CSC 4020 Fundamental of Machine Learning: Introduction to Probability and Information Theory

Baoyuan Wu
School of Data Science, CUHK-SZ

January 18/20, 2021

Outline

1. A brief review of last week
2. Probability distributions of discrete random variables
3. Probability distribution of continuous random variables
4. Some common discrete distributions
5. Some common continuous distributions
6. Information theory

Definition and branches of machine learning

Basic concepts, learning process, model selection

Two clarifications

- Extra score of answering questions: each time 0.5, and 5 maximum; not every question has the extra score, and I will design some scored and non-trivial questions (may be 3) at each course

Two clarifications

- Extra score of answering questions: each time 0.5, and 5 maximum; not every question has the extra score, and I will design some scored and non-trivial questions (may be 3) at each course
- About the course addition: there are 32 extra applications but I rejected all of them, to ensure the course quality. The rejected students are encouraged to choose other related courses or join in the BB system to access the slides.

Discrete random variables

- Binary event:
  - An event is denoted as A. For example, "it will rain tomorrow", "The temperature will increase tomorrow".
  - $P(A)$ denotes the probability that the event A will happen. We require that $0 \leq P(A) \leq 1$, and $P(A) = 0$ means that A definitely will not happen, while $P(A) = 1$ means that A definitely will happen.
  - $P(\bar{A}) = 1 - P(A)$ denotes the probability that the event A will not happen.

Discrete random variables

- Binary event:
  - An event is denoted as A. For example, "it will rain tomorrow", "The temperature will increase tomorrow".
  - $P(A)$ denotes the probability that the event A will happen. We require that $0 \leq P(A) \leq 1$, and $P(A) = 0$ means that A definitely will not happen, while $P(A) = 1$ means that A definitely will happen.
  - $P(\bar{A}) = 1 - P(A)$ denotes the probability that the event A will not happen.
- Discrete random variable:
  - We can extend A to the **discrete random variable** X, which can take any value from a finite set $\mathcal{X}$, which is called **state space**. For example, $\mathcal{X} = \{1, 2, 3, 4, 5\}$.
  - $P(X = x)$ denotes the probability of the event $X = x$, or just $P(x)$ for short. $P(\cdot)$ is called a **probability mass function** or **pmf**.
  - We require that $0 \leq P(x) \leq 1$ and $\sum_{x \in \mathcal{X}} P(x) = 1$.

Joint, marginal, conditional probability

- Probability of a union of two events: Given two events A and B, we define the probability of A or B as follows:

$$\begin{aligned} P(A \vee B) &= P(A) + P(B) - P(A \wedge B) \\ &= P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned} \tag{1}$$

Joint, marginal, conditional probability

- Probability of a union of two events: Given two events A and B, we define the probability of A or B as follows:

$$\begin{aligned} P(A \vee B) &= P(A) + P(B) - P(A \wedge B) \\ &= P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned} \tag{1}$$

- Joint probabilities: The probability of the joint event A and B is defined as follows:

$$P(A, B) = P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A), \tag{2}$$

It is called the **product rule**.

Joint, marginal, conditional probability

- Probability of a union of two events: Given two events A and B, we define the probability of A or B as follows:

$$\begin{aligned} P(A \vee B) &= P(A) + P(B) - P(A \wedge B) \\ &= P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually exclusive} \end{aligned} \tag{1}$$

- Joint probabilities: The probability of the joint event A and B is defined as follows:

$$P(A, B) = P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A), \tag{2}$$

  It is called the **product rule**.

- Marginal distribution: Given the above joint distribution, we can define the **marginal distribution** as follows:

$$P(A) = \sum_b P(A, B) = \sum_b P(A|B = b)P(B = b), \tag{3}$$

  which sums over all possible states of B. It is called the **sum rule**.

Conditional probability and Bayes rule

- Conditional probability: Recalculating probability of event A after someone tells you that event B happened, as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

$$P(A \cap B) = P(A|B)P(B) \tag{5}$$

- Bayes Rule: Combining the definition of conditional probability with the product and sum rules yields Bayes rule, as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, \tag{6}$$

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{7}$$

Conditional probability and Bayes rule

- Conditional probability: Recalculating probability of event A after someone tells you that event B happened, as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

$$P(A \cap B) = P(A|B)P(B) \tag{5}$$

Conditional probability and Bayes rule

- Conditional probability: Recalculating probability of event A after someone tells you that event B happened, as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4}$$

$$P(A \cap B) = P(A|B)P(B) \tag{5}$$

- Bayes Rule: Combining the definition of conditional probability with the product and sum rules yields Bayes rule, as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}, = \frac{P(A,B)}{P(A)} \tag{6}$$

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{7}$$

Application of Bayes rule: medical diagnosis

- Suppose that you do a medical test for breast cancer, the test result could be *positive* or *negative*. We denote $x = 1$ as the event of positive test, while $x = 0$ as the event of negative test. We denote $y = 1$ as the event of having breast cancer, while $y = 0$ as the event of no breast cancer.

Application of Bayes rule: medical diagnosis

- Suppose that you do a medical test for breast cancer, the test result could be *positive* or *negative*. We denote $x = 1$ as the event of positive test, while $x = 0$ as the event of negative test. We denote $y = 1$ as the event of having breast cancer, while $y = 0$ as the event of no breast cancer.
- Suppose that if one has breast cancer, the test will be positive with the probability 0.8, *i.e.*,

$$P(x = 1|y = 1) = 0.8. \tag{8}$$

Application of Bayes rule: medical diagnosis

- Suppose that you do a medical test for breast cancer, the test result could be *positive* or *negative*. We denote $x = 1$ as the event of positive test, while $x = 0$ as the event of negative test. We denote $y = 1$ as the event of having breast cancer, while $y = 0$ as the event of no breast cancer.
- Suppose that if one has breast cancer, the test will be positive with the probability 0.8, *i.e.*,

$$P(x = 1|y = 1) = 0.8. \tag{8}$$

- Then, if one gets a positive test result, what is the probability of having breast cancer? $P(y = 1|x = 1) = 0.8$?

Application of Bayes rule: medical diagnosis

- It is WRONG! It ignores the prior probability of having breast cancer, which is fortunately quite low:

$$P(y = 1) = 0.004. \tag{9}$$

Application of Bayes rule: medical diagnosis

- It is WRONG! It ignores the prior probability of having breast cancer, which is fortunately quite low:

$$P(y = 1) = 0.004. \tag{9}$$

- We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$P(x = 1|y = 0) = 0.1. \tag{10}$$

Application of Bayes rule: medical diagnosis

- It is WRONG! It ignores the prior probability of having breast cancer, which is fortunately quite low:

$$P(y = 1) = 0.004. \tag{9}$$

- We also need to take into account the fact that the test may be a **false positive** or **false alarm**. Unfortunately, such false positives are quite likely (with current screening technology):

$$P(x = 1|y = 0) = 0.1. \tag{10}$$

- Combining all above probabilities using Bayes rule, we can compute $P(y = 1|x = 1)$ as follows:

$$\begin{aligned} P(y = 1|x = 1) &= \frac{P(x = 1|y = 1)P(y = 1)}{P(x = 1|y = 1)P(y = 1) + P(x = 1|y = 0)P(y = 0)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031. \end{aligned} \tag{11}$$

It tells that if you test positive, you have have about a 3% chance of really having breast cancer!

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \Longleftrightarrow P(X, Y) = P(X)P(Y). \tag{12}$$

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y). \tag{12}$$

- Given the marginal independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3-1=2$ and $4-1$ free parameters to define $P(X)$ and $P(Y)$, respectively.

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y). \tag{12}$$

- Given the marginal independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3-1=2$ and $4-1$ free parameters to define $P(X)$ and $P(Y)$, respectively.
- If without the marginal independence, how many free parameters do we need to define the joint probability $P(X, Y)$?

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y). \tag{12}$$

- Given the marginal independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3-1=2$ and $4-1$ free parameters to define $P(X)$ and $P(Y)$, respectively.
- If without the marginal independence, how many free parameters do we need to define the joint probability $P(X, Y)$? $(3 \times 4) - 1 = 11$.

