Lecture 1: Probability and Distributions

Wei Wang
HKUST(GZ)

Adapted from Prof. Yung Yi (KAIST)

Reference:
Mathematics for Machine Learning
Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong
Cambridge 2020

March 11, 2025

Roadmap

(1) Construction of a Probability Space

(2) Discrete and Continuous Probabilities

(3) Sum Rule, Product Rule, and Bayes' Theorem

(4) Change of Variables/Inverse Transform

(5) Entropy and KL Divergence

Probabilistic Model

Motivation: "probable" has many different meanings in NL. Can we have a rigorous treatment (in the spirit of David Hilbert's sixth problem)?

- The standard probability axioms are the foundations of probability theory introduced by Russian mathematician Andrey Kolmogorov in 1933.
- https://en.wikipedia.org/wiki/Probability_axioms [1]

Elements of Probabilistic Model

1. All outcomes of my interest: Sample Space Ω
2. Assigned numbers to each outcome of Ω: Probability Law $\mathbb{P}(\cdot)$

Question: What are the conditions of Ω and $\mathbb{P}(\cdot)$ under which their induced probability model becomes "legitimate"?

[1] See https://www.scottaaronson.com/democritus/lec9.html for extra enlightment.

Sample Space Ω

The set of all outcomes of **my interest**

1. Mutually exclusive
2. Collectively exhaustive
3. At the **right granularity** (not too concrete, not too abstract)

1. Toss a coin. What about this? $\Omega = \{H, T, HT\}$
2. Toss a coin. What about this? $\Omega = \{H\}$
3. (a) Just figuring out prob. of H or T. $\Longrightarrow \Omega = \{H, T\}$

   (b) The impact of the weather (rain or no rain) on the coin's behavior.

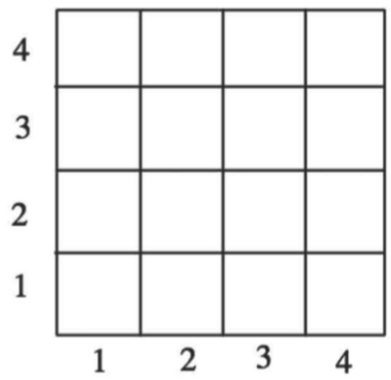
   $$\Longrightarrow \Omega = \{(H, R), (T, R), (H, NR), (T, NR)\},$$

   where R(Rain), NR(No Rain).

Examples: Sample Space Ω

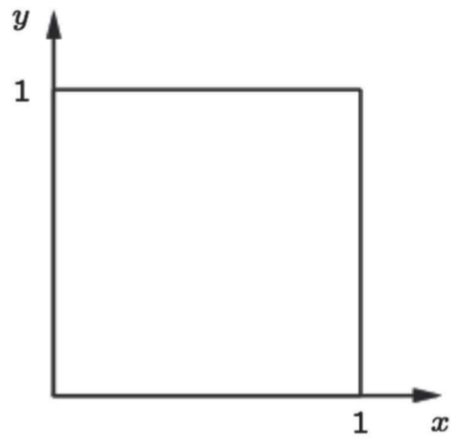
- *Discrete case:* Two rolls of a tetrahedral die

