1)p(O_1|G_1H_1P_1) + p(H_2|G_1P_1)p(O_1|G_1H_2P_1) + p(H_3|G_1P_1)p(O_1|G_1H_3P_1))$
     $+p(G_2)(p(H_1|G_2P_1)p(O_1|G_2H_1P_1) + p(H_2|G_2P_1)p(O_1|G_2H_2P_1) + p(H_3|G_2P_1)p(O_1|G_2H_3P_1))$
     $+p(G_3)(p(H_1|G_3P_1)p(O_1|G_3H_1P_1) + p(H_2|G_3P_1)p(O_1|G_3H_2P_1) + p(H_3|G_3P_1)p(O_1|G_3H_3P_1))$

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G,H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization
   - $p(O_1|P_1) = \sum_{i,j} p(O_1, G_i, H_j|P_1)$
5. Expand probabilities into (conditional) probabilities and evaluate them
   - $p(O_1|P_1) = \sum_{i,j} p(G_i|P_1)p(H_j|P_1,G_i)p(O_1|G_i,H_j,P_1)$
     $= p(G_1)(p(H_1|G_1P_1)p(O_1|G_1H_1P_1) + p(H_2|G_1P_1)p(O_1|G_1H_2P_1) + p(H_3|G_1P_1)p(O_1|G_1H_3P_1))$
     $+p(G_2)(p(H_1|G_2P_1)p(O_1|G_2H_1P_1) + p(H_2|G_2P_1)p(O_1|G_2H_2P_1) + p(H_3|G_2P_1)p(O_1|G_2H_3P_1))$
     $+p(G_3)(p(H_1|G_3P_1)p(O_1|G_3H_1P_1) + p(H_2|G_3P_1)p(O_1|G_3H_2P_1) + p(H_3|G_3P_1)p(O_1|G_3H_3P_1))$
6. Compute sum/integral

Example: Monty Hall problem

Below I will use shorthand like P1, G2, H3 to refer to the case of prize at Door 1, guest picking Door 2, and host open Door 3

1. Introduce helper variables if needed
   - Let's denote O as the other door both guest and host did not pick
2. Identify distributions and condition
   - $P \perp\!\!\!\perp G, O = \{1,2,3\} \setminus \{G, H\}, p(G) = p(H) = \frac{1}{3}$, etc.
3. Identify (conditional) probability to address the question
   - $Pr(Win|switch) = Pr(O = P) = \sum_i p(O_i|P_i)p(P_i) = p(O_1|P_1)$
4. Insert dummy variables to probability by marginalization
   - $p(O_1|P_1) = \sum_{i,j} p(O_1, G_i, H_j|P_1)$
5. Expand probabilities into (conditional) probabilities and evaluate them
   - $p(O_1|P_1) = \sum_{i,j} p(G_i|P_1)p(H_j|P_1, G_i)p(O_1|G_i, H_j, P_1)$
     $= p(G_1)(p(H_1|G_1P_1)p(O_1|G_1H_1P_1) + p(H_2|G_1P_1)p(O_1|G_1H_2P_1) + p(H_3|G_1P_1)p(O_1|G_1H_3P_1))$
     $+p(G_2)(p(H_1|G_2P_1)p(O_1|G_2H_1P_1) + p(H_2|G_2P_1)p(O_1|G_2H_2P_1) + p(H_3|G_2P_1)p(O_1|G_2H_3P_1))$
     $+p(G_3)(p(H_1|G_3P_1)p(O_1|G_3H_1P_1) + p(H_2|G_3P_1)p(O_1|G_3H_2P_1) + p(H_3|G_3P_1)p(O_1|G_3H_3P_1))$
6. Compute sum/integral
   - $p(O_1|P_1) = p(G_2)p(H_3|G_2P_1)p(O_1|G_2H_3P_1) + p(G_3)p(H_2|G_3P_1)p(O_1|G_3H_2P_1) = \frac{1}{3} \cdot 1 \cdot 1 + \frac{1}{3} \cdot 1 \cdot 1 = \frac{2}{3}$

Epilogue: an engineer (dummy) approach to solve probability problems

Our dummy approach can solve virtually solve any probability problems, but

- Identify what variables to introduced may need some experience
- Can solve any problem with only discrete variables, but if there are too many variables, hand calculation not feasible
  $\Rightarrow$ probabilistic programming
- If continuous variables are involved, the last step may involve intractable integral
  $\Rightarrow$ probabilistic programming

Monte Carlo approach

- Our dummy approach involves some understanding of the problem

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning
  - Estimate winning probability = # wins / 10,000

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning
  - Estimate winning probability = # wins / 10,000
- Of course the computed probability won't be exact

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning
  - Estimate winning probability = # wins / 10,000
- Of course the computed probability won't be exact
  - Probability estimate improves with # simulations

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning
  - Estimate winning probability = # wins / 10,000
- Of course the computed probability won't be exact
  - Probability estimate improves with # simulations
  - Problem solved as long as we know how to simulate one time (if we don't need exact probability)

Monte Carlo approach

- Our dummy approach involves some understanding of the problem
- An even dummier approach is by simulation and counting (require even less understanding) $\Rightarrow$ Monte Carlo
- Take Monte Hall as example again
  - Simulate the game many many times (say 10,000 times)
  - Stick to one strategy, always switch or always stay put
  - Count number of winning
  - Estimate winning probability = # wins / 10,000
- Of course the computed probability won't be exact
  - Probability estimate improves with # simulations
  - Problem solved as long as we know how to simulate one time (if we don't need exact probability)
  - Even simulation can be hard and computation can be an issue
    $\Rightarrow$ Markov Chain Monte Carlo (MCMC)
    We will delay this to much later

Monte Hall simulation

Algorithm 1 Simulate one game instance

$P = randint(3)$
$G = randint(3)$
$\mathcal{H} = \{0, 1, 2\} \setminus \{P, G\}$
if $|\mathcal{H}| = 2$ **then**
    $H = \mathcal{H}[randint(2)]$
else
    $H = \mathcal{H}[0]$
end if

More formal treatment: probability space

- More rigorously, a probability model is defined by the **probability space** composed of the triple $(\Omega, \mathcal{F}, p)$
  - Ω is the **sample space** containing all possible outcomes
  - $\mathcal{F}$ is a "σ-field", which is a collection of subsets (events) of Ω
  - p is the (non-negative) **probability measure** on elements of $\mathcal{F}$
- E.g., probability model of unbiased dice
  - $\Omega = \{1, 2, 3, 4, 5, 6\}$
  - $\mathcal{F} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1, 2\}, \{1, 3\}, \cdots, \{1, 2, 3, 4, 5, 6\}\}$
  - $p(S)$ is the probability of an event
    - $p(\{1\}) = p(\{2\}) = p(\{3\}) = p(\{4\}) = p(\{5\}) = p(\{6\}) = 1/6$
    - $p(\{1, 2\}) = p(\{1, 3\}) = \cdots = p(\{5, 6\}) = 2/6$
    - $\cdots$
    - $p(\{1, 2, 3, 4, 5, 6\}) = 1$
- N.B. It could be confusing at first. Be careful that events $\neq$ outcomes. **An event is actually a set of outcomes**

