CT216 Introduction to Communication Systems
Lecture 2: Probability, Information, and Estimation

Yash M. Vasavada

Professor, DA-IICT, Gandhinagar

24th January 2024

Overview of Today's Talk

Overview of Today's Talk

1. Generation of Information
2. A Review of Probability Theory
   - Several Example Random Variables
   - Fundamental Concepts

Overview of Today's Talk

1. Generation of Information
2. A Review of Probability Theory
   - Several Example Random Variables
   - Fundamental Concepts
3. Bayes' Theorem

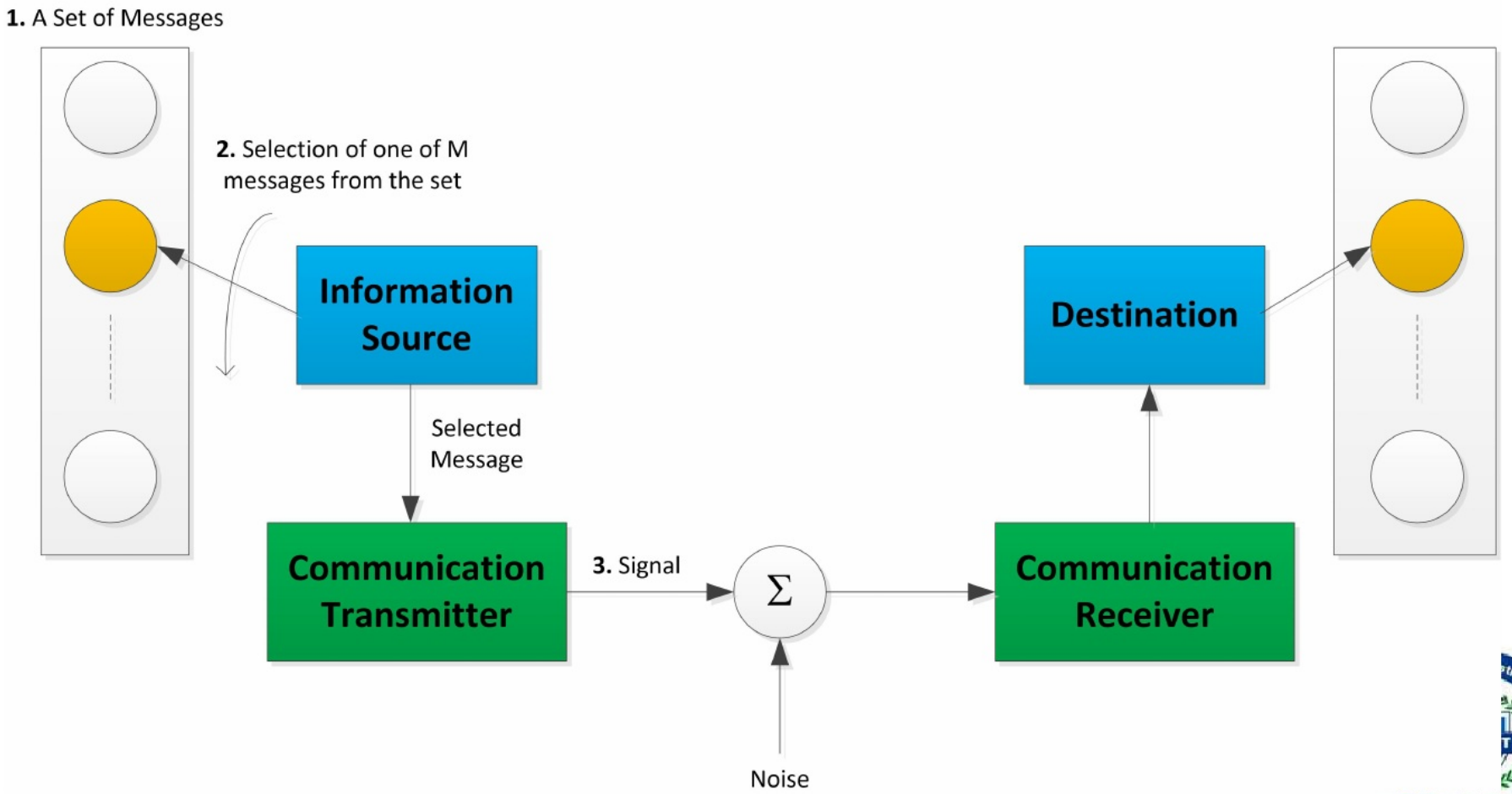
Overview of Today's Talk

Overview of Today's Talk

Overview of Today's Talk

The Model of a Communication System

Crystallized by Shannon



The Definition and Measurement of Information

The Source Alphabet

- The information source is modeled as generating symbols each of which can take K symbols from a set $\{d_1, d_2, \ldots, d_K\}$. This set is the source alphabet.

The Definition and Measurement of Information

The Source Alphabet

- The information source is modeled as generating symbols each of which can take K symbols from a set $\{d_1, d_2, \ldots, d_K\}$. This set is the source alphabet.
  - E.g., in the C disease problem, the source (i.e., the human body) is binary, i.e., it takes one of $K = 2$ symbols; $C = 0$ denotes that there is no disease, $C = 1$ is that the body has the disease

The Definition and Measurement of Information

The Source Alphabet

- The information source is modeled as generating symbols each of which can take K symbols from a set $\{d_1, d_2, \ldots, d_K\}$. This set is the source alphabet.
  - E.g., in the C disease problem, the source (i.e., the human body) is binary, i.e., it takes one of $K = 2$ symbols; $C = 0$ denotes that there is no disease, $C = 1$ is that the body has the disease
  - When the information source (e.g., a writer) uses the English language, K has to be at least equal to the number of letters ($= 26$) of the English alphabet (however, a common method for text encoding, the ASCII character set, uses $K = 128$)

The Definition and Measurement of Information

The Source Alphabet

- The information source is modeled as generating symbols each of which can take K symbols from a set $\{d_1, d_2, \ldots, d_K\}$. This set is the source alphabet.
  - E.g., in the C disease problem, the source (i.e., the human body) is binary, i.e., it takes one of $K = 2$ symbols; $C = 0$ denotes that there is no disease, $C = 1$ is that the body has the disease
  - When the information source (e.g., a writer) uses the English language, K has to be at least equal to the number of letters ($= 26$) of the English alphabet (however, a common method for text encoding, the ASCII character set, uses $K = 128$)

The Definition and Measurement of Information

The Symbol Probabilities

- Probability $P(D = d_k)$ of k^{th} symbol is denoted as p_k. Different values of p_k may not all be the same (i.e., some source symbols may be more likely to occur than the others). The sum of these probabilities equals 1, i.e., $\sum_{k=1}^{K} p_k = 1$.

The Definition and Measurement of Information

The Symbol Probabilities

- Probability $P(D = d_k)$ of k^{th} symbol is denoted as p_k. Different values of p_k may not all be the same (i.e., some source symbols may be more likely to occur than the others). The sum of these probabilities equals 1, i.e., $\sum_{k=1}^{K} p_k = 1$.
  - E.g., in the C disease problem, $p(C = 0) = 0.99$ and $p(C = 1) = 0.01$.

The Definition and Measurement of Information

The Symbol Probabilities

- Probability $P(D = d_k)$ of k^{th} symbol is denoted as p_k. Different values of p_k may not all be the same (i.e., some source symbols may be more likely to occur than the others). The sum of these probabilities equals 1, i.e., $\sum_{k=1}^{K} p_k = 1$.
  - E.g., in the C disease problem, $p(C = 0) = 0.99$ and $p(C = 1) = 0.01$.
  - For the English alphabet, some letters such as `e`, `a`, `i`, `t`, etc. have a much higher probability of occurrence compared to the letters such as `j`, `q`, `x`, `z`.
    - → See, e.g., https://www3.nd.edu/ busi-forc/handouts/cryptography/letterfrequencies.html

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k. Why?

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k. Why?
     - The smaller the p_k, the greater the surprise by the occurrence of k^{th} symbol

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k. Why?
     - The smaller the p_k, the greater the surprise by the occurrence of k^{th} symbol
     - In the C-disease problem, $p(C = 0)$ is very high (0.99), and so disease-free event is not highly surprising or informative. In contrast, $C = 1$ is a low-probability event and so its occurrence conveys a greater surprise, and delivers a bigger information

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k. Why?
     - The smaller the p_k, the greater the surprise by the occurrence of k^{th} symbol
     - In the C-disease problem, $p(C = 0)$ is very high (0.99), and so disease-free event is not highly surprising or informative. In contrast, $C = 1$ is a low-probability event and so its occurrence conveys a greater surprise, and delivers a bigger information
     - Suppose $C = 1$ denotes an event that "there is a snowfall in Gandhinagar". The probability of this event is so small that in case it snows in Gandhinagar, that day would be *extremely* informative! The typical days for which $C = 0$ are non-informative in this sense

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k. Why?
     - The smaller the p_k, the greater the surprise by the occurrence of k^{th} symbol
     - In the C-disease problem, $p(C = 0)$ is very high (0.99), and so disease-free event is not highly surprising or informative. In contrast, $C = 1$ is a low-probability event and so its occurrence conveys a greater surprise, and delivers a bigger information
     - Suppose $C = 1$ denotes an event that "there is a snowfall in Gandhinagar". The probability of this event is so small that in case it snows in Gandhinagar, that day would be *extremely* informative! The typical days for which $C = 0$ are non-informative in this sense
     - In the English language, the letters such as e carry less information compared to the letters such as q

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.
  2. Information should be a logarithmic function (the base of the logarithm would determine the unit of the information). Why?

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.
  2. Information should be a logarithmic function (the base of the logarithm would determine the unit of the information). Why?
     - When all p_k are equal, the information generated is log of K

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.
  2. Information should be a logarithmic function (the base of the logarithm would determine the unit of the information). Why?
     - When all p_k are equal, the information generated is log of K
     - E.g., in the C-disease problem, if $C = 0$ and $C = 1$ events are equally likely, they each would be assigned a single bit 0 or 1

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.
  2. Information should be a logarithmic function (the base of the logarithm would determine the unit of the information). Why?
     - When all p_k are equal, the information generated is log of K
     - E.g., in the C-disease problem, if $C = 0$ and $C = 1$ events are equally likely, they each would be assigned a single bit 0 or 1
     - Similarly, if a source alphabet has $K = 16$ symbols, total $\log_2(16) = 4$ bits are required to represent all of them $(0, 0, 0, 0,\ 0, 0, 0, 1,\ \ldots,$ $1, 1, 1, 0,$ and $1, 1, 1, 1)$.

A Measure of Information Content of the Source

- What is the information I_k generated when the source emits k^{th} symbol d_k with probability p_k? Shannon's answer:
  1. Information should be inversely proportional to p_k.
  2. Information should be a logarithmic function (the base of the logarithm would determine the unit of the information). Why?
     - When all p_k are equal, the information generated is log of K
     - E.g., in the C-disease problem, if $C = 0$ and $C = 1$ events are equally likely, they each would be assigned a single bit 0 or 1
     - Similarly, if a source alphabet has $K = 16$ symbols, total $\log_2(16) = 4$ bits are required to represent all of them $(0, 0, 0, 0,\ 0, 0, 0, 1,\ \ldots,$ $1, 1, 1, 0,$ and $1, 1, 1, 1)$.
- Information generated by k^{th} symbol is defined as:

$$I_k = \log_2\left(\frac{1}{p_k}\right) = -\log_2 p_k \text{ bits.}$$

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF
- The PMF has the following three key properties:

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF
- The PMF has the following three key properties:
  1. $p_k \geq 0$

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF
- The PMF has the following three key properties:
  1. $p_k \geq 0$
  2. $\sum_{k=1}^{K} p_k = 1$

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF
- The PMF has the following three key properties:
  1. $p_k \geq 0$
  2. $\sum_{k=1}^{K} p_k = 1$
  3. $P(k_a \leq k \leq k_b) = \sum_{k=k_a}^{k_b} p_k$

Probability Mass Functions or PMFs

- The set of the probabilities $\{p_1, p_2, \ldots, p_K\}$ of all K symbols of the source is called the Probability Mass Function or PMF
- The PMF has the following three key properties:
  1. $p_k \geq 0$
  2. $\sum_{k=1}^{K} p_k = 1$
  3. $P(k_a \leq k \leq k_b) = \sum_{k=k_a}^{k_b} p_k$

The Random Variables (RVs)

Main Point: The RVs are *always* numbers (integers, fractionals, real-numbers, and sometimes even complex numbers)

The Random Variables (RVs)

Main Point: The RVs are *always* numbers (integers, fractionals, real-numbers, and sometimes even complex numbers)

- Let the symbol d_k be assigned a unique number x_k. This number $X = x_k$ that the source output takes is called a random variable or an RV in short.

The Random Variables (RVs)

Main Point: The RVs are *always* numbers (integers, fractionals, real-numbers, and sometimes even complex numbers)

Let the symbol d_k be assigned a unique number x_k. This number $X = x_k$ that the source output takes is called a random variable or an RV in short.

Only for the RVs, there are two statistical concepts: the expected (or mean) value of the RV and the variance (the square of standard deviation) of the RV

The Random Variables (RVs)

Main Point: The RVs are *always* numbers (integers, fractionals, real-numbers, and sometimes even complex numbers)

- Let the symbol d_k be assigned a unique number x_k. This number $X = x_k$ that the source output takes is called a random variable or an RV in short.
- Only for the RVs, there are two statistical concepts: the expected (or mean) value of the RV and the variance (the square of standard deviation) of the RV
  - $\mathbb{E}[X] = \bar{X} = \sum_{k=1}^{K} x_k p_k$
  - $\text{Var}[X] = \sum_{k=1}^{K} \left[x_k - \bar{X}\right]^2 p_k = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$

Cumulative Mass Function (CMF)

- Defined using the PMF — but only for an RV

Cumulative Mass Function (CMF)

- Defined using the PMF — but only for an RV
- Definition: $F_X(x_m) = P(X \leq x_m) = \sum_{k=1}^{m} p_k$
- Properties:
  1. $F_X(x_m)$ is monotonically nondecreasing
  2. $F(-\infty) = 0$
  3. $F(\infty) = 1$
  4. $P(x_a < X \leq x_b) = F(x_b) - F(x_a)$
- The CMF also completely defines the probability distribution of an RV

A Measure of Information Content of the Source

- Besides the information generated by the individual symbol of the source, Shannon showed that there is an important measure of how much the source is informative in an average sense

A Measure of Information Content of the Source

- Besides the information generated by the individual symbol of the source, Shannon showed that there is an important measure of how much the source is informative in an average sense
- The average informative-ness of the source X is called the Entropy $H(X)$ and it equals

$$H(X) = \mathbb{E}\left[I_k\right]$$

Note that the information I_k is a *number* and so it qualifies to be an RV whose expectation can be taken

A Measure of Information Content of the Source

- Besides the information generated by the individual symbol of the source, Shannon showed that there is an important measure of how much the source is informative in an average sense
- The average informative-ness of the source X is called the Entropy $H(X)$ and it equals

$$H(X) = \mathbb{E}\left[I_k\right]$$

Note that the information I_k is a *number* and so it qualifies to be an RV whose expectation can be taken

- From the definition of the expected value of an RV, $H(X)$ can be written as

$$H(X) = \mathbb{E}\left[I_k\right] = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