  - $\Omega = \{(1,1),(1,2),\ldots,(4,4)\}$



- *Continuous case:* Dropping a needle in a plain

  - $\Omega = \{(x,y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq 1\}$



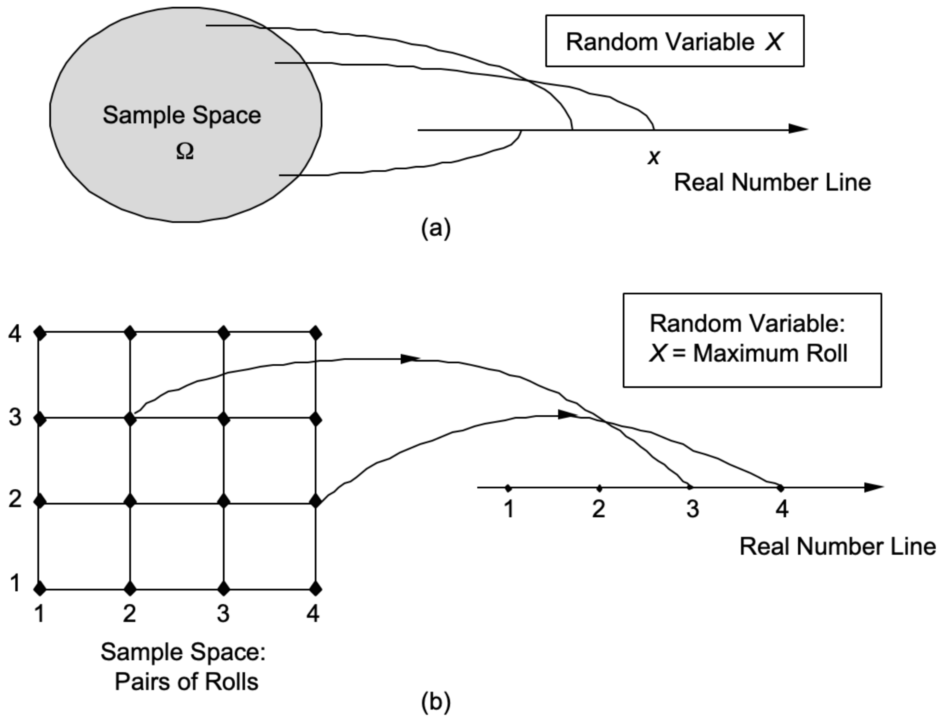
Probability Law

- Assign numbers to what? Each outcome?
- What is the probability of dropping a needle at $(0.5, 0.5)$ over the 1×1 plane?
- Assign numbers to each subset of Ω: A subset of Ω: an event
- $\mathbb{P}(A)$: Probability of an event A.
  - This is where probability meets set theory.
  - Roll a dice. What is the probability of odd numbers?

    $\mathbb{P}(\{1, 3, 5\})$, where $\{1, 3, 5\} \subset \Omega$ is an event.
- Event space $\mathcal{A}$: The collection of subsets of Ω. For example, in the discrete case, the power set of Ω.
- Probability Space $(\Omega, \mathcal{A}, \mathbb{P}(\cdot))$

Random Variable: Idea

- In reality, many outcomes are numerical, e.g., stock price.
- Even if not, very convenient if we map numerical values to random outcomes, e.g., '0' for male and '1' for female.



Random Variable: More Formally

- Mathematically, a random variable X is a function which maps from Ω to $\mathbb{R}$.
- Notation. Random variable X, numerical value x.
- Different random variables X, Y, etc can be defined on the same sample space.
- For a fixed value x, we can associate an event that a random variable X has the value x, i.e., $\{\omega \in \Omega \mid X(w) = x\}$
- Generally,

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S) = \mathbb{P}(X^{-1}(S)) = \mathbb{P}\Big(\{\omega \in \Omega : X(w) \in S\}\Big)$$

Conditioning: Motivating Example

- Pick a person a at random
  - event A: a's age ≤ 20
  - event B: a is married
- (Q1) What is the probability of A?
- (Q2) What is the probability of A, given that B is true?
- Clearly the above two should be different.
- Question. How should I change my belief, given some additional information?
- Need to build up a new theory, which we call conditional probability.

Conditional Probability

- $\mathbb{P}(A \mid B)$: $\mathbb{P}(\cdot|B)$ should be a new probability law.
- Definition.