σ-algebra

- The purpose of σ-field (aka σ-algebra) is to impose restriction on what we can and cannot query regarding probability
- Namely, we can only measure the probability of something inside the σ-field $\mathcal{F}$ (i.e., an event)
- Formal definition of σ-field:
  - **σ-field has to satisfied the following: 1) containing empty set $\varnothing$, 2) closed under complement, countable union, and countable intersection of its element**
- E.g., let $\Omega = \{1, 2, 3, 4\}$
  1. $\{\varnothing, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ is a valid σ-field
  2. $\{\varnothing, \{1\}, \{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$ is NOT a valid σ-field
- N.B., A complement, countable union, or countable intersection of Ω is call a **Borel set**
  - $\varnothing, \{1\}, \{1, 2\}$ are example of Borel sets (an event is a Borel set)
  - Collection of all Borel sets forms a σ-algebra (aka Borel (σ-)algebra)

Probability measure

- Probability measure p is a **measure**. Along with $\mathcal{F}$, the tuple $(\mathcal{F}, p)$ forms a **measure space**. For $\mathbb{P}$ to be a valid probability measure, it has to satisfy the following
  - Requirements to be a measure (in the context of measure theory):
    1. $p(\varnothing) = 0$
    2. Countably additive: $p(\cup_{i\in\mathbb{N}} A_i) = \sum_{i\in\mathbb{N}} p(A_i), \forall i \neq j, A_i \cap A_j = \varnothing$
  - And since p is a probability measure, it also has to satisfy $p(\Omega) = 1$
- The above constraints are sometimes known as the axioms of probability theory

Some properties of probability measure

From the axioms described in the last slides, one can show that probability measure has to satisfies the following:

1. $p(A^c) = 1 - p(A)$
2. $p(A) \leq p(B)$ if $A \subset B$
3. Union bound: $p(\cup_i A_i) \leq \sum_i p(A_i)$
   - Proof hint: use 2) and induction
4. Inclusion-exclusion formula:
   $p(\cup_{i=1}^n A_i) = \sum_{i=1}^n p(A_i) - \sum_{i<j} p(A_i \cap A_j) + \sum_{i<j<k} p(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} p(\cap_{i=1}^n A_i)$
   - Proof hint: show $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ and then use induction. ($p(A \cup B) = p(A) + p(B \setminus A)$ and $p(B) = p(A \cap B) + p(B \setminus A)$).

Why so complex?

- Consider X a uniform random variable defined between $[0,1]$
- Define $Y = \begin{cases} 1 & \text{if } X \text{ is rational} \\ 0 & \text{otherwise} \end{cases}$
- Y is a random variable since X is random. It is reasonable to ask what is the probability that $Y = 1$. From undergrad probability class,

$$Pr(Y=1) = \int_{\{x|x\in[0,1]\cap\mathbb{Q}\}} dx = ?$$

  - The integral above is actually undefined according to undergrad calculus, where the integral is known as a Riemann integral
- Instead, we have to incorporate the idea of "measure" (Lesbeque integral)

$$Pr(Y=1) = \int_{\{x|x\in[0,1]\cap\mathbb{Q}\}} dp(x) = 0$$

  - The Lesbeque integral above is 0 since the measure of $\{x|x \in [0,1] \cap \mathbb{Q}\} = 0$

Some remarks on notation

- In general, we can write

$$p(\Omega') = \int_{\Omega'} dp(\omega)$$

and

$$E[f(X)] = \int_{\Omega} f(X(\omega))dp(\omega)$$

- E.g.,

$$E[X] = \int_{\Omega} X(\omega)dp(\omega) = \int_{\Omega} X(\omega)\ dp = \int_{\Omega} X dp$$

  - Note that p is the probability measure (often people use upper case P instead)
  - People often omit ω as above when context is clear

Lecture 2: ML, MAP, and Bayesian estimation

Inference

o: Observed variable, θ: Parameter, x: Latent variable

Maximum Likelihood (ML)

$\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(o|\theta)$

Maximum A Posteriori (MAP)

$\hat{x} = \arg\max_x p(x|\hat{\theta}), \hat{\theta} = \arg\max_\theta p(\theta|o)$

Bayesian

$$\hat{x} = \sum_x x \underbrace{\sum_\theta p(x|\theta)p(\theta|o)}_{p(x|o)}$$

where $p(\theta|o) = \frac{p(o|\theta)p(\theta)}{p(o)} \propto p(o|\theta)\underbrace{p(\theta)}_{prior}$

Coin Flip

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Prior: Probability of a hypothesis before we make any observations

(Slide credit: University of Washington CSE473)

Coin Flip

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Uniform Prior: All hypothesis are equally likely before we make any observations

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)}$$

$$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.6$

Posterior: Probability of a hypothesis given data

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin:	Best estimate for P(H)
C_2 	$P(H\|C_2) = 0.5$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Your Estimate?

Maximum Likelihood Estimate: The best hypothesis that fits observed data assuming uniform prior

Most likely coin: C_2

Best estimate for P(H)

$P(H|C_2) = 0.5$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

(Slide credit: University of Washington CSE473)

Using Prior Knowledge

- Should we always use Uniform Prior?
- Background knowledge:
  - Heads => you go first in Abalone against TA
  - TAs are nice people
  - => TA is more likely to use a coin biased in your favor

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Using Prior Knowledge

We can encode it in the prior:

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin <u>did</u> I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \alpha P(H|C_1)P(C_1)$$

C_1	C_2	C_3
		
$P(H|C_1) = 0.1$	$P(H|C_2) = 0.5$	$P(H|C_3) = 0.9$
$P(C_1) = 0.05$	$P(C_2) = 0.25$	$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.006$ $P(C_2|H) = 0.165$ $P(C_3|H) = 0.829$

ML posterior after Exp 1:

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.600$

C_1	C_2	C_3
		
$P(H\|C_1) = 0.1$	$P(H\|C_2) = 0.5$	$P(H\|C_3) = 0.9$
$P(C_1) = 0.05$	$P(C_2) = 0.25$	$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1 C_2 C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin <u>did</u> I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Experiment 2: Tails

Which coin <u>did</u> I use?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin:

C_3

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Your Estimate?

Maximum A Posteriori (MAP) Estimate: The best hypothesis that fits observed data assuming a <u>non-uniform prior</u>

Most likely coin:

C_3 

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

(Slide credit: University of Washington CSE473)

Did We Do The Right Thing?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Did We Do The Right Thing?

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_2 and C_3 are almost equally likely

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

A Better Estimate

Recall: $P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$ = 0.680

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data and (generally) assuming a <u>non-uniform prior</u>

$$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) = 0.680$$

$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

(Slide credit: University of Washington CSE473)

Comparison

ML
- Easy to compute

Comparison

ML
- Easy to compute

MAP
- Still relatively easy to compute
- Incorporate prior information

Comparison

ML	• Easy to compute
MAP	• Still relatively easy to compute • Incorporate prior information
Bayesian	• Minimizes expected error $\Rightarrow$ especially shines when little data available • Potentially much harder to compute

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$
- Let's add model type M,
  $p(\theta, o|M) = p(o|M)p(\theta|o, M) = p(\theta|M)p(o|\theta, M)$

Bayes' rule (with model type)

- $p(\theta, o) = p(o)p(\theta|o) = p(\theta)p(o|\theta)$
- Let's add model type M,
  $p(\theta, o|M) = p(o|M)p(\theta|o, M) = p(\theta|M)p(o|\theta, M)$

$$\underbrace{p(\theta|o, M)}_{posterior} = \frac{\overbrace{p(\theta|M)}^{prior}\overbrace{p(o|\theta, M)}^{likelihood}}{\underbrace{p(o|M)}_{model\ evidence}}$$

- M: model type
- θ: model parameter
- o: observation

Lecture 3: Common distributions

Gaussian distribution

- By the Central limit theorem, if we add multiple independent variables together, the sum will become more and more like Gaussian
- Gaussian distribution (aka Normal distribution) has a bell shape
  - It is symmetric w.r.t. mean
  - The mean is also the mode
- The pdf is given by

$$\mathcal{N}(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

  where μ is the mean and σ^2 is the variance

Introduction to Multivariate Gaussian

The probability density function (pdf) of a multivariate Gaussian random variable $\mathbf{X}$ is given by

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

We will also use $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma)$ to denote this pdf.