A Binary Random Variable

Called Bernoulli(q) RV X

- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$

A Binary Random Variable

Called Bernoulli(q) RV X

- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$

  $\rightarrow$ Mean: $m_X = \sum_x x\, p_X(x) = 0 \times (1-q) + 1 \times q = q$

A Binary Random Variable

Called Bernoulli(q) RV X

- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$

  → Mean: $m_X = \sum_x x\, p_X(x) = 0 \times (1-q) + 1 \times q = q$

  → Variance: $\sigma_X^2 = q(1-q)$

A Binary Random Variable

Called Bernoulli(q) RV X

- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$
  - → Mean: $m_X = \sum_x x\, p_X(x) = 0 \times (1-q) + 1 \times q = q$
  - → Variance: $\sigma_X^2 = q(1-q)$
  - → Information $I_k = -\log_2 q$ and $-\log_2(1-q)$ bits

A Binary Random Variable

Called Bernoulli(q) RV X

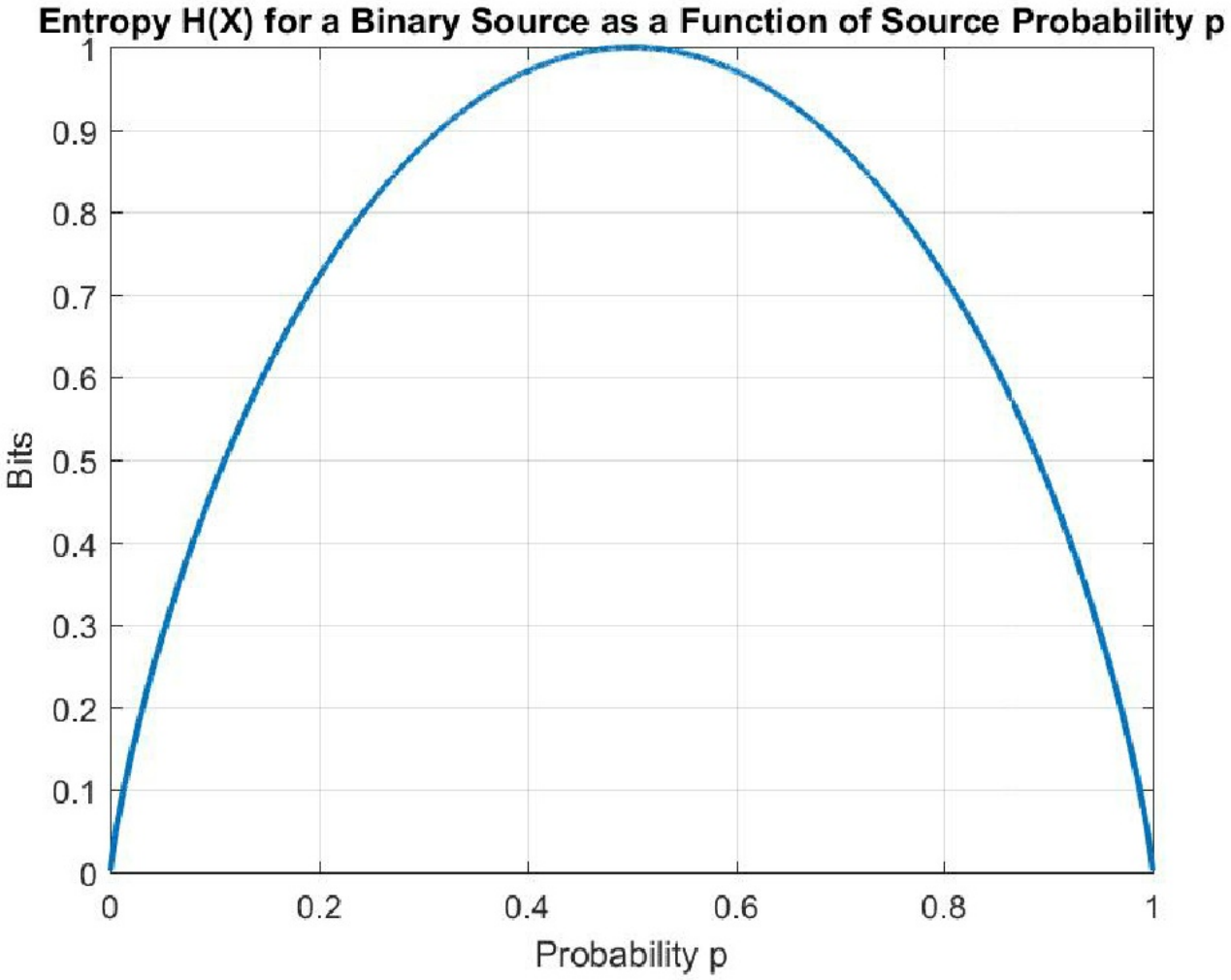
- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$
  - → Mean: $m_X = \sum_x x\, p_X(x) = 0 \times (1-q) + 1 \times q = q$
  - → Variance: $\sigma_X^2 = q(1-q)$
  - → Information $I_k = -\log_2 q$ and $-\log_2(1-q)$ bits
  - → Average information or Entropy
    $H(X) = -q \log_2 q - (1-q) \log_2(1-q)$ bits

A Binary Random Variable

Called Bernoulli(q) RV X

- Outcome of the toss of a coin: $p(X) = \begin{cases} 1-q, & X = 0\,(\text{head}), \\ q, & X = 1\,(\text{tail}) \end{cases}$
  - → Mean: $m_X = \sum_x x\, p_X(x) = 0 \times (1-q) + 1 \times q = q$
  - → Variance: $\sigma_X^2 = q(1-q)$
  - → Information $I_k = -\log_2 q$ and $-\log_2(1-q)$ bits
  - → Average information or Entropy
    $H(X) = -q \log_2 q - (1-q) \log_2(1-q)$ bits
    - ▷ $H(X)$ takes the highest value of 1 bit when $q = 0.5$, i.e., when $X = 0$ and $X = 1$ both are equally likely

Entropy of Bernoulli(p) RV



Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.

Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.
  - The source has to be discrete in nature; someone speaking is also an information source, but his speech waveform is continuous. We have not defined how to define/measure the information generated by him/her

Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.
  - The source has to be discrete in nature; someone speaking is also an information source, but his speech waveform is continuous. We have not defined how to define/measure the information generated by him/her
- This definition agrees with our intuitive notion about how information should be measured

Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.
  - The source has to be discrete in nature; someone speaking is also an information source, but his speech waveform is continuous. We have not defined how to define/measure the information generated by him/her
- This definition agrees with our intuitive notion about how information should be measured
  - the more surprising an event is, a greater amount of information I_k is defined to have been generated

Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.
  - The source has to be discrete in nature; someone speaking is also an information source, but his speech waveform is continuous. We have not defined how to define/measure the information generated by him/her
- This definition agrees with our intuitive notion about how information should be measured
  - the more surprising an event is, a greater amount of information I_k is defined to have been generated
- The definition also provides the units, e.g., in bits

Summary

What we have learnt so far

- We have seen a succinct definition of the information generated by a source
  - The source can be an author writing in English, Hindi, Gujarati, etc., or it can be a camera taking a picture, etc.
  - The source has to be discrete in nature; someone speaking is also an information source, but his speech waveform is continuous. We have not defined how to define/measure the information generated by him/her
- This definition agrees with our intuitive notion about how information should be measured
  - the more surprising an event is, a greater amount of information I_k is defined to have been generated
- The definition also provides the units, e.g., in bits
  - The more informative an event is, a greater number of bits are required to represent it

Summary

What we have learnt so far

- An information source generates, or it causes, one event to occur out of a set of multiple events
- Alternatively, it selects one information-bearing message out of a set of messages (these messages or events are called the symbols)
- While an information measure I_k is defined for each event/message/symbol k, the information generated on average by the source itself is measured using the Entropy $H(X)$

Summary

What we have learnt so far

- The Entropy $H(X)$ captures the notion of information in a better manner compared to I_k — if I_k is large, only an individual source symbol is highly informative, whereas if $H(X)$ is large, the source as a whole is highly informative

Summary

What we have learnt so far

- The Entropy $H(X)$ captures the notion of information in a better manner compared to I_k — if I_k is large, only an individual source symbol is highly informative, whereas if $H(X)$ is large, the source as a whole is highly informative
  - An analogy: a few molecules in a sealed container filled with the gas may be vibrating a lot, but the overall temperature of the gas inside the container will be low if most of the molecules of the gas are not vibrating much;
  - In contrast, if all the molecules are vibrating about equally (i.e., if p_k is equal to $1/K$ for all symbol indices k), the temperature will be highest

Summary

What we have learnt so far

- The Entropy $H(X)$ captures the notion of information in a better manner compared to I_k — if I_k is large, only an individual source symbol is highly informative, whereas if $H(X)$ is large, the source as a whole is highly informative
  - An analogy: a few molecules in a sealed container filled with the gas may be vibrating a lot, but the overall temperature of the gas inside the container will be low if most of the molecules of the gas are not vibrating much;
  - In contrast, if all the molecules are vibrating about equally (i.e., if p_k is equal to $1/K$ for all symbol indices k), the temperature will be highest
  - Thus, the entropy $H(X)$ gives an idea of the "temperature" of the source (the "temperature", here, indicates the informative-ness of the source)

Summary

What we have learnt so far

- The Entropy $H(X)$ captures the notion of information in a better manner compared to I_k — if I_k is large, only an individual source symbol is highly informative, whereas if $H(X)$ is large, the source as a whole is highly informative
  - An analogy: a few molecules in a sealed container filled with the gas may be vibrating a lot, but the overall temperature of the gas inside the container will be low if most of the molecules of the gas are not vibrating much;
  - In contrast, if all the molecules are vibrating about equally (i.e., if p_k is equal to $1/K$ for all symbol indices k), the temperature will be highest
  - Thus, the entropy $H(X)$ gives an idea of the "temperature" of the source (the "temperature", here, indicates the informative-ness of the source)
  - A source that is the *hottest* is the one that has the highest Entropy, i.e., whose all K symbols occur with equal probability (i.e., its output is least predictable)

Summary

What we have learnt so far

- We have obtained our answer to at least first few fundamental questions regarding the information and its generation, except

Summary

What we have learnt so far

- We have obtained our answer to at least first few fundamental questions regarding the information and its generation, except
  - The answer is a bit ad-hoc; we have appealed to intuition, not a rigorous mathematical derivation

Summary

What we have learnt so far

- We have obtained our answer to at least first few fundamental questions regarding the information and its generation, except
  - The answer is a bit ad-hoc; we have appealed to intuition, not a rigorous mathematical derivation
- We will now move on to the question of the information transfer over a communication channel

A Nonbinary Random Variable

Binomial RV

- Let $Y = \sum_{i=1}^{n} X_i$, where $\{X_i\}; i = 1, \ldots, N$ are independent binary RVs with $p(x) = \begin{cases} 1-p, & x = 0(\text{head}), \\ p, & x = 1(\text{tail}) \end{cases}$
- In this case, RV Y follows the Binomial Distribution given as $P_Y(y) = \binom{N}{y} p^y (1-p)^{N-y}$
- Mean $m_y = N \times p$
- Variance $\sigma_y^2 = N \times p \times (1-p)$

Example PMFs

Binomial Distribution

- Let $Y = \sum_{i=1}^{n} X_i$, where $\{X_i\}; i = 1, \ldots, N$ are independent binary RVs with $p(x) = \begin{cases} 1-p, & x = 0(\text{head}), \\ p, & x = 1(\text{tail}) \end{cases}$
- In this case, RV Y follows the Binomial Distribution given as $P_Y(y) = \binom{N}{y} p^y (1-p)^{N-y}$
- Mean $m_y = N \times p$
- Variance $\sigma_y^2 = N \times p \times (1-p)$

Probability Density Function (PDF)

- The concept of PMF generalizes to the PDF when the RV X is continuous
- Definition: $p_X(x) = \dfrac{dF(x)}{dx}$
- Interpretations: PDF measures
  - $\rightarrow$ how likely a random variable is to lie at a particular value
- Properties
  1. $p(x) \geq 0$
  2. $\displaystyle\int_{-\infty}^{\infty} p(x)dx = 1$
  3. $P(a \leq X \leq b) = \displaystyle\int_{a}^{b} p(x)dx$
  4. Unlike the PMFs, the value of PDF for a given value of X can exceed 1.

Cumulative Distribution Function (CDF)

- Definition: $F_X(x) = F(x) = P(X \leq x)$
- Properties:
  1. $F(x)$ is monotonically nondecreasing
  2. $F(-\infty) = 0$
  3. $F(\infty) = 1$
  4. $P(a < X \leq b) = F(b) - F(a)$
- The CDF also completely defines the probability distribution of an RV

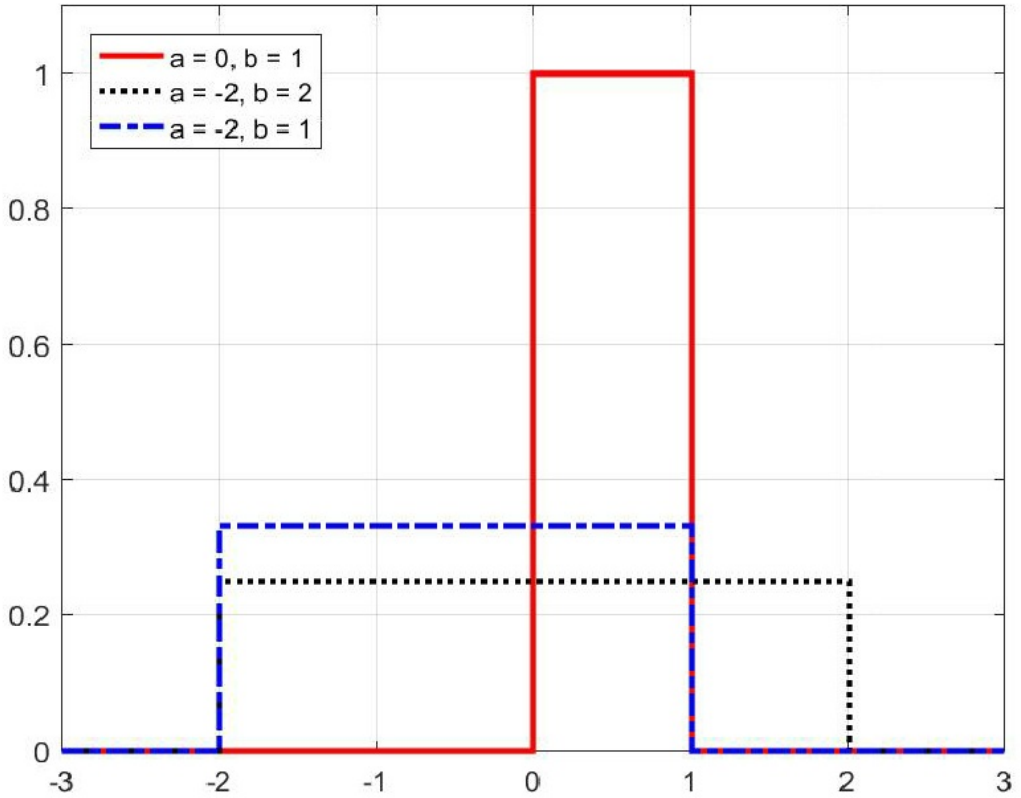
Expected Values

- Sometimes the PDF is cumbersome to specify, or it may not be known
- Expected values are shorthand ways of describing the behavior of RVs
- Most important examples are:
  - → Mean: $E(x) = m_x = \int_{-\infty}^{\infty} x\, p(x)\, dx$
  - → Variance: $E\left((x - m_x)^2\right) = \int_{-\infty}^{\infty} (x - m_x)^2\, p(x)\, dx$
- Expectation operator works with any function $Y = g(X)$.
  - → $E(Y) = E\left(g(X)\right) = \int_{-\infty}^{\infty} g(x)\, p(x)\, dx$