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y). \tag{12}$$

- Given the marginal independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3-1=2$ and $4-1$ free parameters to define $P(X)$ and $P(Y)$, respectively.
- If without the marginal independence, how many free parameters do we need to define the joint probability $P(X, Y)$? $(3 \times 4) - 1 = 11$.
- If given this marginal independence, *i.e.*, $P(X, Y) = P(X)P(Y)$, how many free parameters do we need?

Marginal independence and conditional independence

- Marginal/unconditional independent: If X and Y are independent, denoted as $X \perp Y$, then the joint probability can be represented as the product of two marginals, *i.e.*,

$$X \perp Y \iff P(X, Y) = P(X)P(Y). \tag{12}$$

- Given the marginal independence, we can use fewer parameters to define a joint probability. Suppose that X has 3 states, Y has 4 states, then we need $3-1=2$ and $4-1$ free parameters to define $P(X)$ and $P(Y)$, respectively.
- If without the marginal independence, how many free parameters do we need to define the joint probability $P(X, Y)$? $(3 \times 4) - 1 = 11$.
- If given this marginal independence, *i.e.*, $P(X, Y) = P(X)P(Y)$, how many free parameters do we need? $(3-1) + (4-1) = 5$.

Conditional independence

- Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct.

Conditional independence

- Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct.
- We therefore say X and Y are conditionally independent (CI) given is Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z \iff P(X,Y|Z) = P(X|Z)P(Y|Z). \tag{13}$$

Conditional independence

- Unfortunately, unconditional independence is rare, because most variables can influence most other variables. However, usually this influence is mediated via other variables rather than being direct.
- We therefore say X and Y are conditionally independent (CI) given is Z iff the conditional joint can be written as a product of conditional marginals:

$$X \perp Y|Z \iff P(X,Y|Z) = P(X|Z)P(Y|Z). \tag{13}$$

- CI assumptions allow us to build large probabilistic models from small pieces. We will see more examples in the later section **Graphical models**.

Expectation and variance of discrete random varibales

- Expectation (or mean): $E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$

Expectation and variance of discrete random varibales

- Expectation (or mean): $E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$
- Expectation of a function: $E(f(X)) = \sum_{x \in \mathcal{X}} f(x)P(X = x)$

Expectation and variance of discrete random varibales

- Expectation (or mean): $E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$
- Expectation of a function: $E(f(X)) = \sum_{x \in \mathcal{X}} f(x)P(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$

Expectation and variance of discrete random varibales

- Expectation (or mean): $E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$
- Expectation of a function: $E(f(X)) = \sum_{x \in \mathcal{X}} f(x)P(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2. \tag{14}$$

Expectation and variance of discrete random varibales

- Expectation (or mean): $E(X) = \sum_{x \in \mathcal{X}} xP(X = x)$
- Expectation of a function: $E(f(X)) = \sum_{x \in \mathcal{X}} f(x)P(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2. \tag{14}$$

- Standard deviation: Square root of variance, *i.e.*,

$$\text{Std} = \sqrt{\text{Var}(X)}. \tag{15}$$

- Illustration and examples: [on board]

Continuous random variables

- A random variable X is continuous if its sample space $\mathcal{X}$ is uncountable.
- In this case, $P(X = x) = 0$ for each x.

Continuous random variables

- A random variable X is continuous if its sample space $\mathcal{X}$ is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function (PDF) for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{16}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{17}$$

Continuous random variables

- A random variable X is continuous if its sample space $\mathcal{X}$ is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function (PDF) for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{16}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{17}$$

- The cumulative distribution function (CDF) is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.

Continuous random variables

- A random variable X is continuous if its sample space $\mathcal{X}$ is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function (PDF) for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{16}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{17}$$

- The cumulative distribution function (CDF) is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- More generally: If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx \tag{18}$$

Bivariate continuous distributions: Marginalization, Conditioning and Independence

- $p_{X,Y}(x, y)$, joint probablity density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$

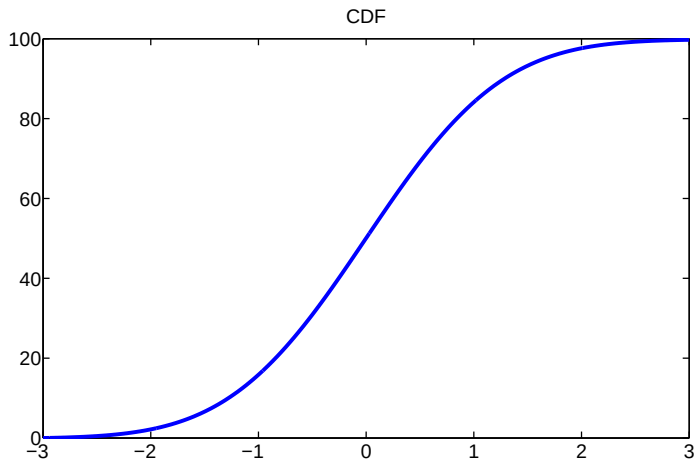
Bivariate continuous distributions: Marginalization, Conditioning and Independence

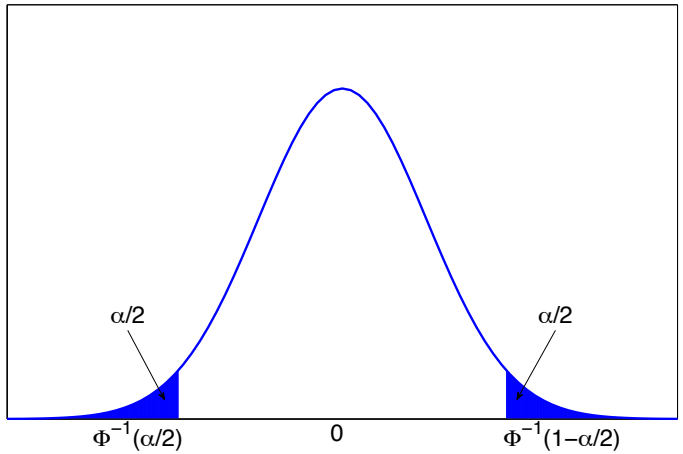
- $p_{X,Y}(x, y)$, joint probablity density function of X and Y
- $\int_x \int_y p(x, y) dx dy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y) dy$

Bivariate continuous distributions: Marginalization, Conditioning and Independence

- $p_{X,Y}(x, y)$, joint probablity density function of X and Y
- $\int_x \int_y p(x, y)dxdy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$! Formally, conditional probability in the continuous case can be derived using infinitesimal events.
- Independence: X and Y are independent if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$

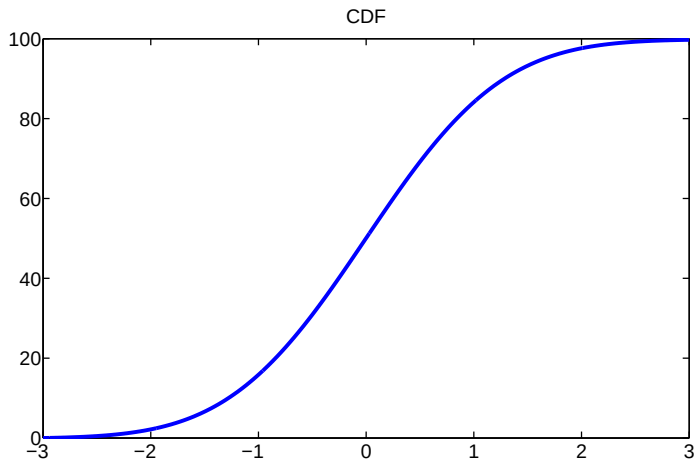
Quantiles

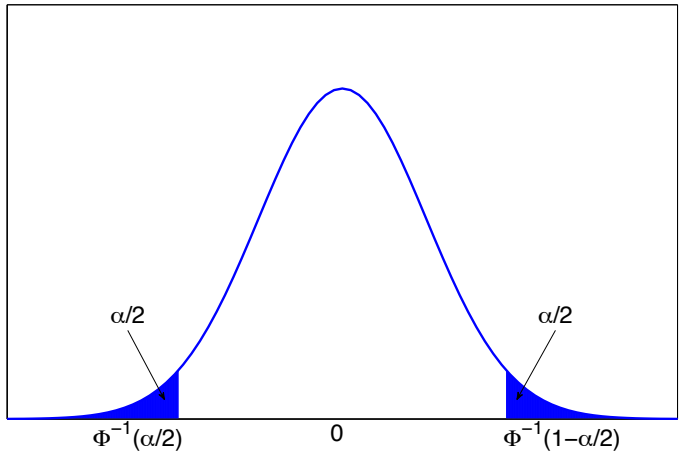




- Since the CDF F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}.