Symmetry and Other Handy Equations

Note that $\mathbf{x}$ and $\boldsymbol{\mu}$ are symmetric in

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\Sigma)=\mathcal{N}(\boldsymbol{\mu};\mathbf{x},\Sigma)=\mathcal{N}(\boldsymbol{\mu}-\mathbf{x};0,\Sigma)=\mathcal{N}(0;\boldsymbol{\mu}-\mathbf{x},\Sigma).$$

These equations are trivial but very handy at times.

Covariance matrix

Σ can be written as $E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top]$

- Eigenvalues are the variance along the principal axes (directions where variable changes the most)
    - $\therefore$ eigenvalues are real and ≥ 0 in general
    - If we don't assume the degenerate case where the vector variables do not vary in some directions, then all eigenvalues $> 0 \Rightarrow \Sigma^{-1}$ exists

Marginalization of normal distribution

- Consider $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and let say $\mathbf{X}$ is a segment of $\mathbf{Z}$. That is, $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ for some $\mathbf{Y}$. Then how should $\mathbf{X}$ behave?

Marginalization of normal distribution

- Consider $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and let say $\mathbf{X}$ is a segment of $\mathbf{Z}$. That is, $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$ for some $\mathbf{Y}$. Then how should $\mathbf{X}$ behave?
- We can find the pdf of $\mathbf{X}$ by just marginalizing that of $\mathbf{Z}$. That is

$$
\begin{aligned}
p(\mathbf{x}) &= \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int \exp\left(-\frac{1}{2} \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} \\ \mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} \\ \mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}} \end{pmatrix} \right) d\mathbf{y}
\end{aligned}
$$

Marginalization of normal distribution

- Denote Σ^{-1} as Λ (also known as the precision matrix). And partition both Σ and Λ into $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}$ and $\Lambda = \begin{pmatrix} \Lambda_{\mathbf{XX}} & \Lambda_{\mathbf{XY}} \\ \Lambda_{\mathbf{YX}} & \Lambda_{\mathbf{YY}} \end{pmatrix}$

Marginalization of normal distribution

- Denote Σ^{-1} as Λ (also known as the precision matrix). And partition both Σ and Λ into $\Sigma = \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix}$ and $\Lambda = \begin{pmatrix} \Lambda_{\mathbf{XX}} & \Lambda_{\mathbf{XY}} \\ \Lambda_{\mathbf{YX}} & \Lambda_{\mathbf{YY}} \end{pmatrix}$
- Then we have

$$
\begin{aligned}
p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \int \exp\Bigg(-\frac{1}{2} \big[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \\
+ (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Lambda_{\mathbf{YX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \Lambda_{\mathbf{XY}} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \\
+ (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Lambda_{\mathbf{YY}} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \big] \Bigg) d\mathbf{y} \\
= \frac{e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T \Lambda_{\mathbf{XX}} (\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int \exp\Bigg(-\frac{1}{2} \big[(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Lambda_{\mathbf{YX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}) \\
+ (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}})^T \Lambda_{\mathbf{XY}} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) + (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})^T \Lambda_{\mathbf{YY}} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}) \big] \Bigg) d\mathbf{y}
\end{aligned}
$$

Marginalization of normal distribution

To proceed, let's apply the completing square trick on $(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})^T\Lambda_{\mathbf{YX}}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})+(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})+(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})^T\Lambda_{\mathbf{YY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})$. For the ease of exposition, let us denote $\tilde{\mathbf{x}}$ as $\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ as $\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}$. We have

Marginalization of normal distribution

To proceed, let's apply the completing square trick on $(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})^T\Lambda_{\mathbf{YX}}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})+(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})+(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})^T\Lambda_{\mathbf{YY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}})$. For the ease of exposition, let us denote $\tilde{\mathbf{x}}$ as $\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ as $\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}$. We have

$$
\begin{aligned}
&\tilde{\mathbf{y}}^T\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}}+\tilde{\mathbf{x}}^T\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}}+\tilde{\mathbf{y}}^T\Lambda_{\mathbf{YY}}\tilde{\mathbf{y}}\\
=&(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})-\tilde{\mathbf{x}}^T\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}},
\end{aligned}
$$

where we use the fact that $\Lambda=\Sigma^{-1}$ is symmetric and so $\Lambda_{\mathbf{XY}}=\Lambda_{\mathbf{YX}}$

Marginalization of normal distribution

$$p(\mathbf{x}) = \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}} d\mathbf{y}$$

Marginalization of normal distribution

$$p(\mathbf{x}) = \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}}\int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}}d\mathbf{y}$$

$$= \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}\right)$$

Marginalization of normal distribution

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}}\int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}}d\mathbf{y} \\
&= \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}\right) \\
&\overset{(a)}{=} \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right)
\end{aligned}
$$

Marginalization of normal distribution

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}} d\mathbf{y} \\
&= \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}\right) \\
&\overset{(a)}{=} \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right) \\
&\overset{(b)}{=} \frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right)
\end{aligned}
$$

Marginalization of normal distribution

$$
\begin{aligned}
p(\mathbf{x}) &= \frac{e^{-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}}}{\sqrt{\det(2\pi\Sigma)}} \int e^{-\frac{(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})^T\Lambda_{\mathbf{YY}}(\tilde{\mathbf{y}}+\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}\tilde{\mathbf{x}})}{2}} d\mathbf{y} \\
&= \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T(\Lambda_{\mathbf{XX}}-\Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}})\tilde{\mathbf{x}}}{2}\right) \\
&\overset{(a)}{=} \frac{\sqrt{\det(2\pi\Lambda_{\mathbf{YY}}^{-1})}}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right) \\
&\overset{(b)}{=} \frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}} \exp\left(-\frac{\tilde{\mathbf{x}}^T\Sigma_{\mathbf{XX}}^{-1}\tilde{\mathbf{x}}}{2}\right) \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma_{\mathbf{XX}})}} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})^T\Sigma_{\mathbf{XX}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}})}{2}\right),
\end{aligned}
$$

where (a) and (b) will be shown next

(a) $\Sigma_{\mathbf{XX}}^{-1} = \Lambda_{\mathbf{XX}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{YY}}^{-1}\Lambda_{\mathbf{YX}}$

Lemma

Assume $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix}$, *then* $A^{-1} = \tilde{A} - \tilde{B}\tilde{D}^{-1}\tilde{C}$

Proof.