Example RVs

Uniform PDF

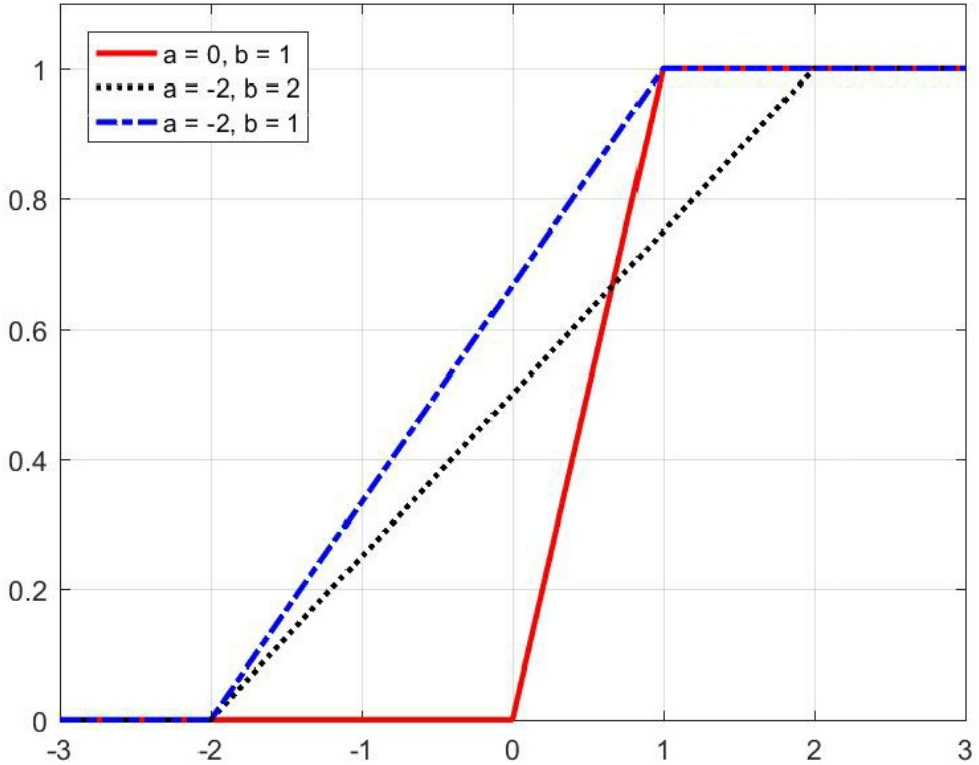
- $p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{else} \end{cases}$



Example PDFs

Uniform CDF

- $F(x) = \begin{cases} 0, & x \leq 0 \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x \geq b \end{cases}$



Example PDFs

Uniform RV

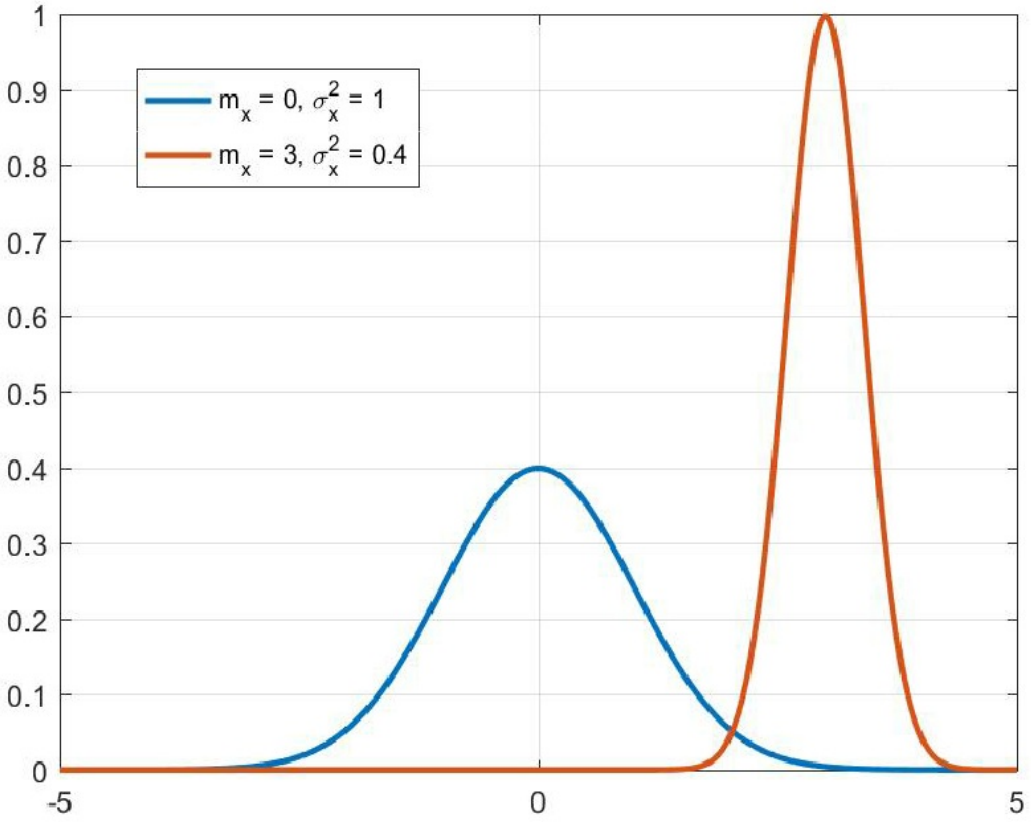
- Mean: $m_x = \int_a^b x\, p(x)\, dx = \frac{1}{b-a} \int_a^b x\, dx = \frac{a+b}{2}$
- Variance: $\sigma_x^2 = \int_a^b (x - m_x)^2\, p(x)\, dx = \frac{(b-a)^2}{12}$
- Probability:

$$P(a_1 \leq x < b_1) = \int_{a_1}^{b_1} p(x)\, dx = \frac{b_1 - a_1}{b-a}, \quad a < a_1, b_1 < b$$

Example RVs

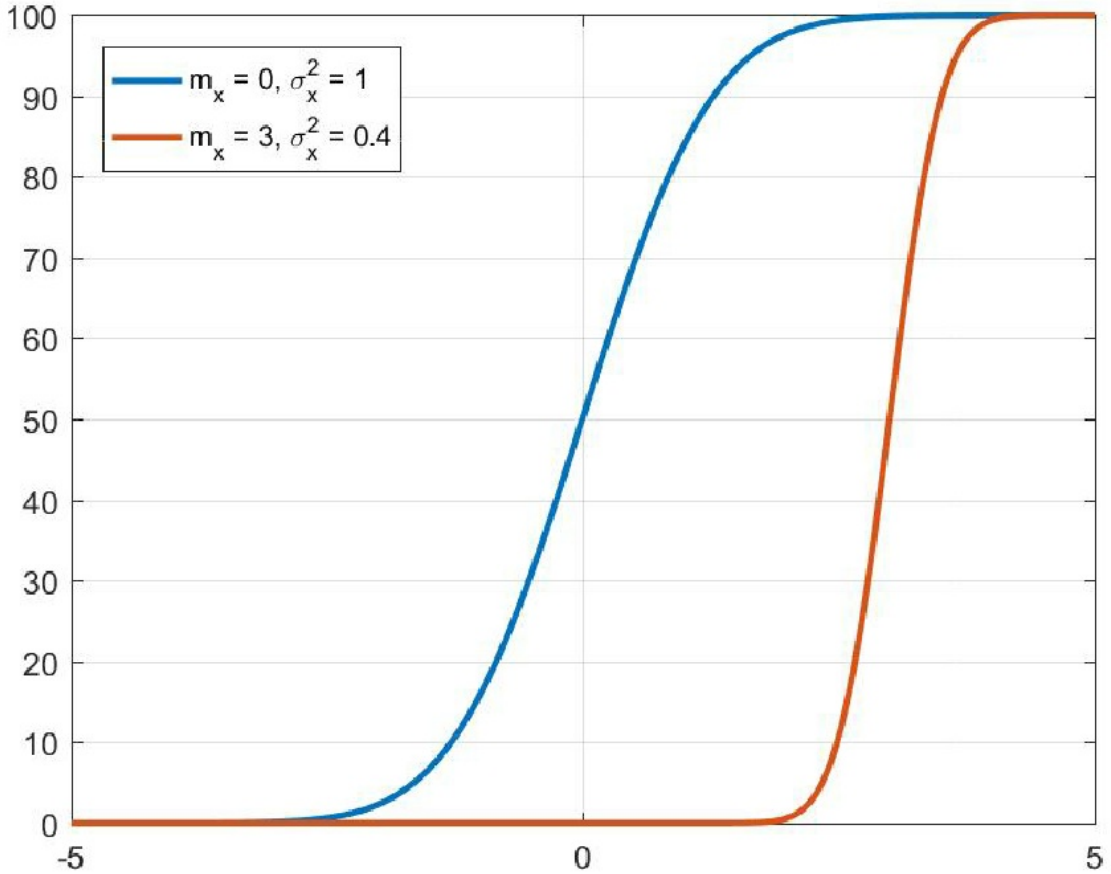
Gaussian PDF

- $p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-m_x)^2}{2\sigma_x^2}\right)$



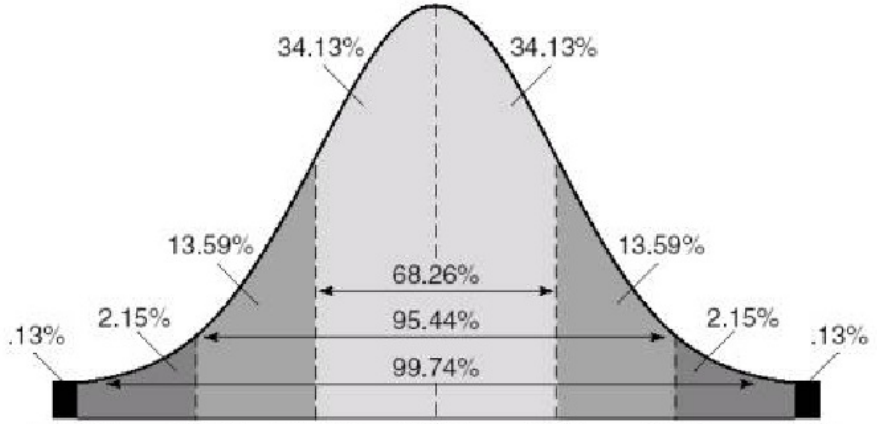
Example RVs

Gaussian CDF



Example RVs

Gaussian PDF



z-scores	−3	−2	−1	0	1	2	3
Percentile ranks	0.1	2	16	50	84	98	99.9
IQ	55	70	85	100	115	130	145
SAT	200	300	400	500	600	700	800

FIGURE 15.8 Percentile ranks and standard scores in relation to the normal curve.

http://turtleinvestor888.blogspot.in/2012_01_01_archive.html

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes
  - What is the probability that the roll of a six-faced die will have an even-numbered outcome?

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes
  - What is the probability that the roll of a six-faced die will have an even-numbered outcome?
  - Favorable outcomes are half of the total outcomes, hence the probability is 0.5.

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes
  - What is the probability that the roll of a six-faced die will have an even-numbered outcome?
  - Favorable outcomes are half of the total outcomes, hence the probability is 0.5.
- Problems with this definition:

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes
  - What is the probability that the roll of a six-faced die will have an even-numbered outcome?
  - Favorable outcomes are half of the total outcomes, hence the probability is 0.5.
- Problems with this definition:
  - Cannot be used when the number of outcomes is infinite

Several Ways of Defining the Probability

Ratio of Favorable to Total Outcomes

- Ratio of favorable outcomes to the total number of outcomes
  - What is the probability that the roll of a six-faced die will have an even-numbered outcome?
  - Favorable outcomes are half of the total outcomes, hence the probability is 0.5.
- Problems with this definition:
  - Cannot be used when the number of outcomes is infinite
  - Does not work when the outcomes are not equally likely (e.g., roll of a biased coin)

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?
  - Toss the coin N times and measure the number of times the event T(ails) occurs.

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?
  - Toss the coin N times and measure the number of times the event T(ails) occurs.
- The probability of occurrence of T is defined as the limit or asymptotic value of the relative frequency as $N \to \infty$, i.e., as $P_T = \lim_{N\to\infty} N_T/N$.

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?
  - Toss the coin N times and measure the number of times the event T(ails) occurs.
- The probability of occurrence of T is defined as the limit or asymptotic value of the relative frequency as $N \to \infty$, i.e., as $P_T = \lim_{N\to\infty} N_T/N$.
- Problems with this definition:

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?
  - Toss the coin N times and measure the number of times the event T(ails) occurs.
- The probability of occurrence of T is defined as the limit or asymptotic value of the relative frequency as $N \to \infty$, i.e., as $P_T = \lim_{N\to\infty} N_T/N$.
- Problems with this definition:
  - No one has time to run the experiment infinite number of times

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - What is the probability that the toss of coin will come up Tails?
  - Toss the coin N times and measure the number of times the event T(ails) occurs.
- The probability of occurrence of T is defined as the limit or asymptotic value of the relative frequency as $N \to \infty$, i.e., as $P_T = \lim_{N\to\infty} N_T/N$.
- Problems with this definition:
  - No one has time to run the experiment infinite number of times
  - When N is not large enough, the relative frequency N_T/N will itself be a random number.

Several Ways of Defining the Probability

Relative Frequency Definition

- Run the probabilistic experiment a large number of times (say, N times) and count the number of times (say, N_T) a particular event T occurs. The relative frequency is the ratio N_T/N
  - ▷ What is the probability that the toss of coin will come up Tails?
  - ▷ Toss the coin N times and measure the number of times the event T(ails) occurs.
- The probability of occurrence of T is defined as the limit or asymptotic value of the relative frequency as $N \to \infty$, i.e., as $P_T = \lim_{N\to\infty} N_T/N$.
- Problems with this definition:
  - ▷ No one has time to run the experiment infinite number of times
  - ▷ When N is not large enough, the relative frequency N_T/N will itself be a random number.
    - → If you run the sequence of N experiments today, you will get one answer; when you run the same sequence of N experiments tomorrow, you may get (at least a slightly) different answer
  - ▷ Cannot define the probability using a ratio whose outcome is itself probabilistic!

Several Ways of Defining the Probability

An Axiomatic Set-Theoretic Definition

- The mathematicians struggled with the slippery-slope definitions of the probability
- Eventually, the following systematic approach was proposed by Kolmogorov, which is now widely accepted
  - This approach does not try to answer the question of "how do we know that the probability of the tails coming up on the toss of a coin is 0.5?"
  - Instead it defines a consistent and methodical way of treating the problems related to the probability

Several Ways of Defining the Probability

An Axiomatic Set-Theoretic Definition

- The mathematicians struggled with the slippery-slope definitions of the probability
- Eventually, the following systematic approach was proposed by Kolmogorov, which is now widely accepted
  - This approach does not try to answer the question of "how do we know that the probability of the tails coming up on the toss of a coin is 0.5?"
  - Instead it defines a consistent and methodical way of treating the problems related to the probability
- This approach considers atomic events and a universal set made up of all the atomic events.
- Each set member, and all the possible subsets, are assigned a number between 0 to 1. These numbers are called the probabilities.
- The approach outlined by Kolmogorov ensured that the assignment of the probability numbers to the set members and all the subsets are such that there are no contradictions or inconsistencies

Several Ways of Defining the Probability

An Axiomatic Set-Theoretic Definition

- Examples of the atomic events and the universal sets
  - → Roll a six-sided die:
    - ▷ Set of atomic events: $S = \{1, 2, \ldots, 6\}$
  - → Toss a coin:
    - ▷ Set of atomic events: $S = \{\text{Head, Tail}\}$
- Set S of all atomic events is a certain event. Its probability is 1.
- Set $\emptyset$ is a null event. Its probability is zero.
- Any subset of the set S denotes a probable event. Its probability is a variable between 0 and 1.