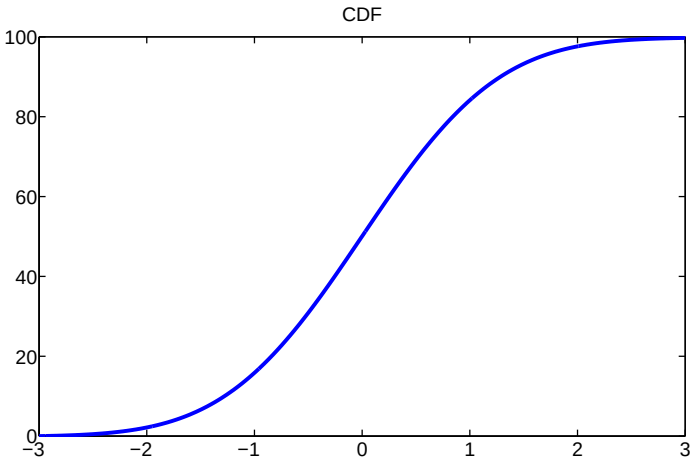
Quantiles

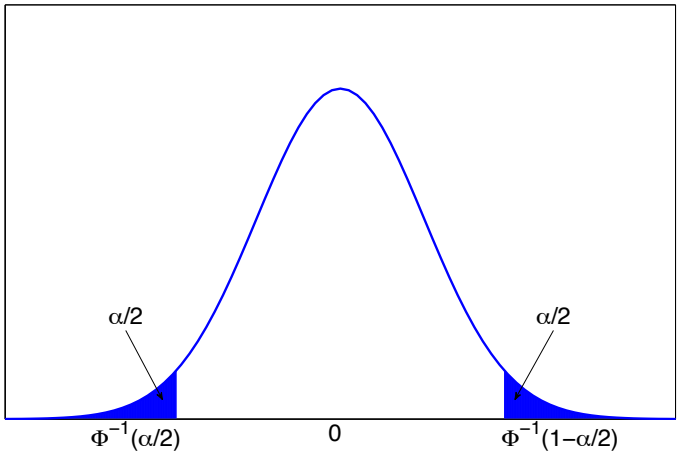




- Since the CDF F is a monotonically increasing function, it has an inverse; let us denote this by F^{-1}.
- If F is the CDF of X, then $F^1(\alpha)$ is the value of x_α such that $P(X \leq x_\alpha) = \alpha$; this is called the a quantile of F. The value $F^1(0.5)$ is the median of the distribution, with half of the probability mass on the left, and half on the right. The values $F^1(0.25)$ and $F^1(0.75)$ are the **lower** and **upper quartiles**.

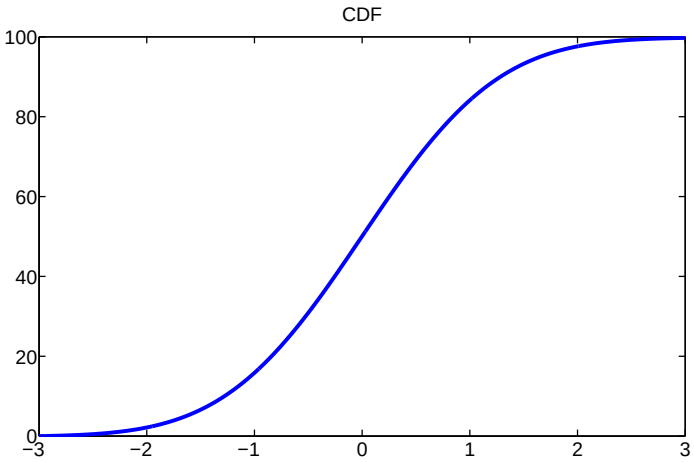
Quantiles

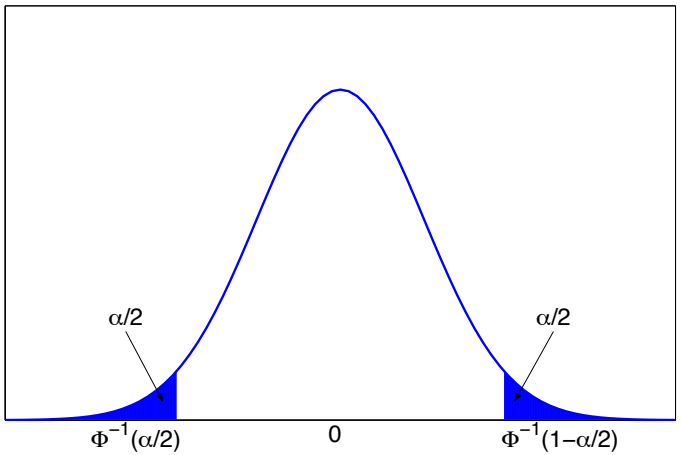




- We can also use the inverse CDF to compute tail area probabilities.

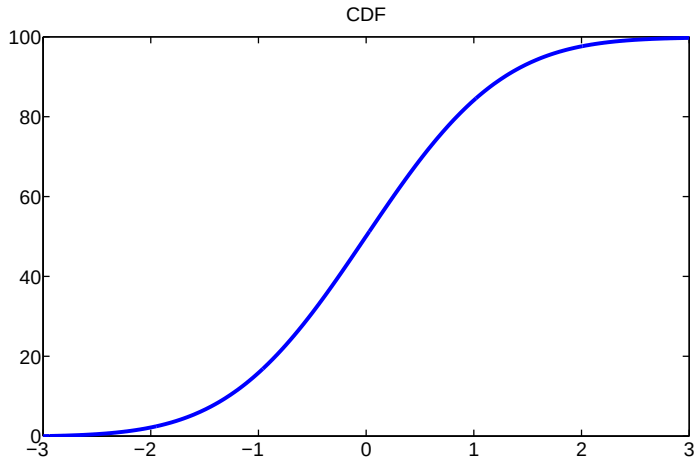
Quantiles

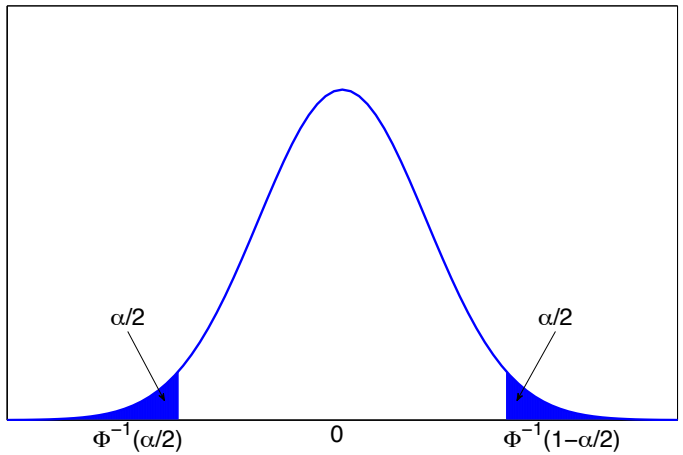




- We can also use the inverse CDF to compute tail area probabilities.
- For example, if Φ is the CDF of the Gaussian distribution $\mathcal{N}(0,1)$, then points to the left of $\Phi^{-1}(\alpha/2)$ contain $\alpha/2$ probability mass. By symmetry, points to the right of $\Phi^{-1}(1-\alpha/2)$ also contain $\alpha/2$ probability mass.