Note that $\begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$. Thus $A\tilde{A} + B\tilde{C} = I$ and $A\tilde{B} + B\tilde{D} = 0$. So

$A(\tilde{A} - \tilde{B}\tilde{D}^{-1}\tilde{C}) = A\tilde{A} - (A\tilde{B})\tilde{D}^{-1}\tilde{C} = A\tilde{A} + B\tilde{D}\tilde{D}^{-1}\tilde{C} = A\tilde{A} + B\tilde{C} = I$ □

(b) $\det(a\Sigma) = \det(a\Sigma_{\mathbf{YY}})\det(a\Lambda_{\mathbf{XX}}^{-1})$

Lemma

Assume $\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{pmatrix}$, *then* $det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = det(D)det(\tilde{A}^{-1})$

Proof.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} A & B \\ D^{-1}C & I \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & B \\ 0 & I \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I \end{pmatrix}$$

$$\Rightarrow det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = det(D)det(A - BD^{-1}C) = det(D)det(\tilde{A}^{-1})$$

□

Remark

N.B. $A - BD^{-1}C$ *is known as Schur complement*

Conditioning mulitivariate Gaussian

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?

Conditioning mulitivariate Gaussian

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?
- Basically, we want to find $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$

Conditioning mulitivariate Gaussian

- Consider the same $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \Sigma_{\mathbf{Z}})$ and $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$. What will $\mathbf{X}$ be like if $\mathbf{Y}$ is observed to be $\mathbf{y}$?
- Basically, we want to find $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{x}, \mathbf{y})/p(\mathbf{y})$
- From previous result, we have $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{YY}})$. Therefore,

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}\left[\begin{pmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{pmatrix} - \tilde{\mathbf{y}}^T \Sigma_{\mathbf{YY}}^{-1} \tilde{\mathbf{y}}\right]\right) \\ &\propto \exp\left(-\frac{1}{2}[\tilde{\mathbf{x}}^T \Lambda_{\mathbf{XX}} \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T \Lambda_{\mathbf{XY}} \tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T \Lambda_{\mathbf{YX}} \tilde{\mathbf{x}}]\right), \end{aligned}$$

where we use $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ as shorthands of $\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}}$ and $\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}$ as before

Conditioning mulitivariate Gaussian

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$
\begin{aligned}
p(\mathbf{x}|\mathbf{y}) &\propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}}+\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T\Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}}+\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)\\
&= \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}+\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}))^T\Lambda_{\mathbf{XX}}\right.\\
&\qquad\left.(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}+\Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{Y}}))\right)
\end{aligned}
$$

Conditioning mulitivariate Gaussian

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T\Lambda_{\mathbf{XX}}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)$$

$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))^T\Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))\right)$$

- Therefore $\mathbf{X}|\mathbf{y}$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\mathbf{X}} - \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})$ and covariance $\Lambda_{\mathbf{XX}}^{-1}$

Conditioning mulitivariate Gaussian

- Completing the square for $\tilde{\mathbf{x}}$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})^T \Lambda_{\mathbf{XX}} (\tilde{\mathbf{x}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}\tilde{\mathbf{y}})\right)$$
$$= \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))^T \Lambda_{\mathbf{XX}} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{X}} + \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}))\right)$$

- Therefore $\mathbf{X}|\mathbf{y}$ is Gaussian distributed with mean $\boldsymbol{\mu}_{\mathbf{X}} - \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}})$ and covariance $\Lambda_{\mathbf{XX}}^{-1}$
- Note that since $\Lambda_{\mathbf{XX}}\Sigma_{\mathbf{XY}} + \Lambda_{\mathbf{XY}}\Sigma_{\mathbf{YY}} = 0 \Rightarrow \Lambda_{\mathbf{XX}}^{-1}\Lambda_{\mathbf{XY}} = -\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}$ and from (a), we have

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}),$$

where $\Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}} \triangleq \Sigma|\Sigma_{\mathbf{YY}}$ is a Schur complement

Conditioning mulitivariate Gaussian

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of $\mathbf{Y}$ is exactly the mean, the conditioned mean does not change

Conditioning mulitivariate Gaussian

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of $\mathbf{Y}$ is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of $\mathbf{Y}$ for the 1-D case.
  - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of $\mathbf{Y}$ to the variation of $\mathbf{X}$

Conditioning mulitivariate Gaussian

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of $\mathbf{Y}$ is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of $\mathbf{Y}$ for the 1-D case.
  - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of $\mathbf{Y}$ to the variation of $\mathbf{X}$
  - In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are negatively correlated, the sign of the adjustment will be reversed

Conditioning mulitivariate Gaussian

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

- When the observation of $\mathbf{Y}$ is exactly the mean, the conditioned mean does not change
- Otherwise, it needs to be modified and the size of the adjustment decreases with $\Sigma_{\mathbf{YY}}$, the variance of $\mathbf{Y}$ for the 1-D case.
  - The observation is less reliable with the increase of $\Sigma_{\mathbf{YY}}$. The adjustment is finally scaled by $\Sigma_{\mathbf{XY}}$, which translates the variation of $\mathbf{Y}$ to the variation of $\mathbf{X}$
  - In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are negatively correlated, the sign of the adjustment will be reversed
- As for the variance of the conditioned variable, it always decreases and the decrease is larger if $\Sigma_{\mathbf{YY}}$ is smaller and $\Sigma_{\mathbf{XY}}$ is larger ($\mathbf{X}$ and $\mathbf{Y}$ are more correlated)

What is a Gaussian Process?

A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$f(x) \sim \mathsf{GP}(m(x), k(x, x'))$$

- $f(x)$ is the function to be modeled.
- $m(x)$ is the mean function, usually zero.
- $k(x, x')$ is the covariance function or kernel.

Advantages and Disadvantages

Advantages:

- Flexible
- Probabilistic Nature
- Non-Parametric

Disadvantages:

- Computational Complexity
- Hyperparameter Sensitivity

Applications

- Regression and function estimation
- Time series forecasting
- Optimization

Uncorrelated implies independence

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{Y}}), \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})$$

If $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated, $\Sigma_{\mathbf{XY}} = 0$. Then

$$\mathbf{X}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{XX}})$$

Note that the statistics of $\mathbf{X}$ does not change with respect to $\mathbf{y}$ and so $\mathbf{X}$ is independent of $\mathbf{Y}$

$X \perp Y | Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Corollary

Given multivariate Gaussian variables X, Y and Z, we have X and Y are conditionally independent given Z if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$, where $\rho_{XZ} = \frac{E[(X-E(X))(Z-E(Z))]}{\sqrt{E[(X-E(X))^2]E[(Z-E(Z))^2]}}$ is the correlation coefficent between X and Z. Similarly, ρ_{YZ} and ρ_{XY} are the correlation coefficients between Y and Z, and X and Y, respectively.

$X \perp Y | Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- From the definition of correlation coefficient, $\Sigma = \begin{pmatrix} \sigma_{XX} & \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} \\ \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sigma_{YY} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \\ \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} & \sigma_{ZZ} \end{pmatrix}$

$X \perp\!\!\!\perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- From the definition of correlation coefficient, $\Sigma = \begin{pmatrix} \sigma_{XX} & \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} \\ \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sigma_{YY} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \\ \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} & \sigma_{ZZ} \end{pmatrix}$
- Then from the conditioning result, we have

$$\Sigma_{\left(\begin{smallmatrix} X \\ Y \end{smallmatrix}\right)|Z} = \begin{pmatrix} \sigma_{XX} & \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} \\ \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sigma_{YY} \end{pmatrix} - \begin{pmatrix} \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \end{pmatrix} \sigma_{ZZ}^{-1} \begin{pmatrix} \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} \\ \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{XX}(1-\rho_{XZ}^2) & \sqrt{\sigma_{XX}\sigma_{YY}}(\rho_{XY} - \rho_{XZ}\rho_{YZ}) \\ \sqrt{\sigma_{XX}\sigma_{YY}}(\rho_{XY} - \rho_{XZ}\rho_{YZ}) & \sigma_{YY}(1-\rho_{YZ}^2) \end{pmatrix}$$

$X \perp Y|Z$ if $\rho_{XZ}\rho_{YZ} = \rho_{XY}$

Proof.