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \textit{for} \quad \mathbb{P}(B) > 0.$$

  - Note that this is a definition, not a theorem.
- All other properties of the law $\mathbb{P}(\cdot)$ is applied to the conditional law $\mathbb{P}(\cdot|B)$.
- For example, for two disjoint events A and C,

$$\mathbb{P}(A \cup C \mid B) = \mathbb{P}(A \mid B) + \mathbb{P}(C \mid B)$$

Roadmap

(1) Construction of a Probability Space

(2) Discrete and Continuous Probabilities

(3) Sum Rule, Product Rule, and Bayes' Theorem

(4) Change of Variables/Inverse Transform

(5) Entropy and KL Divergence

Discrete Random Variables

- The values that a random variable X takes is discrete (i.e., finite or countably infinite).
- Then, $p_X(x) := \mathbb{P}(X = x) := \mathbb{P}\Big(\{\omega \in \Omega \mid X(w) = x\}\Big)$, which we call probability mass function (PMF).
- Examples: Bernoulli, Uniform, Binomial, Poisson, Geometric

Bernoulli X with parameter $p \in [0, 1]$

- Only binary values

$$X = \begin{cases} 0, & \text{w.p.}^2 \quad 1-p, \\ 1, & \text{w.p.} \quad p \end{cases}$$

  In other words, $p_X(0) = 1 - p$ and $p_X(1) = p$ from our PMF notation.

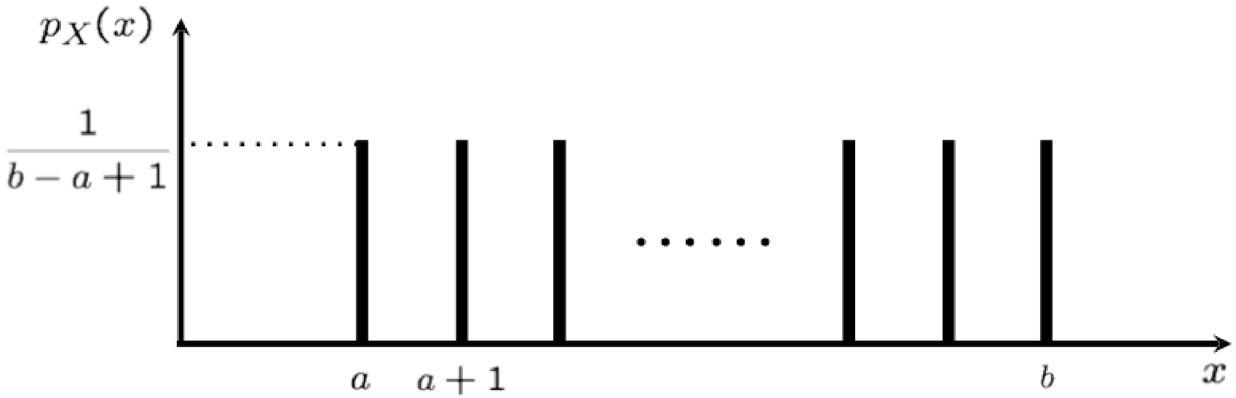
- Models a trial that results in binary results, e.g., success/failure, head/tail
- Very useful for an indicator rv of an event A. Define a rv $\mathbf{1}_A$ as:

$$\mathbf{1}_A = \begin{cases} 1, & \text{if } A \text{ occurs}, \\ 0, & \text{otherwise} \end{cases}$$

[2] with probability

Uniform X with parameter a, b

- integers a, b, where $a \leq b$
- Choose a number of $\Omega = \{a, a+1, \ldots, b\}$ uniformly at random.
- $p_X(i) = \frac{1}{b-a+1}$, $i \in \Omega$.