Several Ways of Defining the Probability

An Axiomatic Set-Theoretic Definition

- An event A can be defined by its two complementary outputs: $X = 1$ denotes that A has occurred and $X = 0$ denotes the complementary event
  - $\rightarrow$ Let A denote the even-numbered outcome when a die is rolled;
  - $\rightarrow$ $X = 1$ denotes that either $0, 2, 4$ or 6 have come up;
  - $\rightarrow$ $X = 0$ denotes that $1, 3$ or 5 are rolled
- Thus, for $X = i$, $i = 1$ denotes that the event A has occurred, and $i = 0$ denotes the complementary event (i.e., that the event A has not occured)

Several Ways of Defining the Probability

An Axiomatic Set-Theoretic Definition

- An event A can be defined by its two complementary outputs: $X = 1$ denotes that A has occurred and $X = 0$ denotes the complementary event
  - $\rightarrow$ Let A denote the even-numbered outcome when a die is rolled;
  - $\rightarrow$ $X = 1$ denotes that either $0, 2, 4$ or 6 have come up;
  - $\rightarrow$ $X = 0$ denotes that $1, 3$ or 5 are rolled
- Thus, for $X = i$, $i = 1$ denotes that the event A has occurred, and $i = 0$ denotes the complementary event (i.e., that the event A has not occured)
  - $\rightarrow$ The certain event S has a complementary event $\emptyset$
- We will drop the event A of the actual world and just focus on the binary variable X

Axioms of Probability

- Probability $P(X = i)$ is a number which measures the likelihood of event $X = i$.
- Following are three axioms of probability:
  1. $P(X = i) \geq 0$ (i.e., no event has a probability less than zero).
  2. $P(X = i) \leq 1$, and $P(X = i) = 1$ only if $(X = i)$ equals S, i.e., if $X = i$ is a certain event.
  3. Let $X = i$ and $Y = j$ be two events (note that j is also either 0 or 1). The probability of the union of these events is given as $P\left((X = i) \bigcup (Y = j)\right) = P(X = i) + P(Y = j) - P((X = i) \bigcap (Y = j))$.
- The entire theory of probability follows from these three axioms.
  - E.g., the probabilities of mutually exclusive events *add*; for some combination of i and j, $(X = i) \bigcap (Y = j) = \emptyset$, i.e., $P((X = i) \bigcap (Y = j)) = 0$.

A Highly Useful Rule of Probability

Marginalization of the Joint Probability

- Given the joint probability $P((X = i) \bigcap (Y = j))$ (i.e., the probability of the joint event $(X = i) \bigcap (Y = j)$) often abbreviated as $P(X = i, Y = j)$, it is possible to obtain the marginal probability $P(X = i)$ by summing or integrating over both the possibilities of Y.

$$\rightarrow \; P(X = i) = \sum_{j=0}^{1} P(X = i, Y_j)$$

Joint Probability

- The joint probability $P(X = i, Y = j) = P\left(X = i \bigcap Y = j\right)$ is the probability that both $X = i$ *and* $Y = j$ occur

Joint Probability

- The joint probability $P(X = i, Y = j) = P\left(X = i \bigcap Y = j\right)$ is the probability that both $X = i$ *and* $Y = j$ occur
  - $\rightarrow$ Suppose X and Y are both Bernoulli sources, i.e., they take two possibilities each (either 0 or 1). Therefore, the joint event (X, Y) takes four possibilities

Joint Probability

- The joint probability $P(X = i, Y = j) = P\left(X = i \bigcap Y = j\right)$ is the probability that both $X = i$ *and* $Y = j$ occur
  - → Suppose X and Y are both Bernoulli sources, i.e., they take two possibilities each (either 0 or 1). Therefore, the joint event (X, Y) takes four possibilities
  - → For simplicity in the notations, the dependence on i, j is dropped; though it is to be noted that (X, Y) is an event with four possibilities

Joint Probability

- The joint probability $P(X = i, Y = j) = P\left(X = i \bigcap Y = j\right)$ is the probability that both $X = i$ *and* $Y = j$ occur
  - → Suppose X and Y are both Bernoulli sources, i.e., they take two possibilities each (either 0 or 1). Therefore, the joint event (X, Y) takes four possibilities
  - → For simplicity in the notations, the dependence on i, j is dropped; though it is to be noted that (X, Y) is an event with four possibilities

An Example

- Let a student from your batch be picked at random and asked whether he/she wishes to be a part of DA-IICT Dance Club, Sports Club, both or none. Let

An Example

- Let a student from your batch be picked at random and asked whether he/she wishes to be a part of DA-IICT Dance Club, Sports Club, both or none. Let
  - X represent the preference regarding the Sports Club

An Example

- Let a student from your batch be picked at random and asked whether he/she wishes to be a part of DA-IICT Dance Club, Sports Club, both or none. Let
  - X represent the preference regarding the Sports Club
    - $\rightarrow$ $X = 1 \Rightarrow$ the respondent wishes to take part in the Sports Club
    - $\rightarrow$ $X = 0 \Rightarrow$ the respondent does not wish to be a sports person

An Example

- Let a student from your batch be picked at random and asked whether he/she wishes to be a part of DA-IICT Dance Club, Sports Club, both or none. Let
  - X represent the preference regarding the Sports Club
    - $\rightarrow$ $X = 1 \Rightarrow$ the respondent wishes to take part in the Sports Club
    - $\rightarrow$ $X = 0 \Rightarrow$ the respondent does not wish to be a sports person
  - Y represent the choice regarding the Dance Club

An Example

- Let a student from your batch be picked at random and asked whether he/she wishes to be a part of DA-IICT Dance Club, Sports Club, both or none. Let
  - X represent the preference regarding the Sports Club
    - $\rightarrow$ $X = 1 \Rightarrow$ the respondent wishes to take part in the Sports Club
    - $\rightarrow$ $X = 0 \Rightarrow$ the respondent does not wish to be a sports person
  - Y represent the choice regarding the Dance Club
    - $\rightarrow$ $Y = 1 \Rightarrow$ the respondent wishes to take part in the Dance Club
    - $\rightarrow$ $Y = 0 \Rightarrow$ the respondent does not wish to be part of the Dance person
- What are the four possibilities? How to express X and Y in terms of them?

A Highly Useful Rule of Probability

Marginalization of the Joint Probability

- Given the joint probability $P((X=i)\bigcap(Y=j))$ (i.e., the probability of the joint event $(X=i)\bigcap(Y=j)$) often abbreviated as $P(X=i, Y=j)$, it is possible to obtain the marginal probability $P(X=i)$ by summing or integrating over both the possibilities of Y.

$$\rightarrow P(X=i) = \sum_{j=0}^{1} P(X=i, Y_j)$$

Conditional Probability

- Conditional probability $P(X \mid Y)$ is defined as the ratio $\dfrac{P(X,Y)}{P(Y)}$

Conditional Probability

- Conditional probability $P(X \mid Y)$ is defined as the ratio $\dfrac{P(X,Y)}{P(Y)}$
- It denotes the probability that X will occur given Y has occurred

Statistical Independence versus Dependence

- Statistical independence: events X and Y are statistically independent if $P(X, Y) = P(X) \times P(Y)$ (for all four possibilities when X and Y are Bernouilli RVs)
  - $\rightarrow$ If X and Y are statistically independent, $P(X \mid Y) = P(X)$ and $P(Y \mid X) = P(Y)$

Statistical Independence versus Dependence

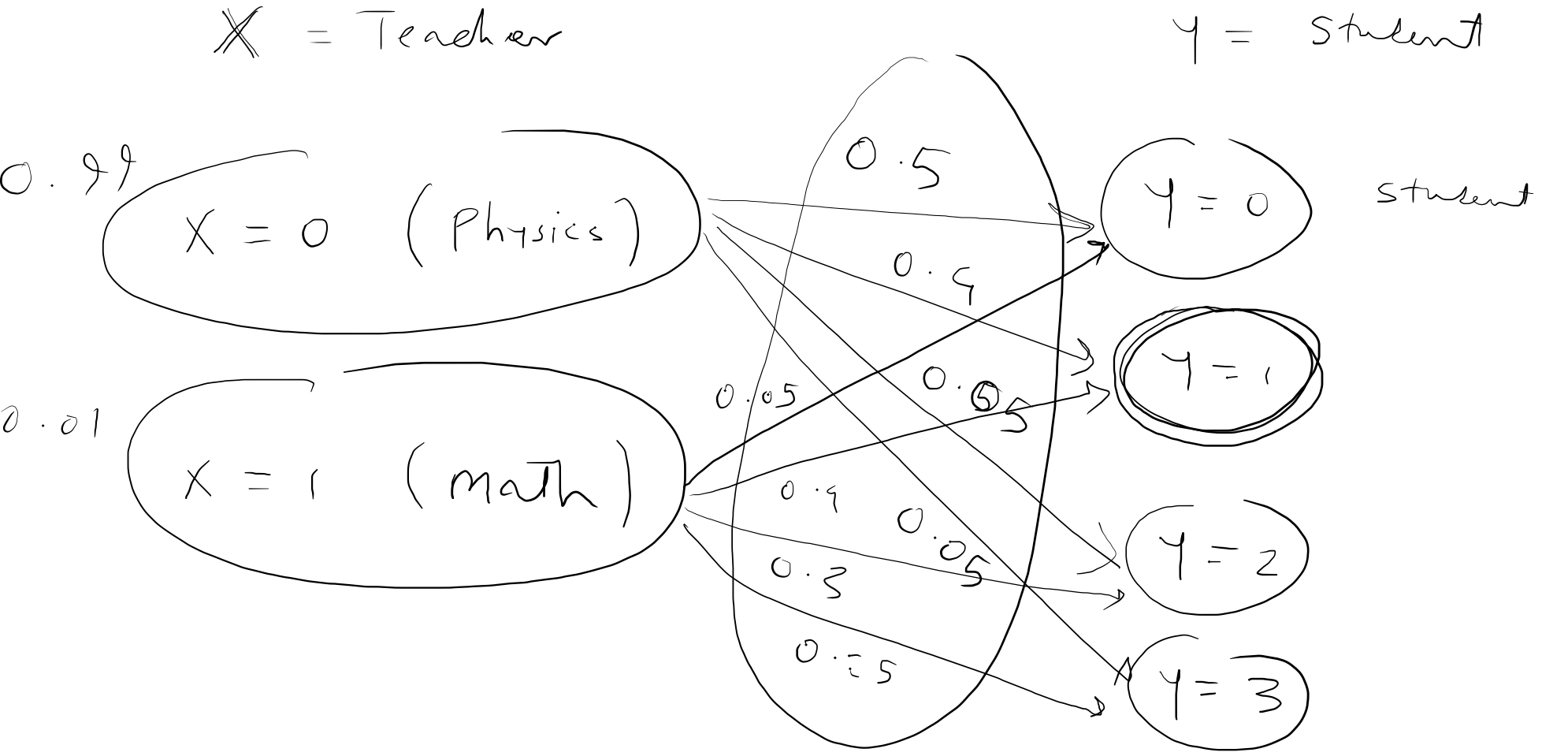
- Statistical independence: events X and Y are statistically independent if $P(X, Y) = P(X) \times P(Y)$ (for all four possibilities when X and Y are Bernouilli RVs)
  - $\rightarrow$ If X and Y are statistically independent, $P(X \mid Y) = P(X)$ and $P(Y \mid X) = P(Y)$
- Statistical Dependence: the joint probabilities cannot be written as the products of the marginal probabilities; the conditional probabilities are not the same as the marginal probabilities

Statistical Independence versus Dependence

- Statistical independence: events X and Y are statistically independent if $P(X, Y) = P(X) \times P(Y)$ (for all four possibilities when X and Y are Bernouilli RVs)
  - $\rightarrow$ If X and Y are statistically independent, $P(X \mid Y) = P(X)$ and $P(Y \mid X) = P(Y)$
- Statistical Dependence: the joint probabilities cannot be written as the products of the marginal probabilities; the conditional probabilities are not the same as the marginal probabilities
- Although the words and the math symbolism may seem abstract and unfamiliar, these are some of the concepts that all of us understand very well!

Statistical Independence versus Dependence

- Statistical independence: events X and Y are statistically independent if $P(X, Y) = P(X) \times P(Y)$ (for all four possibilities when X and Y are Bernouilli RVs)
  - → If X and Y are statistically independent, $P(X \mid Y) = P(X)$ and $P(Y \mid X) = P(Y)$
- Statistical Dependence: the joint probabilities cannot be written as the products of the marginal probabilities; the conditional probabilities are not the same as the marginal probabilities
- Although the words and the math symbolism may seem abstract and unfamiliar, these are some of the concepts that all of us understand very well!
  - → They relate to the notions of influence, causation, correlation, etc.


X = Teacher
Y = Student
0.99
X = 0 (Physics)
0.01
X = 1 (math)
0.5
0.4
0.05
0.05
0.4
0.05
0.3
0.15
Y = 0
Y = 1
Y = 2
Y = 3
student

Statistical Independence versus Dependence

- Statistical independence: events X and Y are statistically independent if $P(X, Y) = P(X) \times P(Y)$ (for all four possibilities when X and Y are Bernouilli RVs)
  - → If X and Y are statistically independent, $P(X \mid Y) = P(X)$ and $P(Y \mid X) = P(Y)$
- Statistical Dependence: the joint probabilities cannot be written as the products of the marginal probabilities; the conditional probabilities are not the same as the marginal probabilities
- Although the words and the math symbolism may seem abstract and unfamiliar, these are some of the concepts that all of us understand very well!
  - → They relate to the notions of influence, causation, correlation, etc.
  - → Probably the best way to come to grips with the concept of the joint and the conditional probabilities is through a framework in which the Bayes' Theorem applies

Bayes Rule

The Causes and the Effects

- The Bayes' Rule deals with the events of the real world which is governed by underlying causes (that are often not observable) and the observed outcomes or the effects

Bayes Rule

The Causes and the Effects

- The Bayes' Rule deals with the events of the real world which is governed by underlying causes (that are often not observable) and the observed outcomes or the effects
- The Bayes' Rule strikes a chord in many ways possibly because it comes close to us, i.e., what we do all the time — our brains probably are programmed to implement the Bayes' Rule without our active awareness