- From the definition of correlation coefficient, $\Sigma = \begin{pmatrix} \sigma_{XX} & \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} \\ \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sigma_{YY} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \\ \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} & \sigma_{ZZ} \end{pmatrix}$
- Then from the conditioning result, we have

$$\begin{aligned}\Sigma_{\left(\begin{smallmatrix}X\\Y\end{smallmatrix}\right)|Z} &= \begin{pmatrix} \sigma_{XX} & \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} \\ \sqrt{\sigma_{XX}\sigma_{YY}}\rho_{XY} & \sigma_{YY} \end{pmatrix} \\ &\quad - \begin{pmatrix} \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} & \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \end{pmatrix} \sigma_{ZZ}^{-1} \begin{pmatrix} \sqrt{\sigma_{XX}\sigma_{ZZ}}\rho_{XZ} \\ \sqrt{\sigma_{YY}\sigma_{ZZ}}\rho_{YZ} \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{XX}(1-\rho_{XZ}^2) & \sqrt{\sigma_{XX}\sigma_{YY}}(\rho_{XY}-\rho_{XZ}\rho_{YZ}) \\ \sqrt{\sigma_{XX}\sigma_{YY}}(\rho_{XY}-\rho_{XZ}\rho_{YZ}) & \sigma_{YY}(1-\rho_{YZ}^2) \end{pmatrix}\end{aligned}$$

- Therefore, X and Y are uncorrelated given Z when the off-diagonal is zero and this gives us $\rho_{XY} = \rho_{XZ}\rho_{YZ}$. Since for Gaussian variables, uncorrelatedness implies independence. This concludes the proof.

□

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$.

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

- More concisely, we can write it as

$$Bern(x|p) = p^x (1-p)^{1-x},$$

Bernoulli distribution

- Consider someone flips a biased coin. The probability of the outcome is described by the Bernoulli distribution. Denote $X = 1$ for a head and $X = 0$ for a tail. Let $Pr(X = 1) = p$. Then the Bernoulli distribution is simply

$$Bern(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

- More concisely, we can write it as

$$Bern(x|p) = p^x(1-p)^{1-x},$$

- The mean and variance are

$$E[X] = p \cdot 1 + (1-p) \cdot 0 = p$$

$$Var[X] = p \cdot (1-p)^2 + (1-p) \cdot p^2 = p(1-p)$$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p, N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p, N-2) = N(N-1)p^2$

Binomial distribution (N trials)

- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
  - $E[X] = \sum_{x=0}^{N} Bin(x|p)x = \sum_{x=1}^{N} \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^{N} \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$
  - Similar, $E[X(X-1)] = \sum_{x=2}^{N} \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p,N-2) = N(N-1)p^2$
  - Therefore,
    $Var[X] = E[X^2] - E[X]^2$

Binomial distribution (N trials)

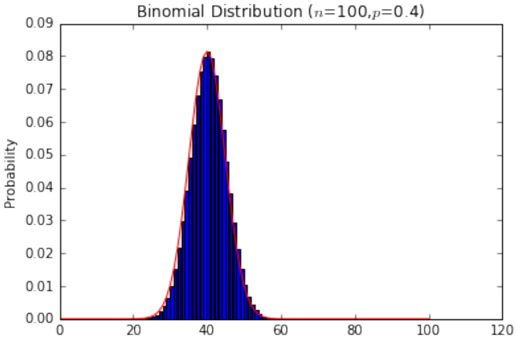
- Repeat the experiment for N times, the probability of the outcome will now be described by the binomial distribution. Note that x is now the number of obtained heads, we have

$$Bin(x|p,N) = \binom{N}{x} p^x (1-p)^{N-x},$$

- Mean and variances are given by
    - $E[X] = \sum_{x=0}^N Bin(x|p)x = \sum_{x=1}^N \frac{N!}{(x-1)!(N-x)!} p^x (1-p)^{N-x}$
    $= Np \sum_{x=1}^N \frac{(N-1)!}{(x-1)!(N-x)!} p^{x-1} (1-p)^{N-x} = Np \sum_{x=0}^{N-1} Bin(x|p,N-1)$
    $= Np$
    - Similar, $E[X(X-1)] = \sum_{x=2}^N \frac{N!}{(x-2)!(N-x)!} p^x (1-p)^{N-x}$
    $= N(N-1)p^2 \sum_{x=0}^{N-2} Bin(x|p,N-2) = N(N-1)p^2$
    - Therefore,
    $Var[X] = E[X^2] - E[X]^2 = E[X(X-1)] + E[X] - E[X]^2 = N(N-1)p^2 + Np - (Np)^2 = Np(1-p)$

Binomial distribution

As shown below, the binomial distribution can be model well with a normal distribution $\mathcal{N}(Np, Np(1-p))$ for large N



The binomial distribution is shown in blue and an approximation by normal distribution is shown in red

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u,v|p) = \arg\max_p p^u(1-p)^v$$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u,v|p) = \arg\max_p p^u(1-p)^v$$

- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead

$$\hat{p} = \arg\max_p p(u,v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u,v|p) = \arg\max_p p^u(1-p)^v$$

- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead

$$\hat{p} = \arg\max_p p(u,v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$

- It is very difficult to determine the prior unanimously. Actually it can be controversial just to determine the form of it

Conjugate prior

- Note that both Bernoulli and binomial distributions have the form $p^u(1-p)^v$
- To estimate p, recall that the ML estimator will try to compute

$$\hat{p} = \arg\max_p p(u,v|p) = \arg\max_p p^u(1-p)^v$$

- Now if we would like to use the MAP estimator instead, we need to introduce a prior $p(p)$ and solve instead

$$\hat{p} = \arg\max_p p(u,v|p)p(p) = \arg\max_p p^u(1-p)^v p(p)$$

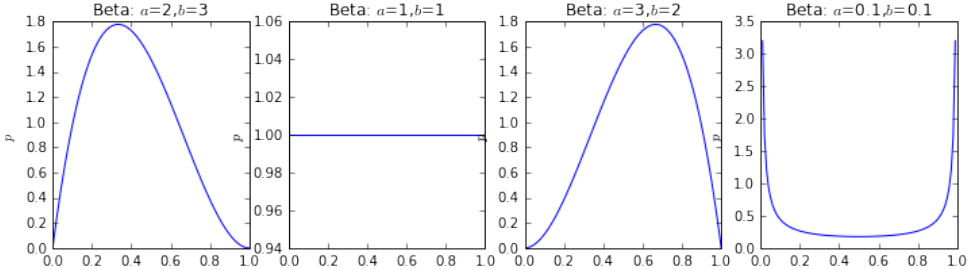
- It is very difficult to determine the prior unanimously. Actually it can be controversial just to determine the form of it
- However, if we select $p(p)$ of a form $p(p) \propto p^a(1-p)^b$, then the resulting posterior distribution with the same form as before. This choice is often chosen for practical purposes, and a prior with same "form" as its likelihood (and thus posterior) is known as the conjugate prior

Beta distribution

- The conjugate prior of both Bernoulli and binomial distributions is the beta distribution. Its pdf is given by