Quantiles





- We can also use the inverse CDF to compute tail area probabilities.
- For example, if Φ is the CDF of the Gaussian distribution $\mathcal{N}(0,1)$, then points to the left of $\Phi^{-1}(\alpha/2)$ contain $\alpha/2$ probability mass. By symmetry, points to the right of $\Phi^{-1}(1-\alpha/2)$ also contain $\alpha/2$ probability mass.
- Hence, the central interval $(\Phi^{-1}(\alpha/2), \Phi^{-1}(1-\alpha/2))$ contains $1-\alpha$ of the mass. If we set $\alpha = 0.05$, the central 95% interval is covered by the range

$$(\Phi^{-1}(0.025), \Phi^{-1}(0.975)) = (-1.96, 1.96). \tag{19}$$

For a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the central 95% interval is $(\mu - 1.96\sigma, \mu + 1.96\sigma)$.

Expectation and variance of continuous random variables

Similar to that of discrete random variables, only change the summation $\sum$ to the integral $\int$.

- Expectation (or mean): $\mu = E(X) = \int_{\mathcal{X}} xP(X = x)$

Expectation and variance of continuous random variables

Similar to that of discrete random variables, only change the summation $\sum$ to the integral $\int$.

- Expectation (or mean): $\mu = E(X) = \int_{\mathcal{X}} xP(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$

Expectation and variance of continuous random variables

Similar to that of discrete random variables, only change the summation $\sum$ to the integral $\int$.

- Expectation (or mean): $\mu = E(X) = \int_{\mathcal{X}} xP(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2. \tag{20}$$

Expectation and variance of continuous random variables

Similar to that of discrete random variables, only change the summation $\sum$ to the integral $\int$.

- Expectation (or mean): $\mu = E(X) = \int_{\mathcal{X}} xP(X = x)$
- Moments: expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2 = M_2 - M_1^2. \tag{20}$$

- Standard deviation: Square root of variance, *i.e.*,

$$\text{Std} = \sqrt{\text{Var}(X)}. \tag{21}$$

- Illustration and examples: [on board]

Binary variables

- We firstly consider the probability of a binary random variable $x \in \{0, 1\}$. Suppose that you toss a coin, and $x = 1$ denotes the event of 'heads', while $x = 0$ indicates the event of 'tails'.

Binary variables

- We firstly consider the probability of a binary random variable $x \in \{0, 1\}$. Suppose that you toss a coin, and $x = 1$ denotes the event of 'heads', while $x = 0$ indicates the event of 'tails'.
- The probability of $x = 1$ is described by a parameter μ,

$$p(x = 1|\mu) = \mu, \tag{22}$$

where $\mu \in [0, 1]$, and we can obtain that $p(x = 0|\mu) = 1 - \mu$.

Binary variables

- We firstly consider the probability of a binary random variable $x \in \{0, 1\}$. Suppose that you toss a coin, and $x = 1$ denotes the event of 'heads', while $x = 0$ indicates the event of 'tails'.
- The probability of $x = 1$ is described by a parameter μ,

$$p(x = 1|\mu) = \mu, \tag{22}$$

  where $\mu \in [0, 1]$, and we can obtain that $p(x = 0|\mu) = 1 - \mu$.
- The probability distribution over x can therefore be written in the form

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \tag{23}$$

  which is called Bernoulli distribution.

Binary variables

- We firstly consider the probability of a binary random variable $x \in \{0, 1\}$. Suppose that you toss a coin, and $x = 1$ denotes the event of 'heads', while $x = 0$ indicates the event of 'tails'.
- The probability of $x = 1$ is described by a parameter μ,

$$p(x = 1|\mu) = \mu, \tag{22}$$

  where $\mu \in [0, 1]$, and we can obtain that $p(x = 0|\mu) = 1 - \mu$.
- The probability distribution over x can therefore be written in the form

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \tag{23}$$

  which is called Bernoulli distribution.
- Its mean and variance are

$$\mathbb{E}[x] = \sum_x x\text{Bern}(x|\mu) = \mu, \tag{24}$$

$$\text{var}[x] = \mathbb{E}[(x - \mu)^2] = \mu(1 - \mu) \tag{25}$$

Binary variables

- Imagine that you toss the coin N times, and each tossing follows the Bernoulli distribution $p(x|\mu)$. We denote the variable m as the numbers of heads, then its distribution is formulated as follows:

$$\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}, \tag{26}$$

which is called Binomial distribution, where

$$\binom{N}{m} = \frac{N!}{(N-m)!m!}. \tag{27}$$

Binary variables

- Imagine that you toss the coin N times, and each tossing follows the Bernoulli distribution $p(x|\mu)$. We denote the variable m as the numbers of heads, then its distribution is formulated as follows:

$$\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}, \tag{26}$$

  which is called Binomial distribution, where

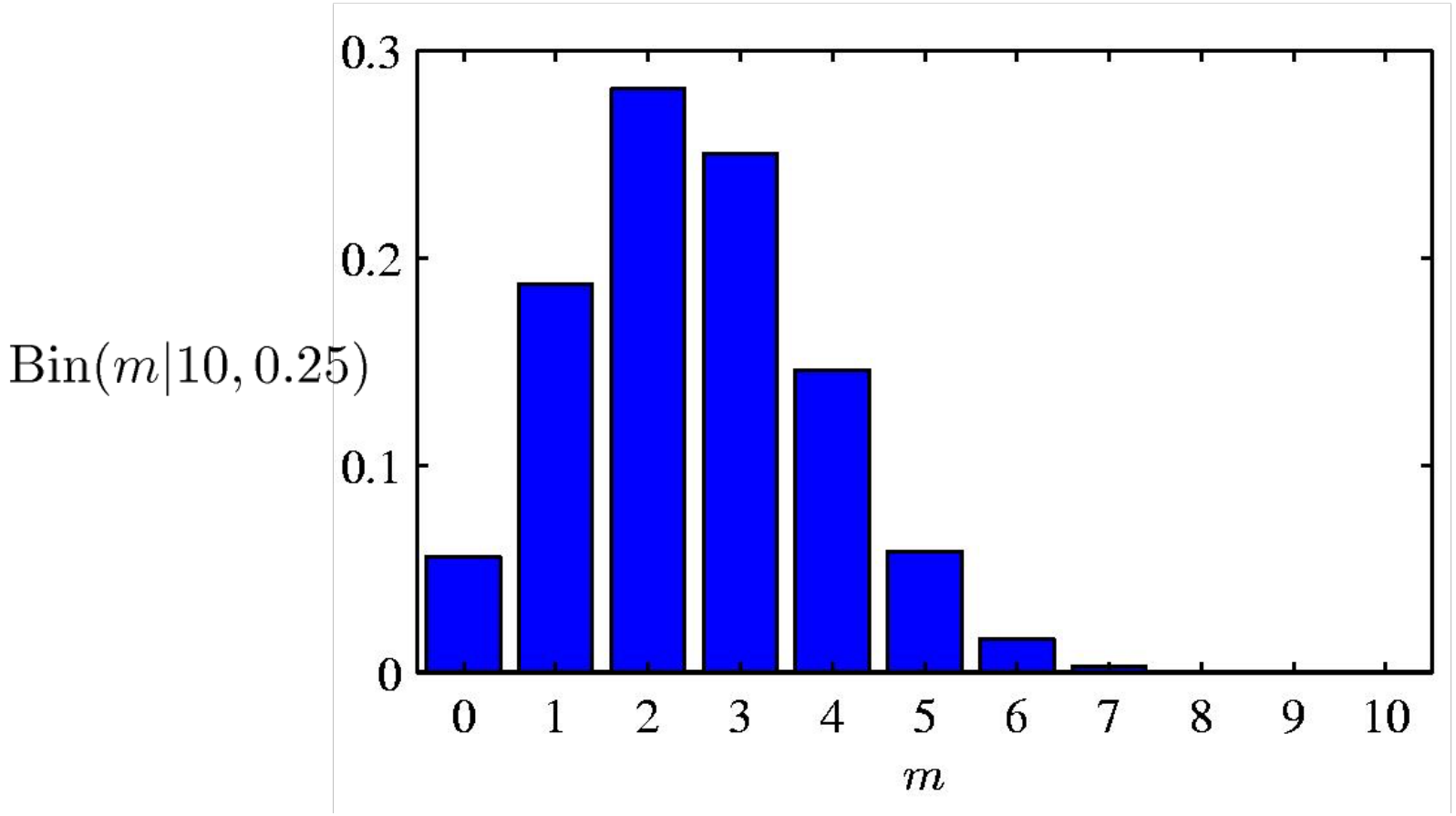
$$\binom{N}{m} = \frac{N!}{(N-m)!m!}. \tag{27}$$