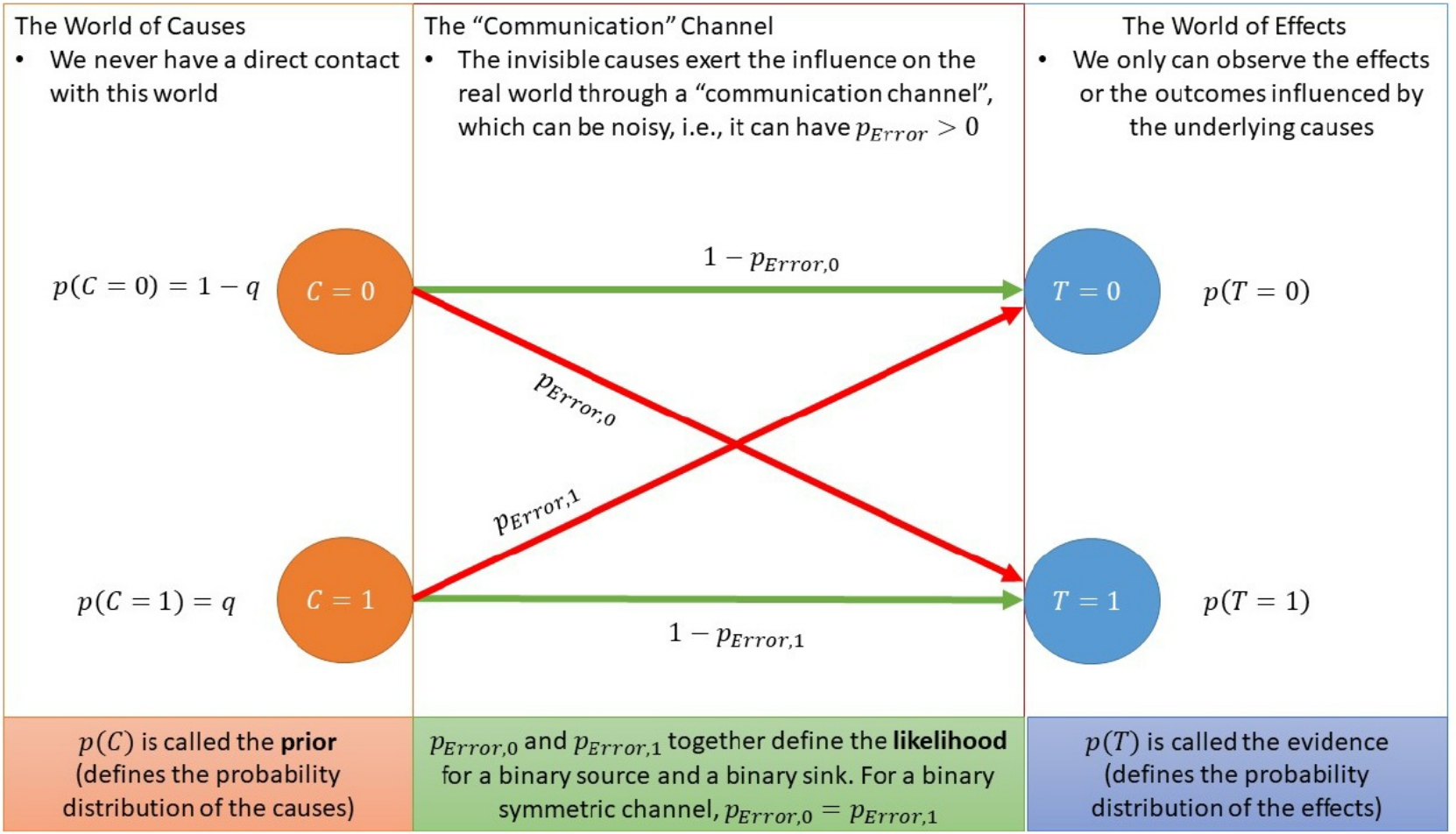
Bayes Rule

The Causes and the Effects

- The Bayes' Rule deals with the events of the real world which is governed by underlying causes (that are often not observable) and the observed outcomes or the effects
- The Bayes' Rule strikes a chord in many ways possibly because it comes close to us, i.e., what we do all the time — our brains probably are programmed to implement the Bayes' Rule without our active awareness
- Due to the Bayes' Rule, the probability theory applies to, e.g., the intensely absorbing episodes of, say, Shakespeare's dramas!

Bayes Rule

A Visualization



Bayes Rule

Terminology

- Let x_i be the cause and y_j be the effect.
  - → $i \in \{0, 1\}$ and $j \in \{0, 1\}$
  - → i.e., the cause can be one of two possibilities, and the observed effect can be one of two possibiliites
  - → We will generalize the number of possibilities later

Bayes Rule

Terminology

- Let x_i be the cause and y_j be the effect.
  - → $i \in \{0, 1\}$ and $j \in \{0, 1\}$
  - → i.e., the cause can be one of two possibilities, and the observed effect can be one of two possibiliites
  - → We will generalize the number of possibilities later
- It is often the case that the conditional probability $p(y_j \mid x_i)$ is available.
  - → This is called the likelihood function or 'forward' probabilities

Bayes Rule

Terminology

- Let x_i be the cause and y_j be the effect.
  - → $i \in \{0, 1\}$ and $j \in \{0, 1\}$
  - → i.e., the cause can be one of two possibilities, and the observed effect can be one of two possibiliites
  - → We will generalize the number of possibilities later
- It is often the case that the conditional probability $p(y_j \mid x_i)$ is available.
  - → This is called the likelihood function
- The probabilities of the causes $p(x_i)$ may be available also — this is called the the prior; and the probabilities of the effect $p(y_j)$ can be computed — this is called the the evidence

Bayes Rule

Terminology

- Let x_i be the cause and y_j be the effect.
  - $\rightarrow$ $i \in \{0, 1\}$ and $j \in \{0, 1\}$
  - $\rightarrow$ i.e., the cause can be one of two possibilities, and the observed effect can be one of two possibiliites
  - $\rightarrow$ We will generalize the number of possibilities later
- It is often the case that the conditional probability $p(y_j \mid x_i)$ is available.
  - $\rightarrow$ This is called the likelihood function
- The probabilities of the causes $p(x_i)$ may be available also — this is called the the prior; and the probabilities of the effect $p(y_j)$ can be computed — this is called the the evidence
- We are oftentimes interested in knowing or *inferring*, given the observation y_j that we have, the most likely x_i that could have caused it. This is called the posterior
- The Bayes' Theorem relates the the likelihood, the prior and the evidence to the posterior

Bayes Rule

The Math Remains the Same; the Problem Description Changes

- Cause x:
  - $\rightarrow$ $x = 1$: patient has (some) disease
  - $\rightarrow$ $x = 0$: patient does not have the disease
- Effect y:
  - $\rightarrow$ $y = 1$: the test that the patient takes is positive
  - $\rightarrow$ $y = 0$: test result is negative
- A person takes the test and $y = 1$, does she/he have the disease, i.e., is $x = 1$?
  - $\rightarrow$ If p and q are known, the posterior on x can be calculated using the Bayes' rule

Toothache and Cavity

The Math Remains the Same; the Problem Description Changes

- Cause x:
  - $\rightarrow$ $x = 1$: your friend has cavity
  - $\rightarrow$ $x = 0$: the friend does not have the cavity
- Effect y:
  - $\rightarrow$ $y = 1$: your friend complains of toothache
  - $\rightarrow$ $y = 0$: she/he does not
- Your friend has the toothache $y = 1$, does she/he have the cavity, i.e., is $x = 1$?
  - $\rightarrow$ If p and q are known, the posterior on x can be calculated using the Bayes' rule

Burglary and Car Alarm

The Math Remains the Same; the Problem Description Changes

- Cause x:
  - $\rightarrow$ $x = 1$: burglars are attempting to steal your car
  - $\rightarrow$ $x = 0$: the car is safe
- Effect y:
  - $\rightarrow$ $y = 1$: you hear the car alarm at 3 am
  - $\rightarrow$ $y = 0$: the car alarm is silent
- You hear the car alarm $y = 1$, are the burglars at the work, i.e., is $x = 1$?
  - $\rightarrow$ Again, the same math can solve the problem provided p and q are known

A Day in the Court

The Math Remains the Same; the Problem Description Changes

- Suppose you're witness to a night-time hit and run accident involving a taxi in Gandhinagar. Of all taxis in Gandhinagar, 1% are light gray, and the rest are white colored. Extensive testing has earlier shown that under dim light conditions, discrimination between the light gray and white is 95% reliable.
- You swear, under oath, that the taxi you saw was gray. Should all the suspects that own the white taxi be released from the jail?

Is It Raining Outside?

The Math Remains the Same; the Problem Description Changes

- Suppose you're a security guard at an underground nuclear facility. Only observation of the outside world you have is when you see the Director coming in. There is a 1% that it rains on any given day. You see the Director coming with an umbrella. When it is raining, there is a 95% chance that the Director has the umbrella with him.
- You see one day Director arriving with his umbrella. Is it raining outside?

Who is the favorite player?

The Math Remains the Same; the Problem Description Changes

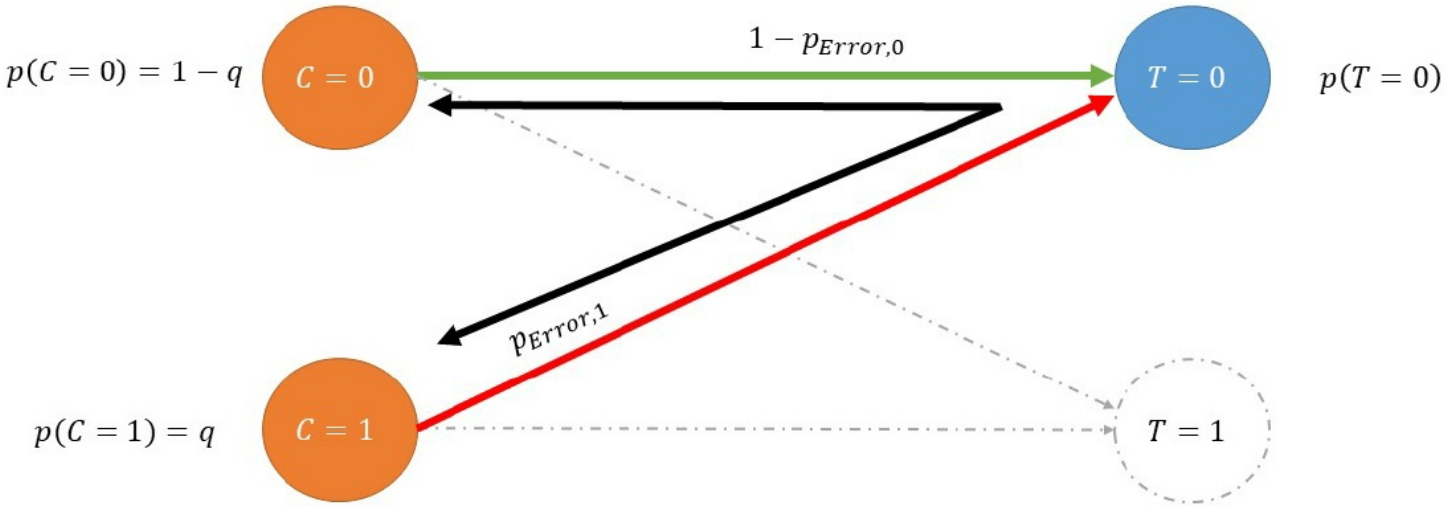
- Suppose who gets to play in the Indian cricket team is influenced to a large extent either by the captain of the team $X = 0$ or the coach $X = 1$. A particular player $Y = 1$ is favored highly (with 95% weight) by the captain, whereas the coach of the team is strongly against including this player in the team, i.e., the coach favors the outcome $Y = 0$ with 95% weight.
- You see that the team declared on the day of the match has the player included in the team, i.e., $Y = 1$. Your immediate reaction is that "oh, this happened only because of ...".

Bayes Rule

A Visualization

Upon the observation of an effect, say, $T = 0$, the task is to determine the likely cause that could have generated this observation, i.e., whether $C = 1$ is more likely compared to $C = 0$ **given** $T = 0$.

- This is called the **posterior** on C given $T = 0$ and it is the conditional probability $\mathrm{p}(C \mid T = 0)$
- This requires the knowledge of $p_{Error,0}$ and $p_{Error,1}$, and the prior $p(C)$

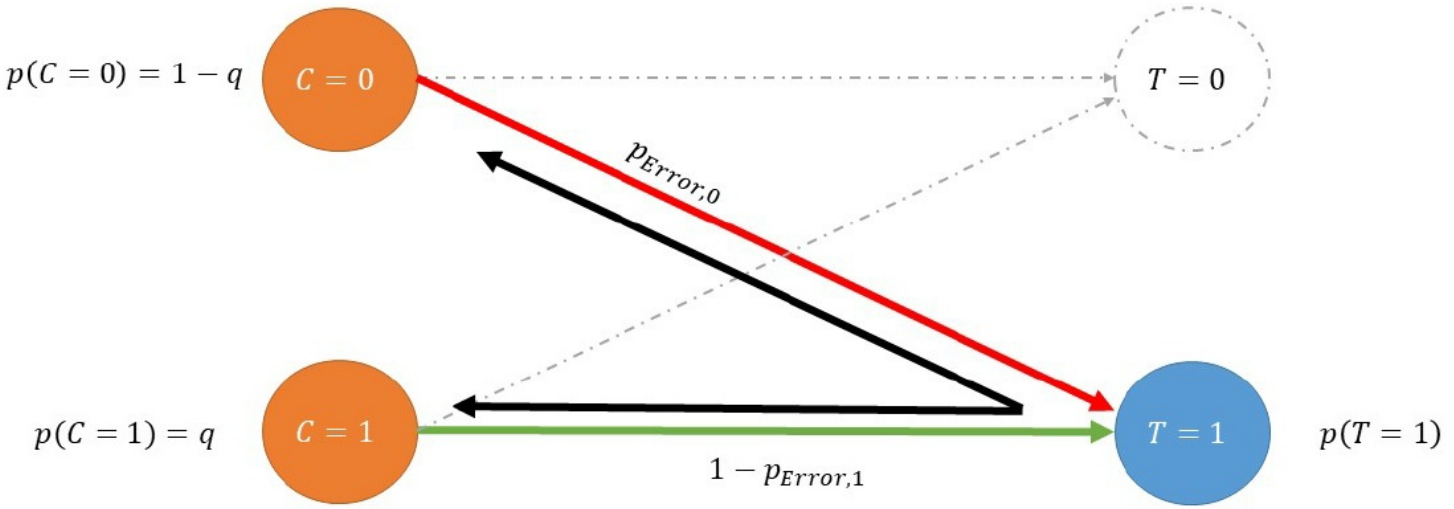


Bayes Rule

A Visualization

If $T = 1$ is observed instead, the posterior on C given $T = 1$, i.e., $p(C \mid T = 1)$, is required to be computed

- This case requires the knowledge of $p_{Error,1}$ and $p_{Error,0}$, and the prior $p(C)$



Bayes Rule

The Main Formula

- Since $p(y_j, x_i) = p(y_j \mid x_i) \times p(x_i) = p(x_i \mid y_j) \times p(y_j)$, we have:

Theorem (Bayes')

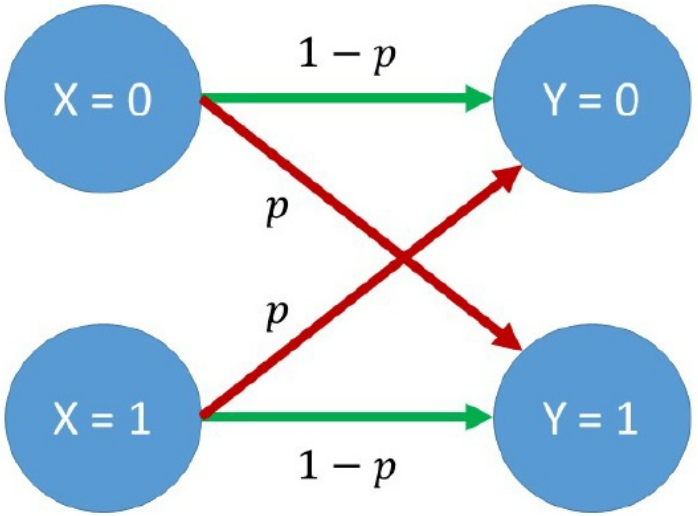
$$p(x_i \mid y_j) = \frac{p(y_j \mid x_i) \times p(x_i)}{p(y_j)}$$