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},$$

where $X \in [0,1]$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

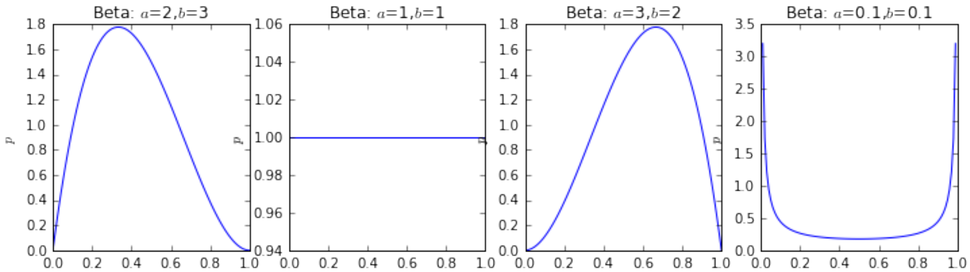


Beta distribution

- The conjugate prior of both Bernoulli and binomial distributions is the beta distribution. Its pdf is given by

$$Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)},$$

where $X \in [0,1]$ and $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$



- Note that with $a=b=1$, $Beta(x|1,1)=1$. It is the same as no prior

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, dx$

- $\Gamma(1) = \int_0^\infty e^{-x} \, dx = -e^{-x}|_0^\infty = 1$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$$

Gamma function

Note that $\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}\,dx$

- $\Gamma(1)=\int_0^\infty e^{-x}\,dx=-e^{-x}|_0^\infty=1$
- For $z>1$, we have $\Gamma(z)=(z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z)=\int_0^\infty x^{z-1}e^{-x}\,dx=-\int_0^\infty x^{z-1}de^{-x}$$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x} = -x^{z-1}e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx$$

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x} = -x^{z-1}e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2}e^{-x}\, dx$$

$$= (z-1)\int_0^\infty x^{z-2}e^{-x}\, dx = (z-1)\Gamma(z-1)$$

□

Gamma function

Note that $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$

- $\Gamma(1) = \int_0^\infty e^{-x}\, dx = -e^{-x}|_0^\infty = 1$
- For $z > 1$, we have $\Gamma(z) = (z-1)\Gamma(z-1)$

Proof.

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx = -\int_0^\infty x^{z-1} de^{-x} = -x^{z-1}e^{-x}|_0^\infty + (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx$$
$$= (z-1)\int_0^\infty x^{z-2} e^{-x}\, dx = (z-1)\Gamma(z-1)$$ □

- Therefore, for integer $z > 1$, $\Gamma(z) = (z-1)!$

Mode of beta distribution

- The mode is the peak of a distribution. Recall that $Beta(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$. Set

$$\frac{\partial Beta(x|a,b)}{\partial x} = \frac{(a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2}}{B(a,b)} = 0,$$

  we have $(a-1)(1-x) = (b-1)x \Rightarrow x = \frac{a-1}{a+b-2}$ when $a, b > 1$
- Note that when a or b is less than or equal to 1, the peak appears at either $x = 0$ or $x = 1$

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. This gives us a handy trick to manipulate beta distribution

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^{1} xBeta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Mean and variance of Beta distribution

Note that $\int_{x=0}^1 p(x|a,b) = 1 \Rightarrow \int_{x=0}^1 x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. This gives us a handy trick to manipulate beta distribution

$$E[X] = \int_{x=0}^1 x Beta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^1 x^a (1-x)^{b-1} dx$$
$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^1 x^{a+1}(1-x)^{b-1}dx$

Mean and variance of Beta distribution

Note that $\int_{x=0}^1 p(x|a,b) = 1 \Rightarrow \int_{x=0}^1 x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. This gives us a handy trick to manipulate beta distribution

$$
\begin{aligned}
E[X] &= \int_{x=0}^1 x Beta(x|a,b) dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^1 x^a (1-x)^{b-1} dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}
\end{aligned}
$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^1 x^{a+1}(1-x)^{b-1} dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}$.

Mean and variance of Beta distribution

Note that $\int_{x=0}^{1} p(x|a,b) = 1 \Rightarrow \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} = B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. This gives us a handy trick to manipulate beta distribution

$$
\begin{aligned}
E[X] &= \int_{x=0}^{1} x Beta(x|a,b)dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a}(1-x)^{b-1}dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} = \frac{a}{a+b}
\end{aligned}
$$

Similarly, $E[X^2] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_{x=0}^{1} x^{a+1}(1-x)^{b-1}dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} = \frac{a(a+1)}{(a+b)(a+b+1)}$. Thus,

$$
\begin{aligned}
Var[X] &= E[X^2] - E[X]^2 = \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)(a+b) - a^2(a+b+1)}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x.

[1] Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$p(p|x,a,b)$$

[1]Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$p(p|x,a,b) = Const1 \cdot Beta(p|a,b) Bern(x|p)$$

[1] Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Consider the coin flipping example again. Let say the prior probability[1] of the coin is beta distributed with parameters a and b. And we flip the coin once to get outcome x. Upon observing x, we can estimate p by

$$\begin{aligned} p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bern(x|p) \\ =& Const2 \cdot p^{a-1+x}(1-p)^{b-1+1-x} \\ =& Beta(p|\tilde{a},\tilde{b}) \end{aligned}$$

So the posterior probability distribution is also beta distributed and the parameters just changed to $\tilde{a} \leftarrow a + x$ and $\tilde{b} \leftarrow b + 1 - x$

[1]Note that this can be very confusing at the beginning. Beware that we are talking about the distribution of the probability of some outcome

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b.

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b. After the experiment x, we can update the distribution of our estimated p by

$$
\begin{aligned}
p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bin(x|p,N) \\
=& Const2 \cdot p^{a-1+x}(1-p)^{b-1+N-x} \\
=& Beta(p|\tilde{a},\tilde{b})
\end{aligned}
$$

Posterior estimate of probability p

Let say we continue our example and we flip the coin by N times and obtain x head. So instead of the Bernoulli likelihood, we have a binomial likelihood. Like the last slide, we have the same beta prior with parameters a and b. After the experiment x, we can update the distribution of our estimated p by

$$\begin{aligned} p(p|x,a,b) =& Const1 \cdot Beta(p|a,b) Bin(x|p,N) \\ =& Const2 \cdot p^{a-1+x}(1-p)^{b-1+N-x} \\ =& Beta(p|\tilde{a},\tilde{b}) \end{aligned}$$

Again, the posterior distribution is still beta but with parameters updated to $\tilde{a} \leftarrow a + x$ and $\tilde{b} \leftarrow b + N - x$

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads

Prior and regularization

- One major reason of introducing prior is for the sake of “regularizing” the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - $3/10$, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is $3/10$

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - $3/10$, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is $3/10$
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?

Prior and regularization

- One major reason of introducing prior is for the sake of “regularizing” the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - 3/10, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is 3/10
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?
  - 0? Okay, the estimate is a bit extreme. We know that it is very difficult to make a coin that always gives a tail

Prior and regularization

- One major reason of introducing prior is for the sake of "regularizing" the answer
- Another coin example
  - Fall back to high school, assume that we flip a coin for 10 times and got 3 heads. We want to estimate the chance of getting heads
  - 3/10, right?
  - And if I asked you chance of getting another head in the future, you will say the chance of getting another head is 3/10
  - Now, if I actually flip the coin for 10 times and got no head, what do you expect the chance of getting a head next time?
  - 0? Okay, the estimate is a bit extreme. We know that it is very difficult to make a coin that always gives a tail
  - How about we first assumed that we actually flipped two times and got 1 head before we did experiment? We will estimate 1/12 instead of 0/10

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

$$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a,10+b) = Beta(2,12)$$

Now, what is the MAP estimate?