- Its mean and variance are

$$\mathbb{E}[m] = \sum_{m=0}^{N} m\text{Bin}(m|N,\mu) = \mu, \tag{28}$$

$$\text{var}[x] = \mathbb{E}[(m-N\mu)^2] = N\mu(1-\mu). \tag{29}$$

Binomial distribution



Beta distribution

- We assume that the probability distribution over the parameter $\mu \in [0, 1]$ can be formulated as follows:

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, \tag{30}$$

  which is called distribution, with $\Gamma(\cdot)$ being the gamma function

$$\Gamma(a) = \int_0^\infty \mu^{a-1}e^{-\mu}d\mu. \tag{31}$$

Beta distribution

- We assume that the probability distribution over the parameter $\mu \in [0, 1]$ can be formulated as follows:

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, \tag{30}$$

  which is called distribution, with $\Gamma(\cdot)$ being the gamma function

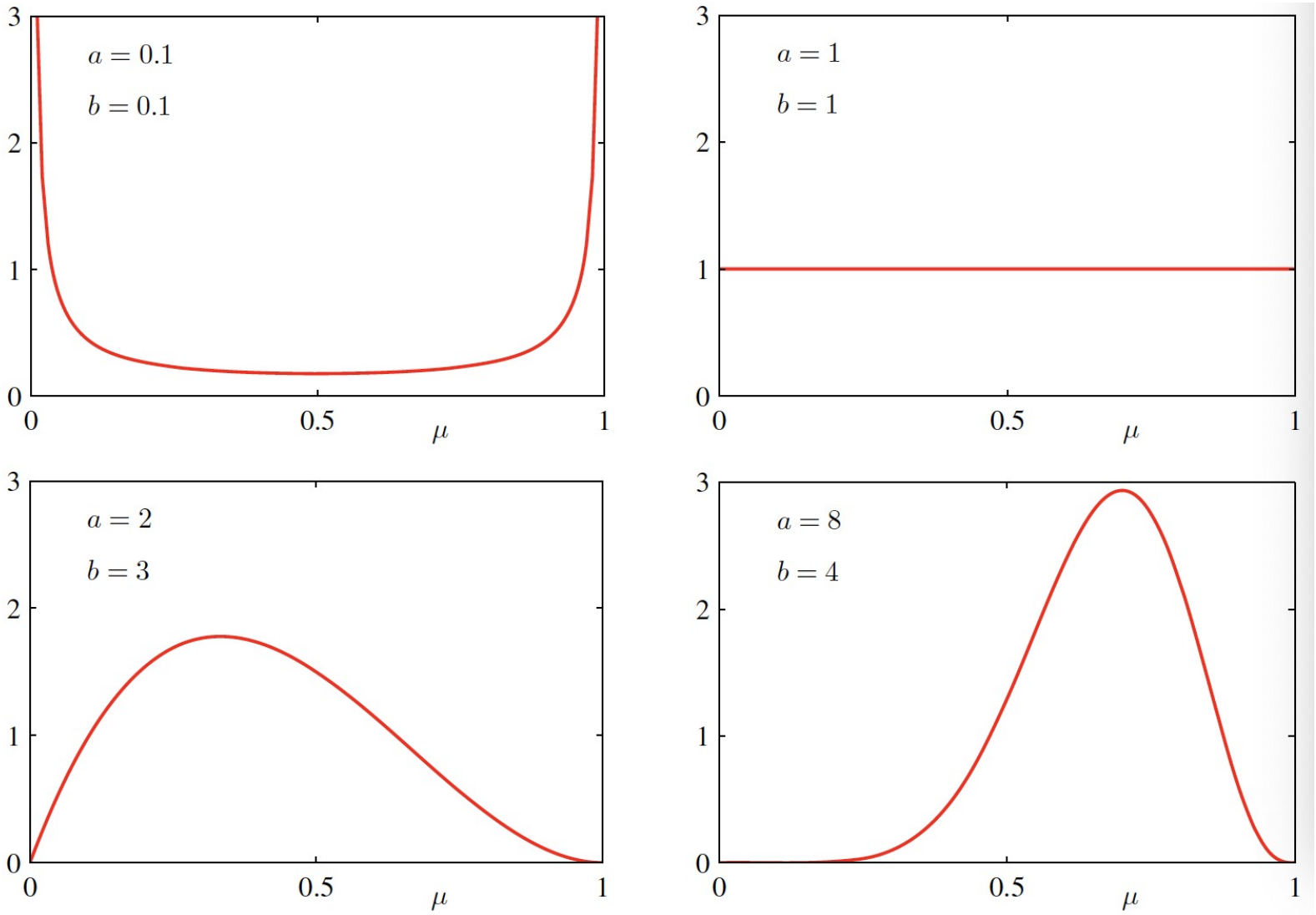
$$\Gamma(a) = \int_0^\infty \mu^{a-1}e^{-\mu}d\mu. \tag{31}$$

- Its mean and variance are

$$\mathbb{E}[\mu] = \frac{a}{a+b}, \tag{32}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}. \tag{33}$$

Beta distribution



Gaussian distribution

- The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{34}$$

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

Gaussian distribution

- The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the Gaussian distribution can be written in the form

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{34}$$

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

- For a D-dimensional vector $\boldsymbol{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \tag{35}$$

  where $\boldsymbol{\mu}$ is a D-dimensional mean vector, and $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

Student t distribution

- One problem with the Gaussian distribution is that it is sensitive to outliers, since the log probability only decays quadratically with distance from the center. A more robust distribution is the Student t distribution. Its pdf is as follows:

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{(2)}}, \tag{36}$$

where μ is the mean, σ^2 is the scale parameter, and $\nu > 0$ is called the **degrees of freedom**.

Student t distribution

- One problem with the Gaussian distribution is that it is sensitive to outliers, since the log probability only decays quadratically with distance from the center. A more robust distribution is the Student t distribution. Its pdf is as follows:

$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{(2)}}, \tag{36}$$

  where μ is the mean, σ^2 is the scale parameter, and $\nu > 0$ is called the **degrees of freedom**.
- It has the following properties:

$$\text{mean} = \mu, \text{mode} = \mu, \text{var} = \frac{\nu\sigma^2}{\nu-2}. \tag{37}$$

Laplace distribution

- Another distribution with heavy tails is the Laplace distribution. This has the following pdf:

$$\text{Lap}(x|\mu, b) = \frac{1}{2b} \exp\left(- \frac{|x - \mu|}{b} \right), \tag{38}$$

where μ is a location parameter and $b > 0$ is a scale parameter.