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

Binary Symmetric Channel or BSC

- The likelihood function $p(Y = 1 \mid X = 0)$ is identical to $p(Y = 0 \mid X = 1)$. Both equal p
- The prior on the Bernouilli variable X is denoted as q

$1 - q = 0.5$



$q = 0.5$

(1) $P(Y)$

(2) $P(X|Y)$

(3) $H(X)$

(4) $H(X|Y)$

(5) $I(X;Y)$

$P(Y/X)$		$Y=0$	$Y=1$
$X=0$	0.5	$1-p$	p
$X=1$	0.5	p	$1-p$
	↑	↑	
	Prior	Likelihood $P(Y/X)$	

$P(X,Y)$	$Y=0$	$Y=1$
$X=0$	$(1-p)0.5$	$p(0.5)$
$X=1$	$p(0.5)$	$(1-p)0.5$
	0.5	0.5

$$P(X,Y) = P(Y/X)P(X)$$

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

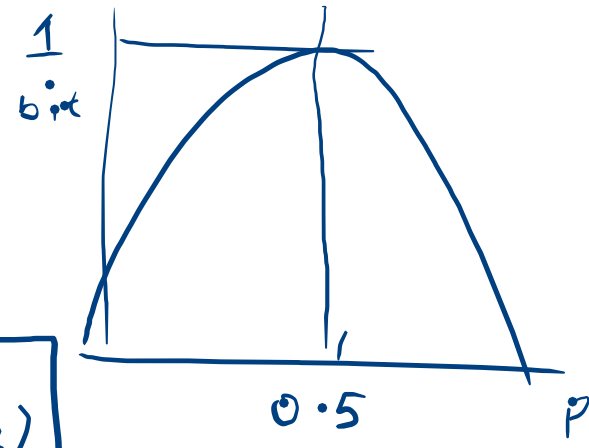
	$y=0$	$y=1$
$x=0$	$1-p$	p
$x=1$	p	$1-p$

$I(\tilde{X}_0 = 0)$ $I(\tilde{X}_0 = 1)$

$-\log_2(1-p)$ $-\log_2 p$

$\tilde{X}_0$ $\tilde{X}_1$

(3) $H(X) = 1$ bit



$$H_2(x) = -x \log_2(x) - (1-x) \log_2(1-x)$$

← Binary Entropy Function

(4) $H(X|Y)$

(4.1) $H(\tilde{X}_0) = H(X|Y=0) = H_2(p)$

(4.2) $H(\tilde{X}_1) = H(X|Y=1) = H_2(1-p) = H_2(p)$

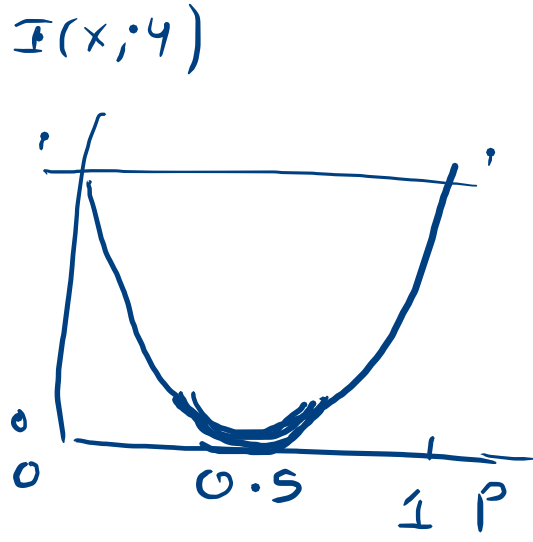
(4) $H(X|Y) = P(Y=0)\,H(X|Y=0) + P(Y=1)\,H(X|Y=1)$

$= 0.5\,H_2(p) + 0.5\,H_2(p)$

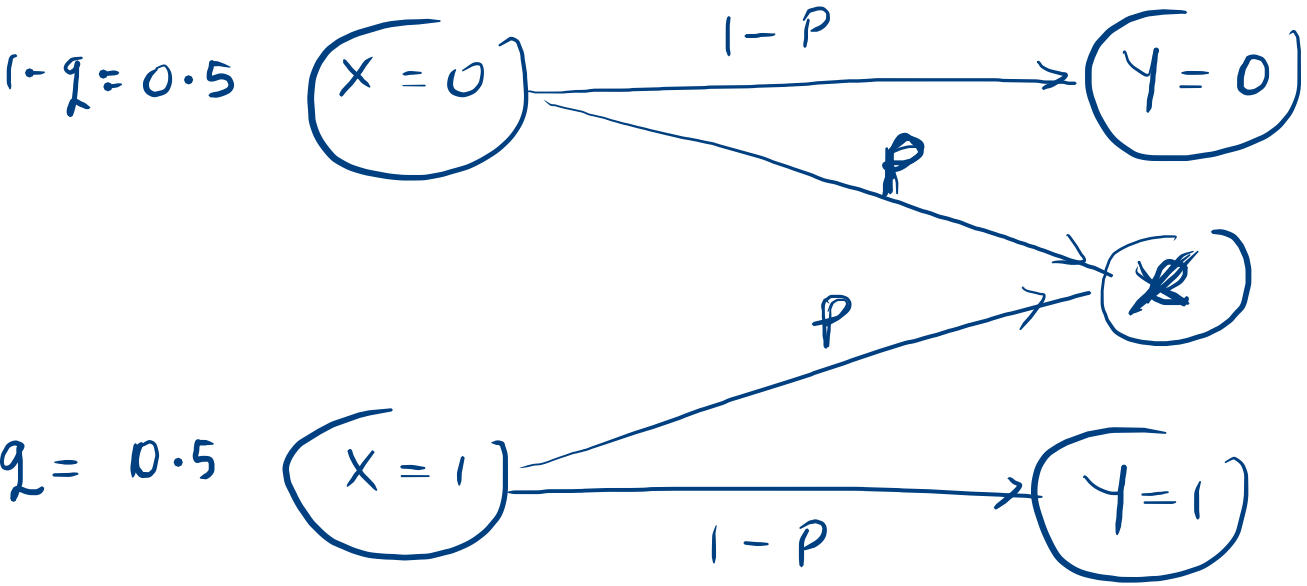
$= H_2(p)$

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

$= 1 - H_2(p)$



Binary Erasure channel (BEC(p))



(1) $P(Y)$

(2) $P(X|Y)$

(3) $H(X)$

(4) $H(X|Y)$

(5) $I(X;Y)$

(Plot This function $I(X;Y)$)

Bayes Rule

The MAP Decision Rule

- Sometimes we don't care about the exact probabilities; all that we want to know, given $Y = 1$, is whether $X = 1$ is more likely than $X = 0$

Bayes Rule

The MAP Decision Rule

- Sometimes we don't care about the exact probabilities; all that we want to know, given $Y = 1$, is whether $X = 1$ is more likely than $X = 0$
  - E.g., the only allowed answers to the patient's query "do I have the disease?" may be "yes" or "no"; the former implies that the patient has to be hospitalized, the latter means the hospitalization is not needed. Note that the hospitalization is a binary decision, the patient cannnot be partially hosptialized!

Bayes Rule

The MAP Decision Rule

- Sometimes we don't care about the exact probabilities; all that we want to know, given $Y = 1$, is whether $X = 1$ is more likely than $X = 0$
  - E.g., the only allowed answers to the patient's query "do I have the disease?" may be "yes" or "no"; the former implies that the patient has to be hospitalized, the latter means the hospitalization is not needed. Note that the hospitalization is a binary decision, the patient cannnot be partially hosptialized!
- Given this question, the obvious strategy, when $Y = 1$ is observed, is to say "yes" if $p(X = 1 \mid Y = 1) > p(X = 0 \mid Y = 1)$ and say "no" otherwise.

Bayes Rule

The MAP Decision Rule

- Sometimes we don't care about the exact probabilities; all that we want to know, given $Y = 1$, is whether $X = 1$ is more likely than $X = 0$
  - E.g., the only allowed answers to the patient's query "do I have the disease?" may be "yes" or "no"; the former implies that the patient has to be hospitalized, the latter means the hospitalization is not needed. Note that the hospitalization is a binary decision, the patient cannnot be partially hosptialized!
- Given this question, the obvious strategy, when $Y = 1$ is observed, is to say "yes" if $p(X = 1 \mid Y = 1) > p(X = 0 \mid Y = 1)$ and say "no" otherwise.
- This is called the Maximum A-Posteriori (MAP) decision rule

Bayes Rule

The MAP Decision Rule

- The MAP rule:

Bayes Rule

The MAP Decision Rule

- The MAP rule:
  - $\rightarrow$ Given an observation Y_j, evaluate $p(X_i \mid Y_j)$ for all i, and select that cause i for which $p(X_i|Y_j)$ is the maximum

Bayes Rule

The MAP Decision Rule

- The MAP rule:
  - → Given an observation Y_j, evaluate $p(X_i \mid Y_j)$ for all i, and select that cause i for which $p(X_i|Y_j)$ is the maximum
  - → When X is binary, this amounts to evaluating the ratio of the posteriors when $Y = 1$

$$\lambda_1 = \frac{p(X=1|Y=1)}{p(X=0|Y=1)} = \frac{p(Y=1|X=1)p(X=1)}{p(Y=0|X=1)p(X=0)}$$
$$= \frac{(1-p)\,q}{p\,(1-q)}$$

This λ_1 is called the "odds in favor of $X = 1$".

Bayes Rule

The MAP Decision Rule

- The MAP rule:
  - → Given an observation Y_j, evaluate $p(X_i \mid Y_j)$ for all i, and select that cause i for which $p(X_i|Y_j)$ is the maximum
  - → When X is binary, this amounts to evaluating the ratio of the posteriors when $Y = 1$

$$\lambda_1 = \frac{p(X=1|Y=1)}{p(X=0|Y=1)} = \frac{p(Y=1|X=1)p(X=1)}{p(Y=0|X=1)p(X=0)}$$
$$= \frac{(1-p)\,q}{p\,(1-q)}$$

Prior on X

This λ_1 is called the "odds in favor of $X = 1$".

  - → If the odds in favor of $X = 1$ exceeds 1, $X = 1$ is more likely than $X = 0$.
  - → According to the MAP rule, the receiver decides $\hat{X} = 1$; else it sets $\hat{X} = 0$. (here, $\hat{X}$ is the decision formed by the receiver about the transmitted bit X)

Bayes Rule

The MAP and the ML Decision Rules

- The MAP rule simplifies to Maximum Likelihood (ML) rule when all the priors are equally likely.
  - → Given an observation Y_j, select that i for which $p(Y_j \mid X_i)$ is maximum
  - → Note that Bayes rule is not explicitly invoked
  - → When X is binary, the ML rule amounts to evaluating the ratio of the likelihoods:

$$\lambda_1 = \frac{1-p}{p}$$

Thus, the ratio of the posteriors becomes simply the likelihood ratio (LR): $(1-p)/p$

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

$$\underline{\lambda_0} = \frac{p(X = 0 \,|\, Y = 0)}{p(X = 1 \,|\, Y = 0)} = \frac{p(Y = 0 \,|\, X = 0)p(X = 0)}{p(Y = 0 \,|\, X = 1)p(X = 0)}$$
$$= \frac{(1 - p)\,(1 - q)}{p\,q}$$

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

$$\lambda_0 = \frac{p(X=0 \,|\, Y=0)}{p(X=1 \,|\, Y=0)} = \frac{p(Y=0 \,|\, X=0)p(X=0)}{p(Y=0 \,|\, X=1)p(X=0)}$$
$$= \frac{(1-p)\,(1-q)}{p\,q}$$

- The MAP receiver decides $\hat{X} = 0$ if $\lambda_0 \geq 1$; else it decides $\hat{X} = 1$

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

$$\lambda_0 = \frac{p(X = 0 \,|\, Y = 0)}{p(X = 1 \,|\, Y = 0)} = \frac{p(Y = 0 \,|\, X = 0)p(X = 0)}{p(Y = 0 \,|\, X = 1)p(X = 0)}$$
$$= \frac{(1 - p)\,(1 - q)}{p\,q}$$

- The MAP receiver decides $\hat{X} = 0$ if $\lambda_0 \geq 1$; else it decides $\hat{X} = 1$
- If $q = 0.5$, the above degenerates to the ML rule:

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

$$\lambda_0 = \frac{p(X = 0 \,|\, Y = 0)}{p(X = 1 \,|\, Y = 0)} = \frac{p(Y = 0 \,|\, X = 0)p(X = 0)}{p(Y = 0 \,|\, X = 1)p(X = 0)}$$
$$= \frac{(1 - p)\,(1 - q)}{p\,q}$$

- The MAP receiver decides $\hat{X} = 0$ if $\lambda_0 \geq 1$; else it decides $\hat{X} = 1$
- If $q = 0.5$, the above degenerates to the ML rule:

$$\lambda_0 = \frac{1 - p}{p}$$

The MAP and the ML Rules for the BSC

The Ratio of the Posteriors when $Y = 0$ is received

- The ratio of the posteriors when $Y = 0$ is received:

$$\lambda_0 = \frac{p(X = 0 \,|\, Y = 0)}{p(X = 1 \,|\, Y = 0)} = \frac{p(Y = 0 \,|\, X = 0)p(X = 0)}{p(Y = 0 \,|\, X = 1)p(X = 0)}$$
$$= \frac{(1 - p)\,(1 - q)}{p\,q}$$

- The MAP receiver decides $\hat{X} = 0$ if $\lambda_0 \geq 1$; else it decides $\hat{X} = 1$
- If $q = 0.5$, the above degenerates to the ML rule:

$$\lambda_0 = \frac{1 - p}{p}$$

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$
  - Similarly, if $Y = 0$ is received, λ_0 typically exceeds 1, i.e., the source PMF favors the possibility that $X = 0$

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$
  - Similarly, if $Y = 0$ is received, λ_0 typically exceeds 1, i.e., the source PMF favors the possibility that $X = 0$
  - Note that the receiver never compares λ_1 and λ_0 with each other — after a given reception, the receiver either calculates λ_1 (if $Y = 1$ is received) or λ_0 (if $Y = 0$). The receiver compares either λ_1 or λ_0 against 1

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$
  - Similarly, if $Y = 0$ is received, λ_0 typically exceeds 1, i.e., the source PMF favors the possibility that $X = 0$
  - Note that the receiver never compares λ_1 and λ_0 with each other — after a given reception, the receiver either calculates λ_1 (if $Y = 1$ is received) or λ_0 (if $Y = 0$). The receiver compares either λ_1 or λ_0 against 1
- We have here assumed that the receiver knows p and q. Is this a fair assumption? Maybe. In many cases, the exact knowledge of p and q may not be available and so they may be estimated

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$
  - Similarly, if $Y = 0$ is received, λ_0 typically exceeds 1, i.e., the source PMF favors the possibility that $X = 0$
  - Note that the receiver never compares λ_1 and λ_0 with each other — after a given reception, the receiver either calculates λ_1 (if $Y = 1$ is received) or λ_0 (if $Y = 0$). The receiver compares either λ_1 or λ_0 against 1
- We have here assumed that the receiver knows p and q. Is this a fair assumption? Maybe. In many cases, the exact knowledge of p and q may not be available and so they may be estimated
- The values of λ_1 or λ_0 indicate the belief — a value of 1 implies both the causes are equally likely, values either much greater or much less than 1 imply a strong belief about the likely cause