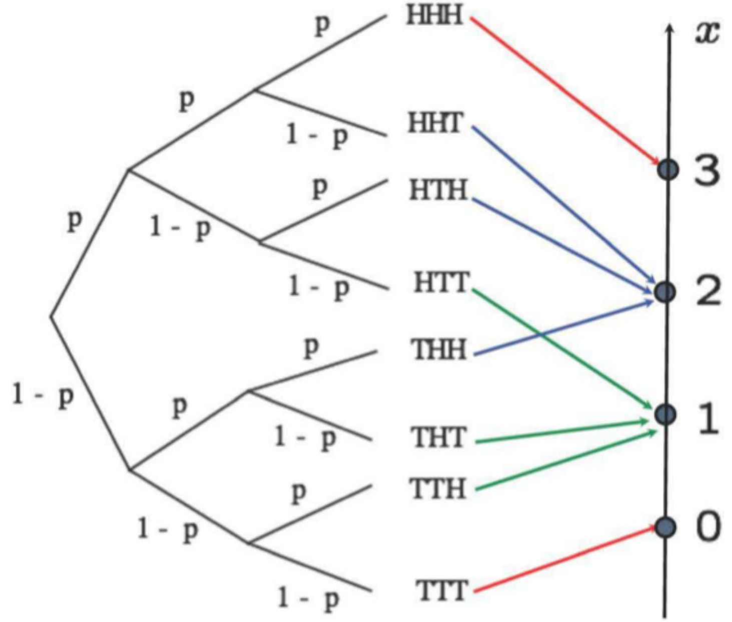


- Models complete ignorance (I don't know anything about X)

Binomial X with parameter n, p

- Models the number of successes in a given number of independent trials
- n independent trials, where one trial has the success probability p.

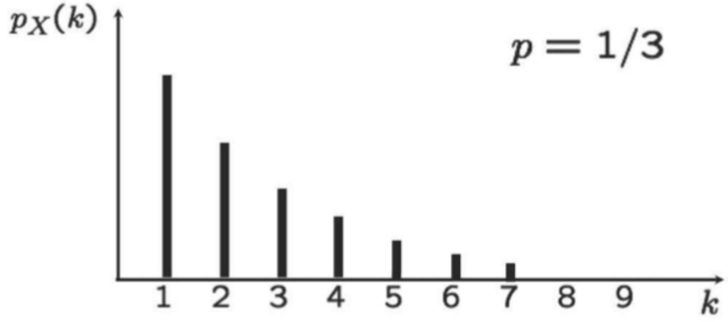
$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Geometric X with parameter p

- Experiment: infinitely many independent Bernoulli trials, where each trial has success probability p
- Random variable: number of trials until the first success.
- Models waiting times until something happens.

$$p_X(k) = (1-p)^{k-1}p$$



Joint PMF

- Joint PMF. For two random variables X, Y, consider two events $\{X = x\}$ and $\{Y = y\}$, and

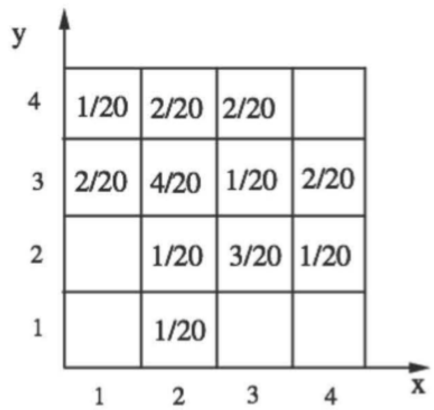
$$p_{X,Y}(x,y) \;:=\; \mathbb{P}\Big(\{X = x\} \cap \{Y = y\}\Big)$$

- $\sum_x \sum_y p_{X,Y}(x,y) = 1$
- Marginal PMF.

$$p_X(x) = \sum_y p_{X,Y}(x,y),$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Example.



$p_{X,Y}(1,3) = 2/20$

$p_X(4) = 2/20 + 1/20 = 3/20$

$\mathbb{P}(X = Y) = 1/20 + 4/20 + 3/20 = 8/20$

Conditional PMF

- Conditional PMF

$$p_{X|Y}(x|y) \; := \; \mathbb{P}(X = x | Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

  for y such that $p_Y(y) > 0$.

- $\sum_x p_{X|Y}(x|y) = 1$
- Multiplication rule.