Laplace distribution

- Another distribution with heavy tails is the Laplace distribution. This has the following pdf:

$$\text{Lap}(x|\mu, b) = \frac{1}{2b} \exp\left(- \frac{|x - \mu|}{b} \right), \tag{38}$$

  where μ is a location parameter and $b > 0$ is a scale parameter.
- This distribution has the following properties:

$$\text{mean} = \mu, \text{mode} = \mu, \text{var} = 2b^2. \tag{39}$$

Laplace distribution

- Another distribution with heavy tails is the Laplace distribution. This has the following pdf:

$$\text{Lap}(x|\mu, b) = \frac{1}{2b} \exp\left(- \frac{|x - \mu|}{b} \right), \tag{38}$$

  where μ is a location parameter and $b > 0$ is a scale parameter.
- This distribution has the following properties:

$$\text{mean} = \mu, \text{mode} = \mu, \text{var} = 2b^2. \tag{39}$$

- It puts more probability density at 0 than the Gaussian. This property is a useful way to encourage sparsity in a model

Plots of above three distribution

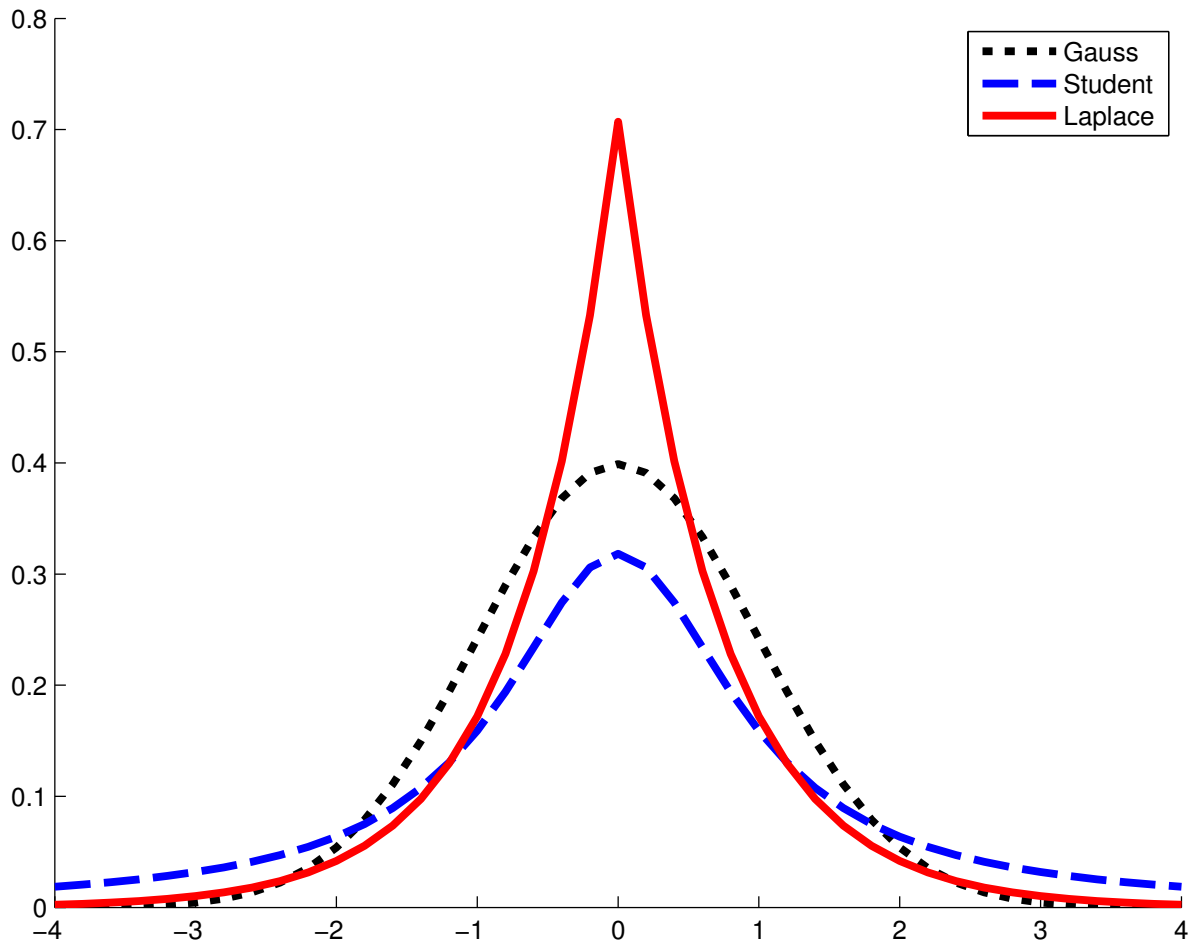


Figure: The pdf's for a $\mathcal{N}(0,1)$, $\mathcal{T}(0,1,1)$ and $\text{Lap}(0,\frac{\sqrt{2}}{2})$. The mean is 0 and the variance is 1 for both the Gaussian and Laplace. The mean and variance of the Student is undefined when $\nu = 1$.

What is information

- It is a measure that quantifies the *uncertainty* of an event with given probability - Shannon 1948.
- For a discrete source with finite alphabet $\mathcal{X} = \{x_0, x_1, \ldots, x_{M-1}\}$ where the probability of each symbol is given by $P(X = x_k) = p_k$

$$I(x_k) = \log \frac{1}{p_k} = -\log(p_k)$$

- If logarithm is base 2, information is given in bits.

What is information

- It represents the *surprise* of seeing the outcome (a highly probable outcome is not surprising).

event	probability	surprise
one equals one	1	0 bits
wrong guess on a 4-choice question	3/4	0.415 bits
correct guess on true-false question	1/2	1 bit
correct guess on a 4-choice question	1/4	2 bits
seven on a pair of dice	6/36	2.58 bits
win any prize at Euromilhões	1/24	4.585 bits
win Euromilhões Jackpot	$\approx 1/76$ million	≈ 26 bits
gamma ray burst mass extinction today	$< 2.7 \cdot 10^{-12}$	> 38 bits

Entropy

- Expected value of information from a source.

$$H(X) = E[I(x_k)] = \sum_{x \in \mathcal{X}} p_x(x) I(x_k)$$

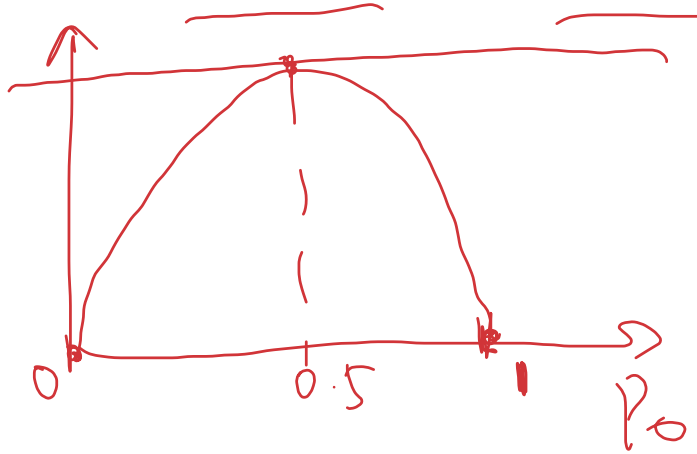
$$= -\sum_{x \in \mathcal{X}} p_x(x) \log p_x(x)$$