Observations

- In typical cases, the source (i.e., its PMF) appears vastly different depending on whether $Y = 0$ is received versus $Y = 1$ is received
  - If $Y = 1$ is received, λ_1 typically exceeds 1, i.e., the source PMF becomes *biased* toward the possibility that $X = 1$
  - Similarly, if $Y = 0$ is received, λ_0 typically exceeds 1, i.e., the source PMF favors the possibility that $X = 0$
  - Note that the receiver never compares λ_1 and λ_0 with each other — after a given reception, the receiver either calculates λ_1 (if $Y = 1$ is received) or λ_0 (if $Y = 0$). The receiver compares either λ_1 or λ_0 against 1
- We have here assumed that the receiver knows p and q. Is this a fair assumption? Maybe. In many cases, the exact knowledge of p and q may not be available and so they may be estimated
- The values of λ_1 or λ_0 indicate the belief — a value of 1 implies both the causes are equally likely, values either much greater or much less than 1 imply a strong belief about the likely cause
- The prior acts as a "dead weight" that prevents these beliefs to move by much

An Example Calculation

The Ratio of the Posteriors = LR $\times$ the Ratio of the Priors

- For the C-disease problem,

An Example Calculation

The Ratio of the Posteriors = LR $\times$ the Ratio of the Priors

- For the C-disease problem,
  - $\rightarrow$ $p = 0.05$ and $q = 0.01$

An Example Calculation

The Ratio of the Posteriors = LR $\times$ the Ratio of the Priors

- For the C-disease problem,
  - $\rightarrow$ $p = 0.05$ and $q = 0.01$
- Therefore, if $Y = 1$, the posterior ratio

$$\lambda_1 = \frac{(1-p)\,q}{p\,(1-q)} \approx \frac{q}{p} = 0.01/0.05 = 0.2$$

An Example Calculation

The Ratio of the Posteriors = LR $\times$ the Ratio of the Priors

- For the C-disease problem,
  - $\rightarrow$ $p = 0.05$ and $q = 0.01$
- Therefore, if $Y = 1$, the posterior ratio

$$\lambda_1 = \frac{(1-p)\,q}{p\,(1-q)} \approx \frac{q}{p} = 0.01/0.05 = 0.2$$

- Let $p(X = 1 \mid Y = 1)$ be denote as q'. Thus,

$$\lambda_1 = \frac{q'}{1-q'} = 0.2 \Rightarrow q' = 0.16$$

- The person detected positive is quite unlikely to have the C-disease.

An Example Calculation

When the Ratio of the Priors is Unity, The Ratio of the Posteriors = LR

- For the C-disease problem, suppose

An Example Calculation

When the Ratio of the Priors is Unity, The Ratio of the Posteriors = LR

- For the C-disease problem, suppose
  - $\rightarrow$ $p = 0.05$ and $q = 0.5$

An Example Calculation

When the Ratio of the Priors is Unity, The Ratio of the Posteriors = LR

- For the C-disease problem, suppose
  - $\rightarrow$ $p = 0.05$ and $q = 0.5$
- Now, the posterior ratio $\lambda_1 = q'/(1-q')$ becomes identical to the likelihood ratio $(1-p)/p = 0.95/0.05 = 19$

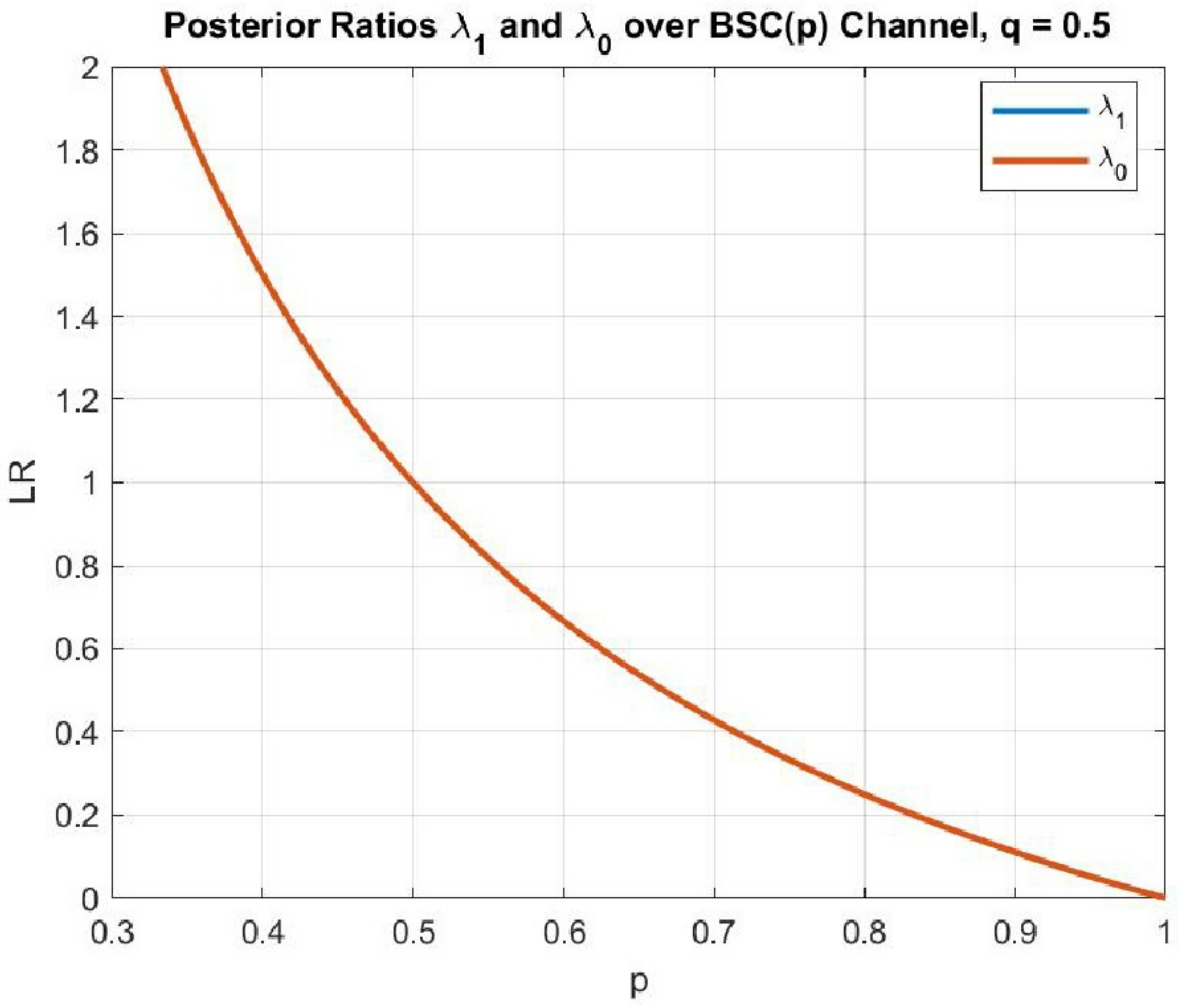
An Example Calculation

When the Ratio of the Priors is Unity, The Ratio of the Posteriors = LR

- For the C-disease problem, suppose
  - $\rightarrow$ $p = 0.05$ and $q = 0.5$
- Now, the posterior ratio $\lambda_1 = q'/(1-q')$ becomes identical to the likelihood ratio $(1-p)/p = 0.95/0.05 = 19$
- This shows that $q' = 0.95$. The person detected positive is now highly likely to have the C-disease

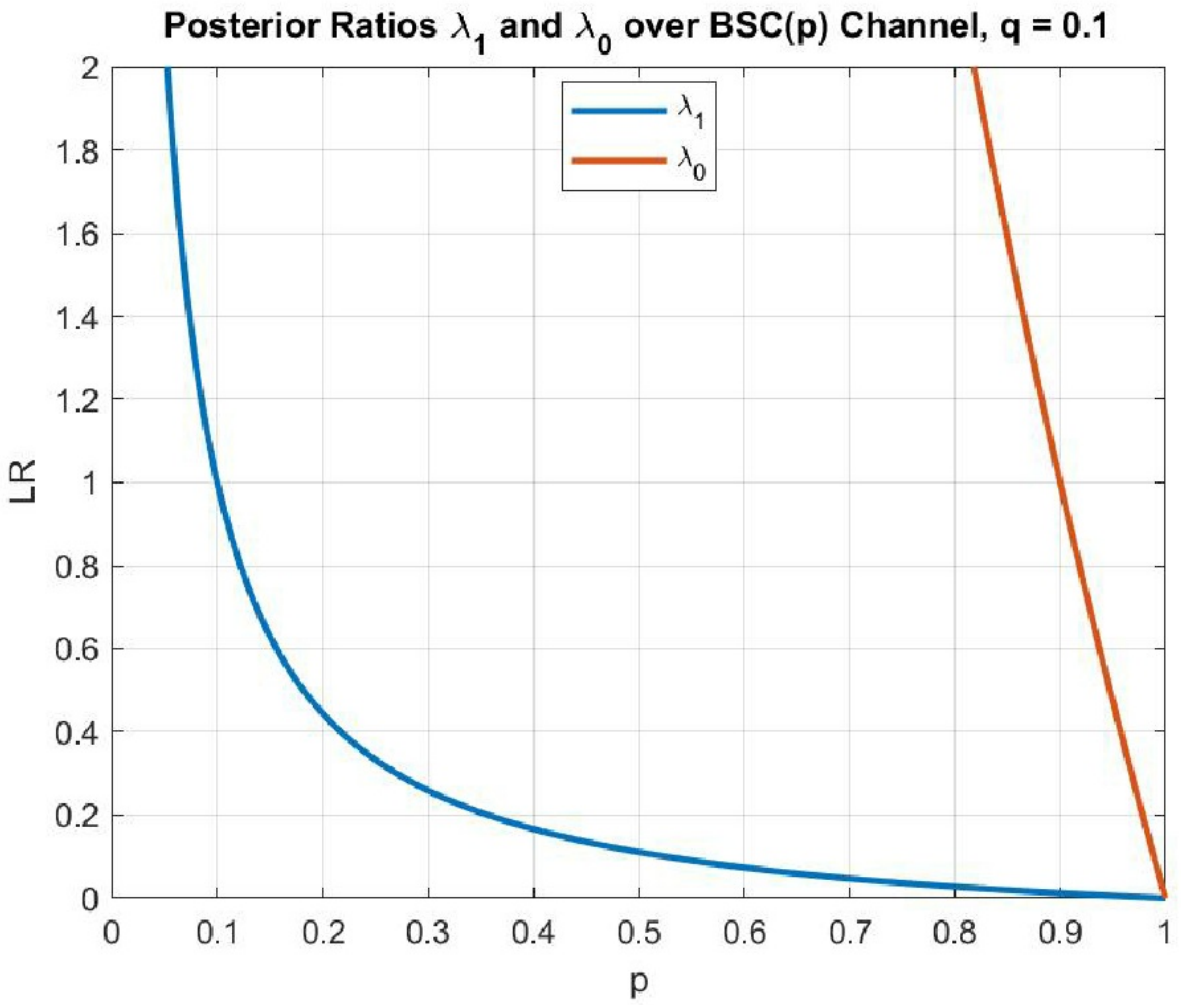
Posterior Ratios $\lambda_1 = \lambda_0 =$ Likelihood Ratio $(1-p)/p$