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y) p_{X|Y}(x|y) \\ &= p_X(x) p_{Y|X}(y|x) \end{aligned}$$

- $p_{X,Y,Z}(x,y,z) = p_X(x) p_{Y|X}(y|x) p_{Z|X,Y}(z|x,y)$

y \ x	1	2	3	4
4	1/20	2/20	2/20	
3	2/20	4/20	1/20	2/20
2		1/20	3/20	1/20
1		1/20		

$p_{X|Y}(2|2) = \frac{1}{1+3+1}$

$p_{X|Y}(3|2) = \frac{3}{1+3+1}$

$\mathbb{E}[X|Y=3] = 1(2/9)+2(4/9)+3(1/9)+4(2/9)$

Continuous RV and Probability Density Function (PDF)

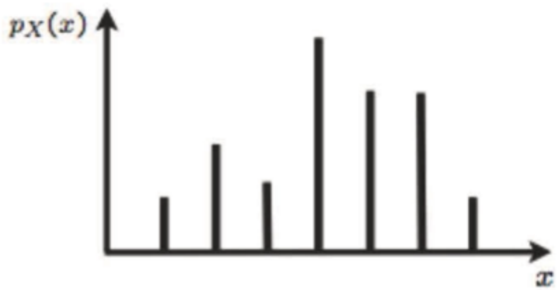
- How to handle random variables that have continuous values, e.g., velocity of a car?

Continuous Random Variable

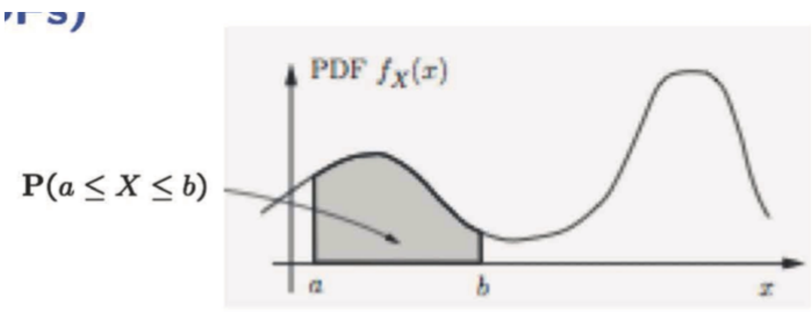
A rv X is continuous if $\exists$ a function f_X, called probability density function (PDF), s.t.

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx$$

- All of the concepts and methods (expectation, PMFs, and conditioning) for discrete rvs have continuous counterparts

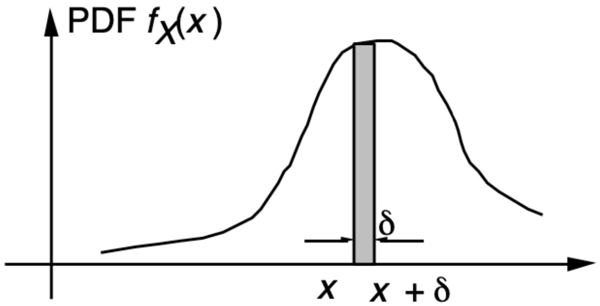


- $\mathbb{P}(a \le X \le b) = \sum_{x: a \le x \le b} p_X(x)$
- $p_X(x) \ge 0$, $\sum_x p_X(x) = 1$



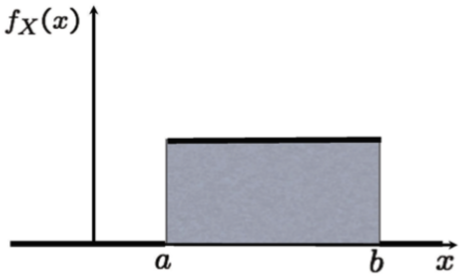
- $\mathbb{P}(a \le X \le b) = \int_a^b f_X(x)dx$
- $f_X(x) \ge 0$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$

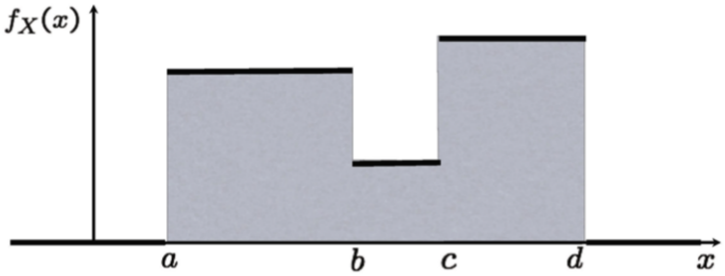
PDF and Examples



- $\mathbb{P}(a \leq X \leq a + \delta) \approx \boxed{f_X(a) \cdot \delta}$
- $\mathbb{P}(X = a) = 0$