$$-\sum_{x} q(x) \log P(x)$$

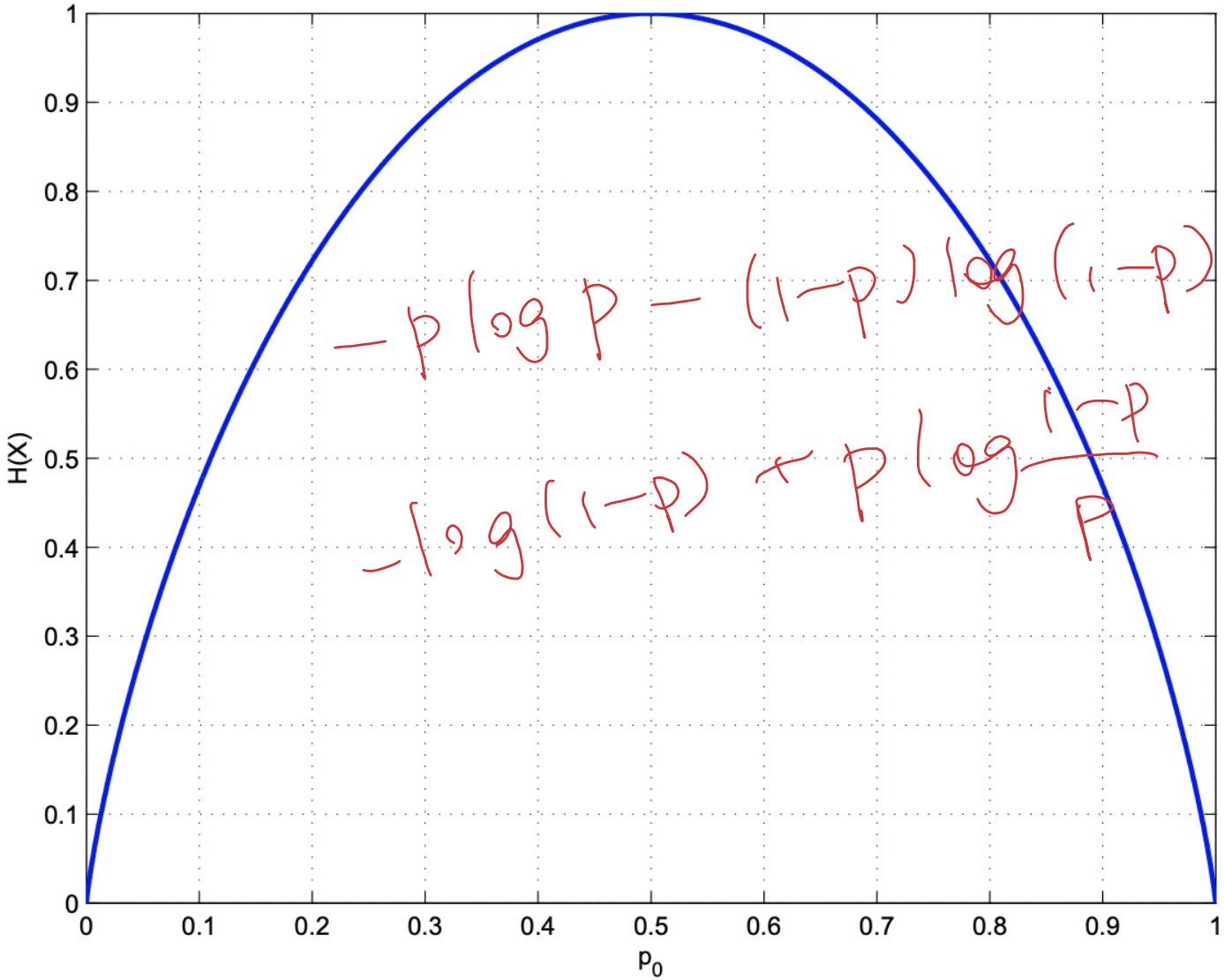
Entropy of binary source

- Let X be a binary source with p_0 and p_1 being the probabililty of symbols x_0 and x_1 respectively.

$$\begin{aligned} H(X) &= -p_0 \log p_0 - p_1 \log p_1 \\ &= -p_0 \log p_0 - (1-p_0)\log(1-p_0) \end{aligned}$$



Entropy of binary source



Joint entropy

- The joint entropy of a pair of random variables X and Y is given by:

$$H(X,Y) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{XY}(x,y)\log p_{X,Y}(x)$$

Conditional entropy

- Average amount of information of a random variable given the occurence of other.

$$
\begin{aligned}
H(X|Y) &= \sum_{y\in\mathcal{Y}} p_Y(y) H(X|Y=y) \\
&= -\sum_{y\in\mathcal{Y}} p_Y(y) \sum_{x\in\mathcal{X}} p_{X|Y=y}(x) \log p_{x|Y=y}(x) \\
&= -\sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p_{XY}(x,y) \log p_{X|Y=y}(x)
\end{aligned}
$$

$p(x|y)$

Conditional and joint entropy

- The entropy of a pair of random variables is equal to the entropy of one of them plus the conditional entropy.

$$H(X, Y) = H(X) + H(Y|X)$$

- Corollary

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$$

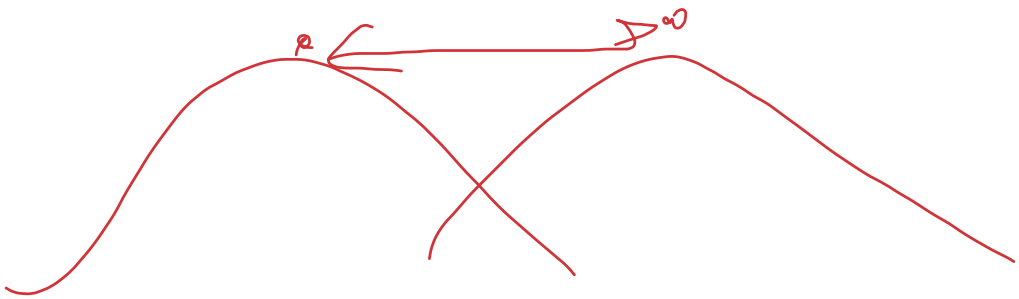
$$-\sum_{x,y} p(x,y) \log p(x,y) \quad -\sum_{x} p(x) \log p(x) \quad -\sum_{x,y} p(y|x) \log p(y|x)$$

Chain rule of joint entropy

The joint entropy also follows the chain rule like the joint probability:

$$H(X_1, X_2, \ldots, X_M) = \sum_{j=1}^{M} H(X_j | X_1, \ldots, X_{j-1})$$

Relative entropy: Kullback-Leibler divergence



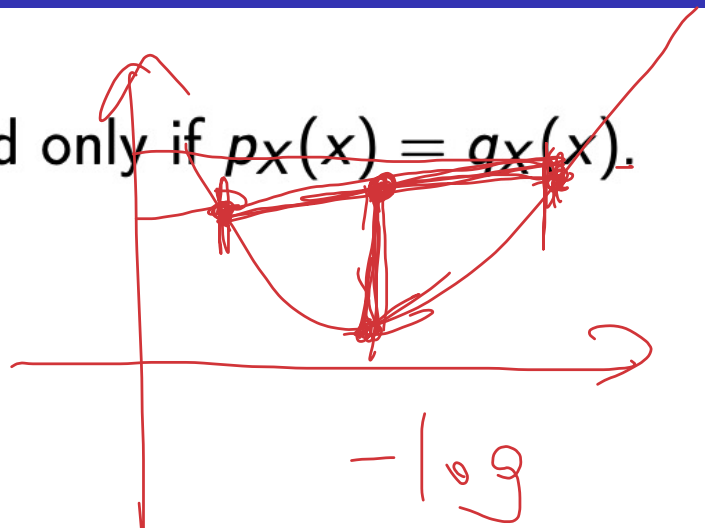
- Is a measure of the distance between two distributions.
- The relative entropy between two probability density functions $p_X(x)$ and $q_X(x)$ is defined as:

$$D(p_X(x)||q_X(x)) = \sum_{x \in \mathcal{X}} p_X(x) \log \frac{p_X(x)}{q_X(x)}$$

$$D(p \| q) = \sum_x p \log \frac{p}{q}$$

Properties of KL divergence

- $D(p_X(x)||q_X(x)) \geq 0$ with equality if and only if $p_X(x) = q_X(x)$.
- $D(p_X(x)||q_X(x)) \neq D(q_X(x)||p_X(x))$