over BSC(p) Channel, $q = 0.5$



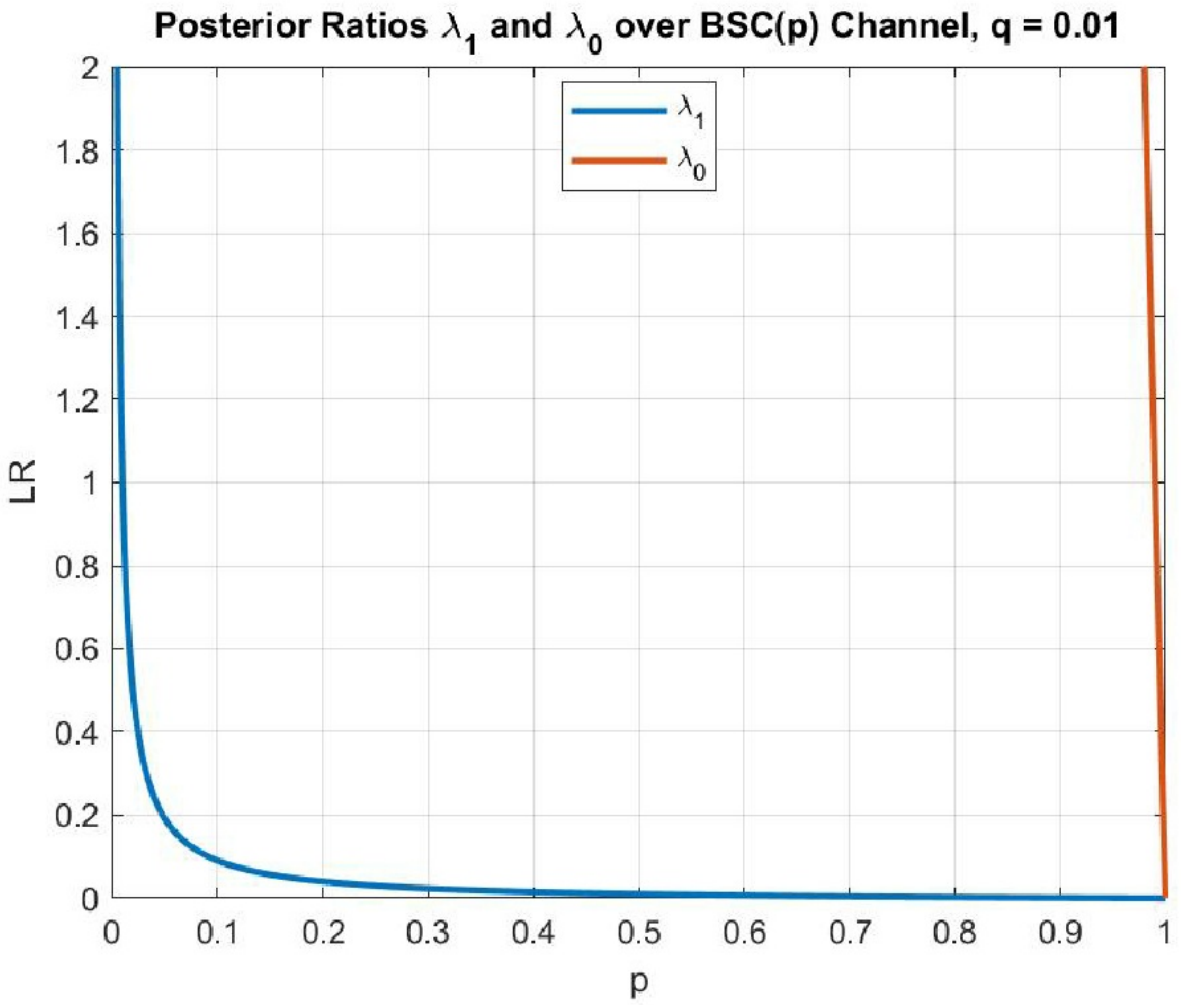
Posterior Ratios λ_1, and λ_0

over BSC(p) Channel, $q = 0.1$



Posterior Ratios λ_1, and λ_0

over BSC(p) Channel, $q = 0.01$



Repeated Tests

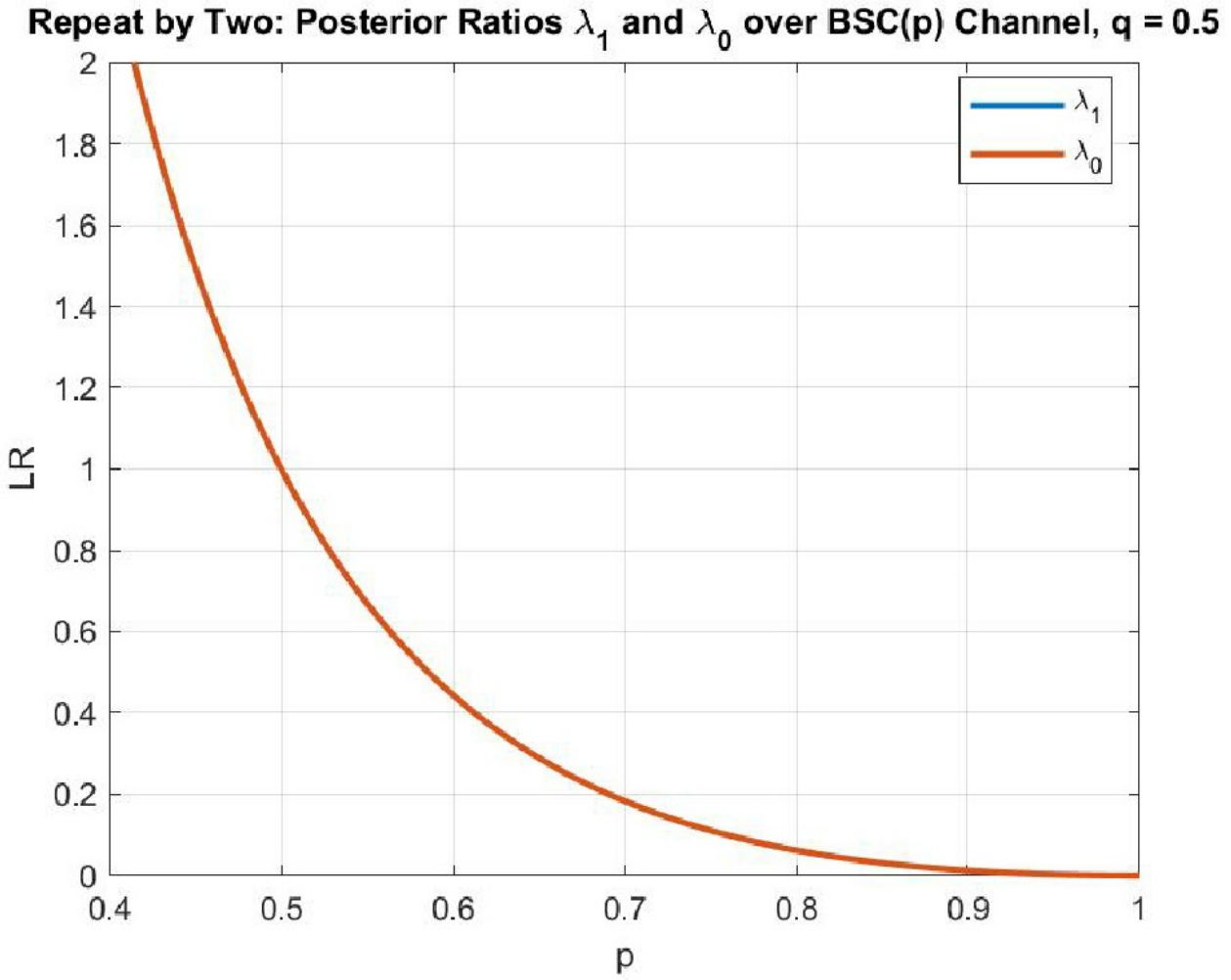
- Suppose your friend tests positive for C-disease. For further confirmation, she decides to take another test in which she again tests positive.

Repeated Tests

- Suppose your friend tests positive for C-disease. For further confirmation, she decides to take another test in which she again tests positive.
- What is the probability that she has the disease? (Take $p = 0.05$. Use $q = 0.5, 0.1$, and 0.01).

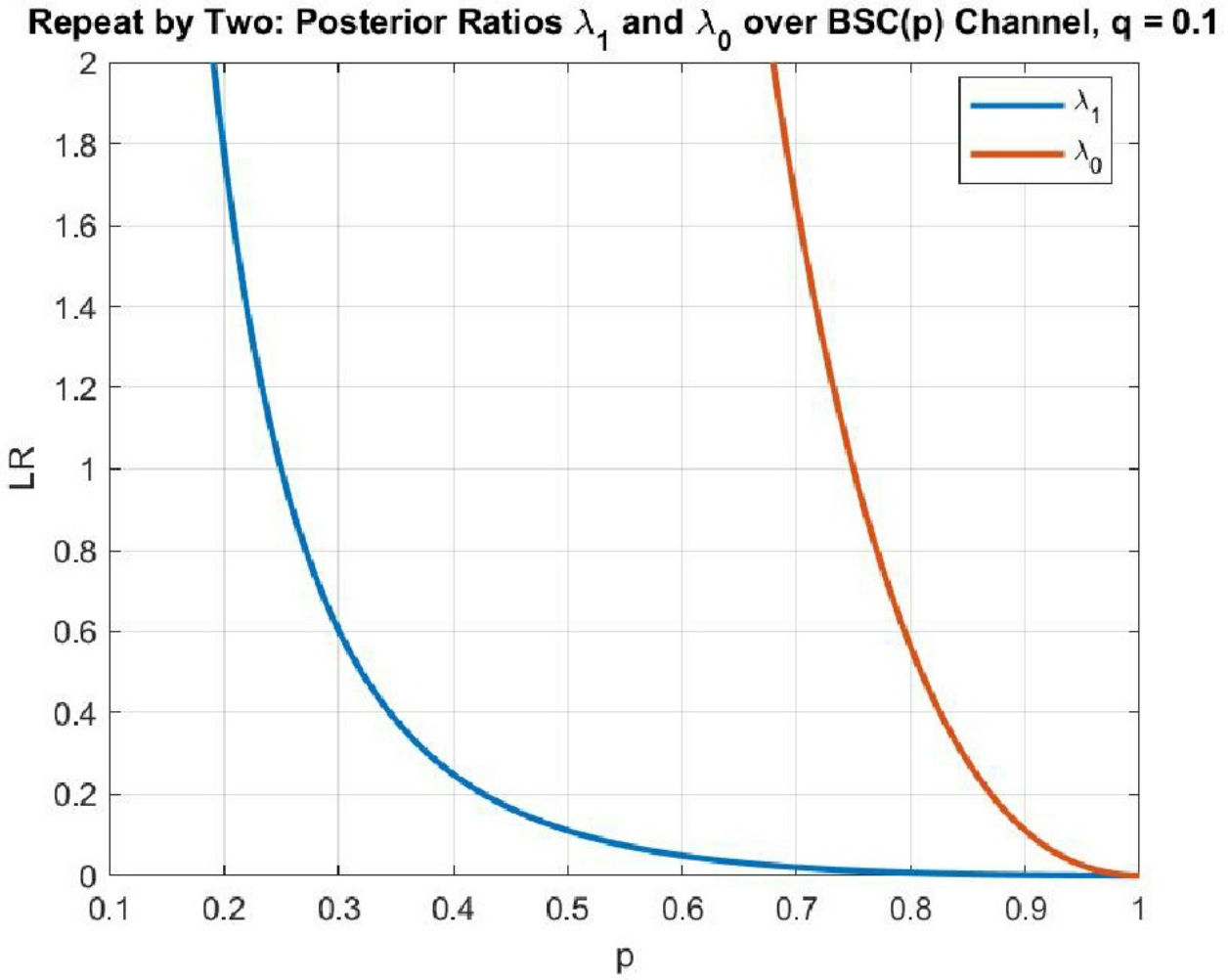
Posterior Ratios $\lambda_1 = \lambda_0$

over BSC(p) Channel, $q = 0.5$. Repeated Transmission.



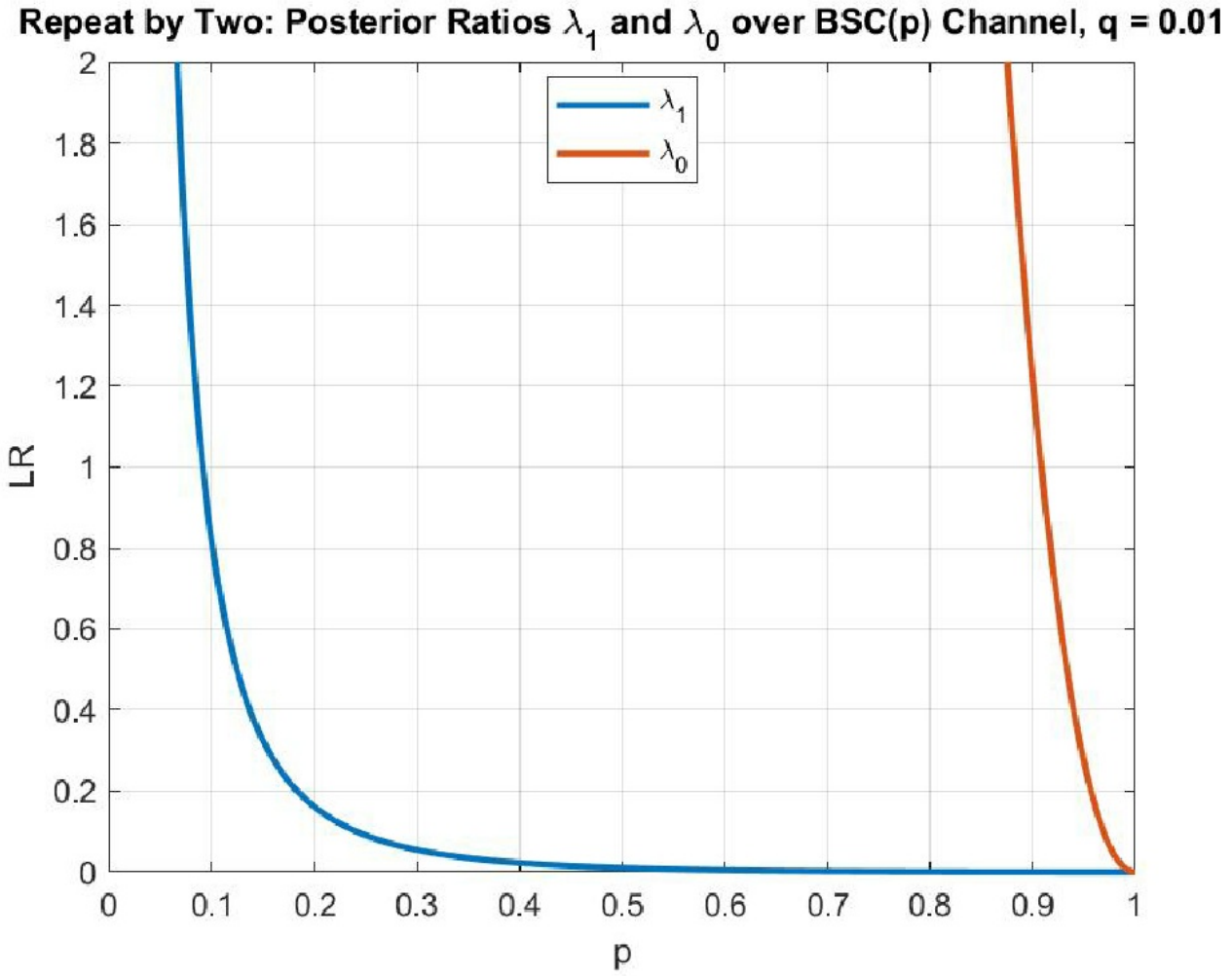
Posterior Ratios λ_1, and λ_0

over BSC(p) Channel, $q = 0.1$. Repeated Transmission.



Posterior Ratios λ_1, and λ_0

over BSC(p) Channel, $q = 0.01$. Repeated Transmission.



The Conditional Entropy

- We have earlier seen that the Entropy $H(X)$ of the source can be thought of as its temperature
- Bayesian framework tells us that we never observe a source X directly; we observe it through the evidence Y we obtain

The Conditional Entropy

- We have earlier seen that the Entropy $H(X)$ of the source can be thought of as its temperature
- Bayesian framework tells us that we never observe a source X directly; we observe it through the evidence Y we obtain
- A question: what is the Entropy of the source after we obtain the evidence Y?

The Conditional Entropy

- We have earlier seen that the Entropy $H(X)$ of the source can be thought of as its temperature
- Bayesian framework tells us that we never observe a source X directly; we observe it through the evidence Y we obtain
- A question: what is the Entropy of the source after we obtain the evidence Y?
  - This is called the Conditional Entropy $H(X \mid Y)$, called the entropy of X given Y.

The Conditional Entropy

- We have earlier seen that the Entropy $H(X)$ of the source can be thought of as its temperature
- Bayesian framework tells us that we never observe a source X directly; we observe it through the evidence Y we obtain
- A question: what is the Entropy of the source after we obtain the evidence Y?
  - This is called the Conditional Entropy $H(X \mid Y)$, called the entropy of X given Y.
- Hopefully, the source temperature, after we receive the information-bearing message Y, should reduce. If not, the information transfer cannot be said to have occurred

The Conditional Entropy

- The conditional entropy when the observed effect Y takes a particular value j is given as

$$H(X \mid Y = j) = \mathbb{E}\left[I_k \mid Y = j\right] = -\sum_{k=1}^{K} p(X_k \mid Y = j) \log_2 \left(p(X_k \mid Y = j)\right)$$

The Conditional Entropy

- The conditional entropy when the observed effect Y takes a particular value j is given as

$$H(X \mid Y = j) = \mathbb{E}\left[I_k \mid Y = j\right] = -\sum_{k=1}^{K} p(X_k \mid Y = j) \log_2 \left(p(X_k \mid Y = j)\right)$$

- The conditional entropy is obtained by averaging over all realizations of the variable Y, i.e., as

$$\begin{aligned} H(X \mid Y) &= \mathbb{E}\left[\mathbb{E}\left[I_k \mid Y = j\right]\right] \\ &= -\sum_{j=1}^{J} p(Y = j) \sum_{k=1}^{K} p(X_k \mid Y = j) \log_2 \left(p(X_k \mid Y = j\right. \end{aligned}$$

The Information Transfer

- Having defined the entropies $H(X)$ and $H(X \mid Y)$ of the source *before* and *after* the reception of the message Y, we can now define the information that is transferred to the receiver after it receives the message

The Information Transfer

- Having defined the entropies $H(X)$ and $H(X \mid Y)$ of the source *before* and *after* the reception of the message Y, we can now define the information that is transferred to the receiver after it receives the message
  - In our analogy, the information transferred equals the amount by which the *temperature* of the source reduces because it has received the message

The Information Transfer

- Having defined the entropies $H(X)$ and $H(X \mid Y)$ of the source *before* and *after* the reception of the message Y, we can now define the information that is transferred to the receiver after it receives the message
  - In our analogy, the information transferred equals the amount by which the *temperature* of the source reduces because it has received the message
- The information transfer achieved is defined as the difference between $H(X)$ and $H(X \mid Y)$; specifically,

$$I(X \mid Y = j) = H(X) - H(X \mid Y = j)$$
$$I(X \mid Y) = H(X) - H(X \mid Y)$$

How can the bits take fractional values?

- We have seen that the concepts of information is obtained in terms of fraction of bits. However, how can the bits be in fractional units?

How can the bits take fractional values?

- We have seen that the concepts of information is obtained in terms of fraction of bits. However, how can the bits be in fractional units?
- Shannon's answer: one bit at a time cannot be in fraction. However, the best way to think about the information generation and its transfer is to consider a large sequence of bits