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

$$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a, 10+b) = Beta(2,12)$$

Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

$$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a, 10+b) = Beta(2,12)$$

Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$. Thus,

$$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a = 2$ and $b = 2$. Note that the posterior distribution is

$$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a,10+b) = Beta(2,12)$$

Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$. Thus,

$$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

- Recall that $Beta(1,1) = 1$ and so likelihood function is equivalent to $Beta(p|1,1)Bin(0|p,10) \sim Beta(1,11)$.

Prior and regularization

- We can verify that this is exactly what we got for a Beta prior with $a=2$ and $b=2$. Note that the posterior distribution is

$$Beta(p|2,2)Bin(x=0|p,N=10) \sim Beta(0+a,10+b) = Beta(2,12)$$

Now, what is the MAP estimate? It should be the p that maximize the posterior probability. That is the mode of $Beta(2,12)$. Thus,

$$p_{Head}^{(MAP)} = \frac{a-1}{a+b-2} = \frac{1}{12}$$

- Recall that $Beta(1,1)=1$ and so likelihood function is equivalent to $Beta(p|1,1)Bin(0|p,10) \sim Beta(1,11)$. Thus the ML estimate is the mode of $Beta(1,11) \Rightarrow p_{Head}^{(ML)} = \frac{1-1}{1+11-2} = \frac{0}{10} = 0$
  - This indeed is the same as our high school naïve estimate

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p,

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean.

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean. Thus

$$p_{Head}^{(Bayesian)} = \frac{a}{a+b} = \frac{1}{11}$$

Bayesian estimation and regularization

- Now let's consider the Bayesian estimate. Even for the case with no prior (equivalently an uniform prior or Beta prior with $a = 1$ and $b = 1$), recall that the "posterior distribution" is $Beta(1, 11)$
- The Bayesian estimate should be the average p summing all possibility of p, which is essentially just, $\int pBeta(p|1, 11)dp = E[p]$, i.e., the mean. Thus

$$p_{Head}^{(Bayesian)} = \frac{a}{a+b} = \frac{1}{11}$$

- Note that Bayesian estimation is "self-regularized" (i.e., giving less extreme results) since it inherently averages out all possible cases

Remark

Note that we used the non-informative prior above just to illustrate the self-regularization property of Bayesian estimation. When you are given a prior, you should always use the given prior instead for an actual problem

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

$$Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) = \binom{N}{x_1 x_2 \cdots x_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n},$$

Multinomial distribution

- Binomial distribution models the probability of a binary outcome. For a random event with discrete but non-binary (more than two) outcomes, we can model the event with a multinomial distribution
- Let say the probability of each possible outcome i is p_i. And we have conducted N different experiments, let say x_i is the number of times we obtain outcome i. Then the probability of such even is given by

$$Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) = \binom{N}{x_1 x_2 \cdots x_n} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n},$$

- Just make sure we are in the same pace. Note that $p_1 + p_2 + \cdots + p_n = 1$ and $x_1 + x_2 + \cdots + x_n = N$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$
- It turns out that the distribution is the so-called Dirichlet distribution. Its pdf is given by

$$
\begin{aligned}
&Dir(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n)\\
=&\frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}
\end{aligned}
$$

Dirichlet distribution

- Note that the conjugate prior of multinomial distribution should take the form $x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$
- It turns out that the distribution is the so-called Dirichlet distribution. Its pdf is given by

$$\begin{aligned} & Dir(x_1,\cdots,x_n|\alpha_1,\cdots,\alpha_n) \\ =& \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)} x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1} \end{aligned}$$

- As usual since pdf should be normalized to 1, we have

$$\int x_1^{\alpha_1-1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n)}$$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1 + 1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1 + \cdots + \alpha_n + 1)} = \frac{\alpha_1}{\alpha_1 + \cdots + \alpha_n}$$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}\int x_1^{\alpha_1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$$
$$= \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}\frac{\Gamma(\alpha_1+1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+1)} = \frac{\alpha_1}{\alpha_1+\cdots+\alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}\int x_1^{\alpha_1+1}x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)}\frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$.

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2-1} \cdots x_n^{\alpha_n-1}$$
$$= \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+1)} = \frac{\alpha_1}{\alpha_1+\cdots+\alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1+1} x_2^{\alpha_2-1} \cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$. Thus, $Var(X_1) = E[X_1^2] - E[X_1^2] = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)} - \frac{\alpha_1^2}{(\alpha_1+\cdots+\alpha_n)^2} = \frac{\alpha_1(\alpha_0-\alpha_1)}{\alpha_0^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1+\cdots+\alpha_n$

Mean, mode, variance of Dirichlet distribution

- Mean:

$$E[X_1] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1} x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1}$$
$$= \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+1)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+1)} = \frac{\alpha_1}{\alpha_1+\cdots+\alpha_n}$$

- Similarly, $E[X_1^2] = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \int x_1^{\alpha_1+1} x_2^{\alpha_2-1}\cdots x_n^{\alpha_n-1} = \frac{\Gamma(\alpha_1+\cdots+\alpha_n)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_n)} \frac{\Gamma(\alpha_1+2)\cdots\Gamma(\alpha_n)}{\Gamma(\alpha_1+\cdots+\alpha_n+2)} = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)}$. Thus,
  $Var(X_1) = E[X_1^2] - E[X_1^2] = \frac{(\alpha_1+1)\alpha_1}{(\alpha_1+\cdots+\alpha_n+1)(\alpha_1+\cdots+\alpha_n)} - \frac{\alpha_1^2}{(\alpha_1+\cdots+\alpha_n)^2} = \frac{\alpha_1(\alpha_0-\alpha_1)}{\alpha_0^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1+\cdots+\alpha_n$
- Mode: one can show that the mode of $Dir(\alpha_1,\cdots,\alpha_n)$ for $\alpha_1,\cdots,\alpha_n > 1$ is

$$\frac{\alpha_i - 1}{\alpha_1+\cdots+\alpha_n - n}.$$

We will not show it now but will leave as an exercise

Summary of Dirichlet distribution

- Pdf:

$$Dir(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_n)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_n^{\alpha_n - 1}$$

- Mean:

$$\frac{\alpha_i}{\alpha_1 + \cdots + \alpha_n}$$

- Variance:

$$\frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

- Mode:

$$\frac{\alpha_i - 1}{\alpha_1 + \cdots + \alpha_n - n}$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n)$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$
\begin{aligned}
&p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n) \\
=&Const1 \cdot Dir(p_1, \cdots, p_n | \alpha_1, \cdots, \alpha_n) Mult(x_1, \cdots, x_n | p_1, \cdots, p_n)
\end{aligned}
$$

Posterior probability given Multinomial likelihood and Dirichlet prior

Upon observing $x_1, \cdots, x_n$, the posterior distribution of $p_1, \cdots, p_n$ becomes

$$\begin{aligned}
& p(p_1, \cdots, p_n | x_1, \cdots, x_n, \alpha_1, \cdots, \alpha_n) \\
=& Const1 \cdot Dir(p_1, \cdots, p_n | \alpha_1, \cdots, \alpha_n) Mult(x_1, \cdots, x_n | p_1, \cdots, p_n) \\
=& Const2 \cdot p_1^{x_1+\alpha_1} \cdots p_n^{x_n+\alpha_n} \\
=& Dir(p_1, \cdots, p_n | \tilde{\alpha}_1, \cdots, \tilde{\alpha}_n)
\end{aligned}$$

So the posterior distribution is Dirichlet with parameters updated to $\tilde{\alpha}_1 \leftarrow x_1 + \alpha_1, \cdots, \tilde{\alpha}_n \leftarrow x_n + \alpha_n$

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store.