$\ln a \leq a - 1$

$-D(p||q) = -\sum p \log\left(\frac{q}{p}\right)$

$\leq -\sum p \left(\frac{q}{p} - 1\right)$

$= -\sum q + \sum p$

$-1 \quad 1$

$= 0$

$f(\alpha x_1 + (1-\alpha)x_2)$

$\leq \alpha f(x_1) + (1-\alpha) f(x_2)$

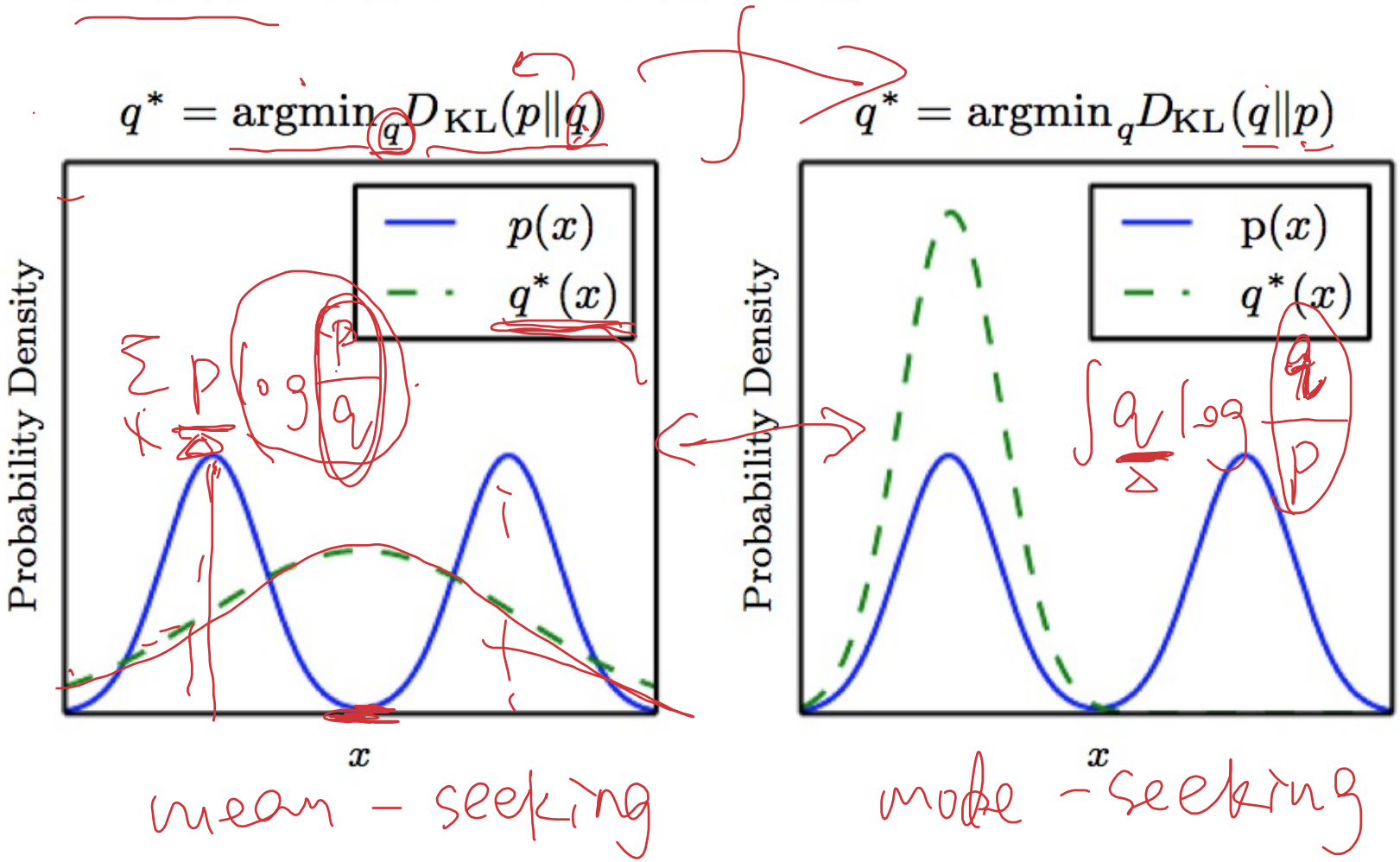
$-D(p||q) = \sum p \left(-\log \frac{q}{p}\right)$

$\leq -\log \sum p \frac{q}{p}$

$= 0$

Properties of KL divergence

- $D(p_X(x)||q_X(x)) \geq 0$ with equality if and only if $p_X(x) = q_X(x)$.
- $D(p_X(x)||q_X(x)) \neq D(q_X(x)||p_X(x))$



Mutual information

- The mutual information of two random variables X and Y is defined as the relative entropy between the joint probability density $p_{XY}(x, y)$ and the product of the marginals $p_X(x)$ and $p_Y(y)$

$$\begin{aligned} I(X;Y) &= D(p_{XY}(x,y)||p_X(x)p_Y(y)) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{XY}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \end{aligned}$$

Conditional mutual information

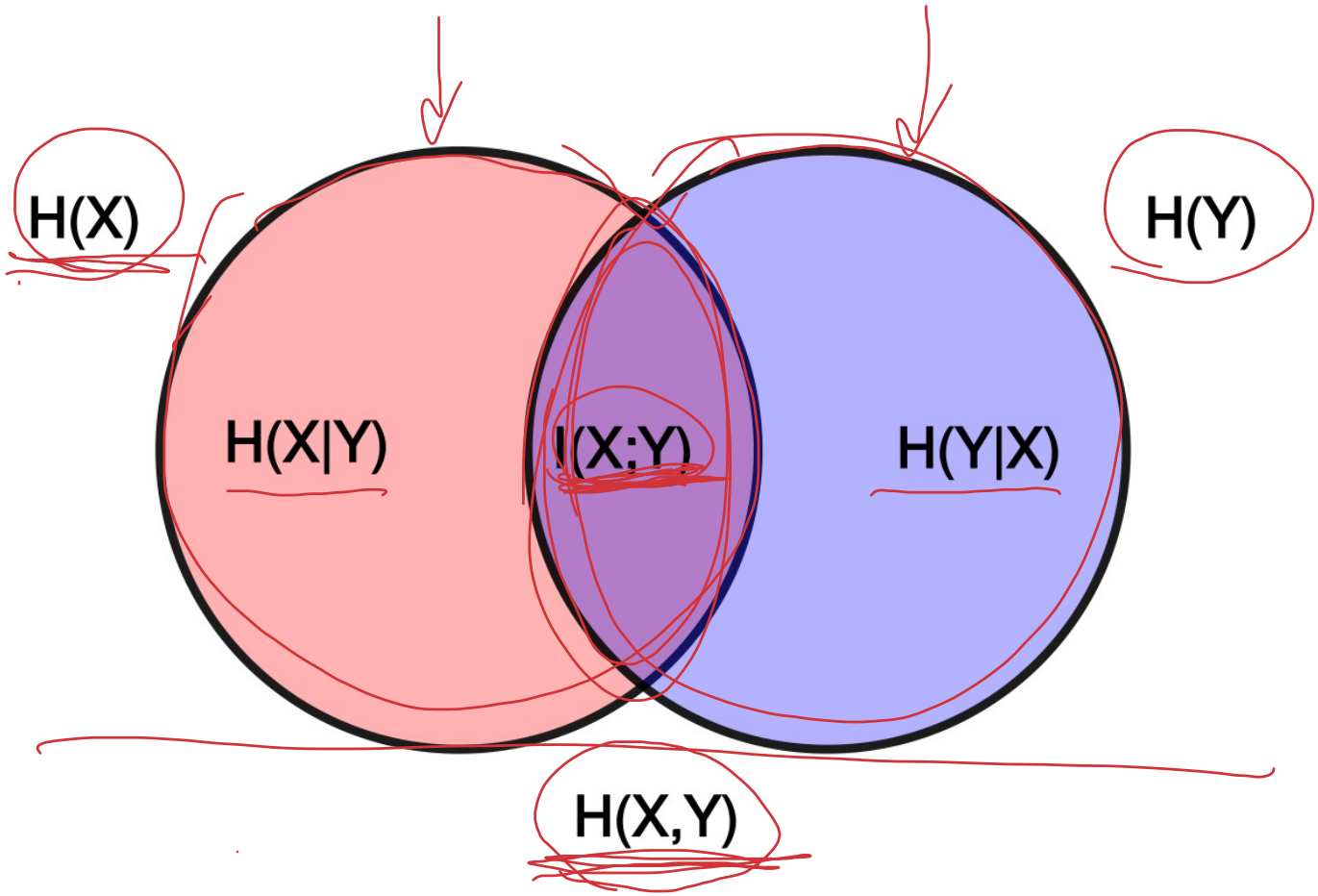
- Conditional Mutual Information:

$$I(X; Y|Z) = H(X|Z) - H(X|Y,Z)$$

- Chain Rule for Mutual Information

$$I(X_1, X_2, \ldots, X_M; Y) = \sum_{j=1}^{M} I(X_j; Y|X_1, \ldots, X_{j-1})$$

Mutual information and entropy



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