Examples



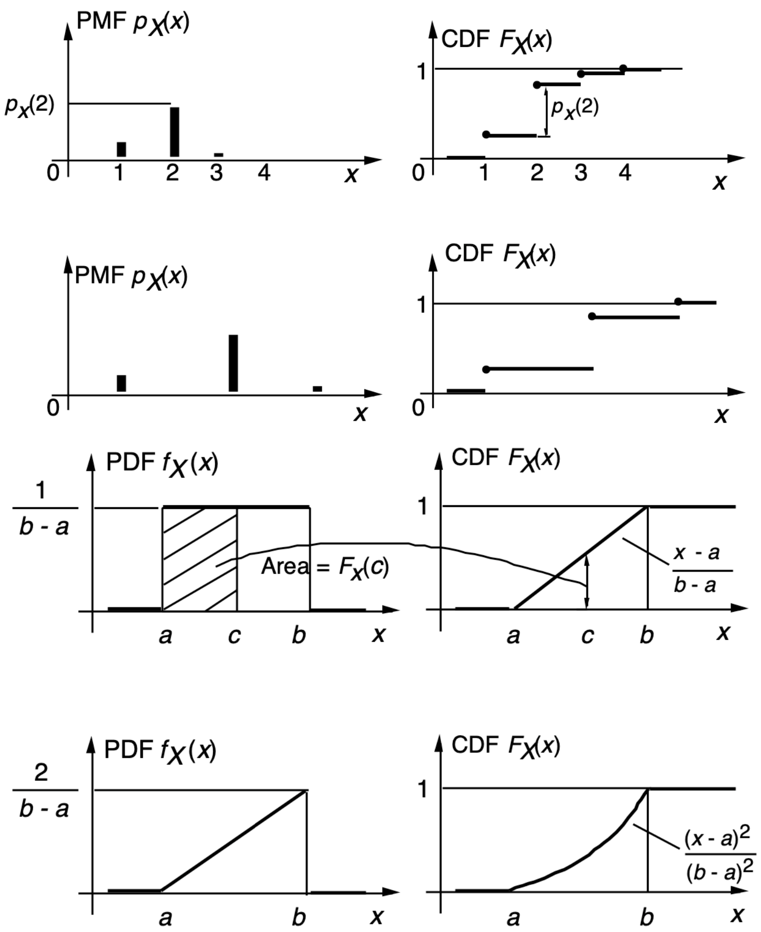


Cumulative Distribution Function (CDF)

- Discrete: PMF, Continuous: PDF
- Can we describe all types of rvs with a single mathematical concept?

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{discrete} \\ \int_{-\infty}^{x} f_X(t)dt, & \text{continuous} \end{cases}$$

- always well defined, because we can always compute the probability for the event $\{X \leq x\}$
- CCDF (Complementary CDF): $\mathbb{P}(X > x)$



CDF Properties

- Non-decreasing
- $F_X(x)$ tends to 1, as $x \to \infty$
- $F_X(x)$ tends to 0, as $x \to -\infty$

Continuous: Joint PDF and CDF (1)

Jointly Continuous

Two continuous rvs are **jointly continuous** if a non-negative function $f_{X,Y}(x, y)$ (called joint PDF) satisfies: for **every** subset B of the two dimensional plane,

$$\mathbb{P}((X, Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x, y)dxdy$$

1. The joint PDF is used to calculate probabilities

$$\mathbb{P}((X, Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x, y)dxdy$$

Our particular interest: $B = \{(x, y) \mid a \leq x \leq b, c \leq y \leq d\}$

Continuous: Joint PDF and CDF (2)

2. The marginal PDFs of X and Y are from the joint PDF as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

3. The joint CDF is defined by $F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$, and determines the joint PDF as:

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{x,y}}{\partial x \partial y}(x,y)$$

4. A function $g(X, Y)$ of X and Y defines a new random variable, and

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy$$

Continuous: Conditional PDF given a RV

- (discrete) $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

- (continuous) for $f_Y(y) > 0$,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Remember: For a fixed event A, $\mathbb{P}(\cdot|A)$ is a legitimate probability law.