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store. Its pdf is given by

$$Poisson(k|\lambda T) = \frac{e^{-\lambda T}(\lambda T)^k}{k!},$$

where k is a non-negative integer, λ is rate of arrival and T is the length of the observed period.

Poisson distribution

Poisson distribution describes the number of arrival K within some period. For example, one can use Poisson distribution to model the arrival process (Poisson process) of customers into a store. Its pdf is given by

$$Poisson(k|\lambda T) = \frac{e^{-\lambda T}(\lambda T)^k}{k!},$$

where k is a non-negative integer, λ is rate of arrival and T is the length of the observed period. It is easy to check that (please verify)

$$Mean = \lambda T$$

$$Variance = \lambda T$$

N.B. the parameters λT comes as a group and so we can consider it as a single parameter

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
   - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
    - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
    - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease
    - It makes sense to model say customers to a department store

Poisson process

Poisson process is probably the simplest random process to model event arrivals. It is based on two simple assumptions

1. Arrival rate is invariant over time
   - That is, λ is a constant that does not change with time
2. Each arrival is independent of the other
   - For example, even though we just have one customer coming in, the probability that the next customer to come in immediately should not decrease
   - It makes sense to model say customers to a department store
   - It can be less perfect to model the times my car broke down. The events are likely to be related

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$.

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T)$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $\approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $\approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k} = \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k}$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $\approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k} = \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^N$

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $\approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k} = \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^N = \frac{(\lambda T)^k}{k!}\exp(-\lambda T)$,
  where we use $(1+a/N)^N = \exp(a)$ for the last equality

Poisson process and Poisson distribution

- Consider a period T and let's the arrival rate be λ as before. Let's partition T into N different very short intervals of length Δ. Hence, $T = N\Delta$. We will also assume $N \to \infty$ and thus $\Delta \to 0$. The probability of getting an arrival in any interval Δ is thus $\lambda\Delta$. Moreover, since $\Delta \to 0$, the probability of getting two arrivals $\propto \Delta^2$ and is negligible compared to $\lambda\Delta$
- Then, the probability of k arrivals
  $Pr(k \text{ arrivals in } T) = \binom{N}{k}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k} = \frac{N(N-1)\cdots(N-k+1)}{k!}(\lambda\Delta)^k(1-\lambda\Delta)^{N-k}$
  $\approx \frac{N^k}{k!}\lambda^k\frac{T^k}{N^k}(1-\lambda\Delta)^{N-k} = \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N-k} \approx \frac{(\lambda T)^k}{k!}(1-\frac{\lambda T}{N})^{N} = \frac{(\lambda T)^k}{k!}\exp(-\lambda T)$,
  where we use $(1+a/N)^N = \exp(a)$ for the last equality

Note that indeed $Pr(k \text{ arrivals in } T) = Poisson(k|\lambda T)$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$.

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, Pr(next event happened within in time $[t, t+\Delta]$) $= Pr$(next event happened within in time $[n\Delta, (n+1)\Delta]$)

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, $Pr(\text{next event happened within in time } [t, t + \Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n + 1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n + 1 \text{ interval})$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, $Pr(\text{next event happened within in time } [t, t+\Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n+1 \text{ interval}) = (1-\lambda\Delta)^n(\lambda\Delta)$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, $Pr(\text{next event happened within in time } [t, t+\Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n+1 \text{ interval}) = (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then, $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta}$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, $Pr(\text{next event happened within in time } [t, t+\Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n+1 \text{ interval}) = (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then, $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1-\lambda\frac{t}{n})^n$

Interarrival time of Poisson process

Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, $Pr(\text{next event happened within in time } [t, t+\Delta])$
  $= Pr(\text{next event happened within in time } [n\Delta, (n+1)\Delta])$
  $= Pr(\text{no event in first } n \text{ intervals}) Pr(\text{event happened in } n+1 \text{ interval}) = (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then, $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1-\lambda\frac{t}{n})^n = \lambda\exp(-\lambda t)$, where we use $(1+a/n)^n = \exp(a)$ again for $n \to \infty$

Interarrival time of Poisson process

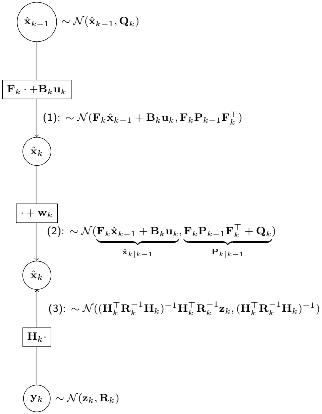
Using the similar analysis, we can also easily evaluate the distribution of interarrival time, the time that the next event will happen given that an event just happened. Let $t = n\Delta$ and use the same notation as before

- Note that $t > 0$ and $\Delta \to 0$ and so $n \to \infty$. Now, Pr(next event happened within in time $[t, t+\Delta]$)
  $= Pr$(next event happened within in time $[n\Delta, (n+1)\Delta]$)
  $= Pr$(no event in first n intervals)Pr(event happened in $n+1$ interval) $= (1-\lambda\Delta)^n(\lambda\Delta)$
- Let $f_T(t)$ be the pdf of the interval time. Then, $f_T(t) = \frac{(1-\lambda\Delta)^n(\lambda\Delta)}{\Delta} = \lambda(1-\lambda\frac{t}{n})^n = \lambda\exp(-\lambda t)$, where we use $(1+a/n)^n = \exp(a)$ again for $n \to \infty$

Exponential distribution

$f_T(t) = \lambda\exp(-\lambda t) \triangleq Exp(t|\lambda)$ is the pdf of the exponential distribution with parameter λ. It is easy to verify that (as exercise)

- $E[T] = 1/\lambda$
- $Var(T) = 1/\lambda^2$


$\hat{\mathbf{x}}_{k-1} \sim \mathcal{N}(\hat{\mathbf{x}}_{k-1}, \mathbf{Q}_k)$
$\mathbf{F}_k \cdot + \mathbf{B}_k \mathbf{u}_k$
(1): $\sim \mathcal{N}(\mathbf{F}_k \hat{\mathbf{x}}_{k-1} + \mathbf{B}_k \mathbf{u}_k, \mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^\top)$
$\tilde{\mathbf{x}}_k$
$\cdot + \mathbf{w}_k$
(2): $\sim \mathcal{N}(\underbrace{\mathbf{F}_k \hat{\mathbf{x}}_{k-1} + \mathbf{B}_k \mathbf{u}_k}_{\hat{\mathbf{x}}_{k|k-1}}, \underbrace{\mathbf{F}_k \mathbf{P}_{k-1} \mathbf{F}_k^\top + \mathbf{Q}_k}_{\mathbf{P}_{k|k-1}})$
$\hat{\mathbf{x}}_k$
(3): $\sim \mathcal{N}((\mathbf{H}_k^\top \mathbf{R}_k^{-1} \mathbf{H}_k)^{-1} \mathbf{H}_k^\top \mathbf{R}_k^{-1} \mathbf{z}_k, (\mathbf{H}_k^\top \mathbf{R}_k^{-1} \mathbf{H}_k)^{-1})$
$\mathbf{H}_k \cdot$
$\mathbf{y}_k \sim \mathcal{N}(\mathbf{z}_k, \mathbf{R}_k)$