- Similarly, For a fixed y, $f_{X|Y}(x|y)$ is a legitimate PDF, since

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}{f_Y(y)} = 1$$

Sum Rule and Product Rule

- Sum Rule

$$p_X(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) & \text{if discrete} \\ \int_{y \in \mathcal{Y}} f_{X,Y}(x,y)dy & \text{if continuous} \end{cases}$$

  - Generally, for $X = (X_1, X_2, \ldots, X_D)$,

$$p_{X_i}(x_i) = \int p_X(x_1, \ldots, x_i, \ldots, x_D) d\boldsymbol{x}_{-i}$$

  - Computationally challenging, because of high-dimensional sums or integrals

- Product Rule

$$p_{X,Y}(x,y) = p_X(x) \cdot p_{Y|X}(y|x)$$

joint dist. = marginal of the first × conditional dist. of the second given the first

  - Same as $p_Y(y) \cdot p_{X|Y}(x|y)$

Bayes Rule

- X: state/cause/original value $\to$ Y: result/resulting action/noisy measurement
- Model: $\mathbb{P}(X)$ (prior) and $\mathbb{P}(Y|X)$ (cause $\to$ result)
- Inference: $\mathbb{P}(X|Y)$?

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)p_{Y|X}(y|x) \\ &= p_Y(y)p_{X|Y}(x|y) \\ p_{X|Y}(x|y) &= \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)} \\ p_Y(y) &= \sum_{x'} p_X(x')p_{Y|X}(y|x') \end{aligned}$$

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)f_{Y|X}(y|x) \\ &= f_Y(y)f_{X|Y}(x|y) \\ f_{X|Y}(x|y) &= \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} \\ f_Y(y) &= \int f_X(x')f_{Y|X}(y|x')dx' \end{aligned}$$

$$\underbrace{p_{X|Y}(x|y)}_{\text{posterior}} = \frac{\overbrace{p_{Y|X}(y|x)}^{\textit{likelihood}}\overbrace{p_X(x)}^{\textit{prior}}}{\underbrace{p_Y(y)}_{\textit{evidence}}}$$

Bayes Rule for Mixed Case

K: discrete, Y: continuous

- Inference of K given Y

$$p_{K|Y}(k|y) = \frac{p_K(k) f_{Y|K}(y|k)}{f_Y(y)}$$

$$f_Y(y) = \sum_{k'} p_K(k') f_{Y|K}(y|k')$$

- Inference of Y given K

$$f_{Y|K}(y|k) = \frac{f_Y(y) p_{K|Y}(k|y)}{p_K(k)}$$

$$p_K(k) = \int f_Y(y') p_{K|Y}(k|y') dy'$$

Roadmap

(1) Construction of a Probability Space

(2) Discrete and Continuous Probabilities

(3) Sum Rule, Product Rule, and Bayes' Theorem

(4) Change of Variables/Inverse Transform

(5) Entropy and KL Divergence

Normal (also called Gaussian) Random Variable

- Why important?
  - Central limit theorem (CLT): One of the most remarkable findings in the probability theory
  - Convenient analytical properties
  - Modeling aggregate noise with many small, independent noise terms

- Standard Normal $\mathcal{N}(0, 1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

- $\mathbb{E}[X] = 0$
- $\text{var}[X] = 1$

- General Normal $\mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $\mathbb{E}[X] = \mu$
- $\text{var}[X] = \sigma^2$

Power of Gaussian Random Vectors

- Marginals of Gaussians are Gaussians
- Conditionals of Gaussians are Gaussians
- Products of Gausssian Densities are Gaussians.
- A sum of two Gassuaians is Gaussian if they are independent
- Any linear/affine transformation of a Gaussian is Gaussian.

Marginals and Conditionals of Gaussians

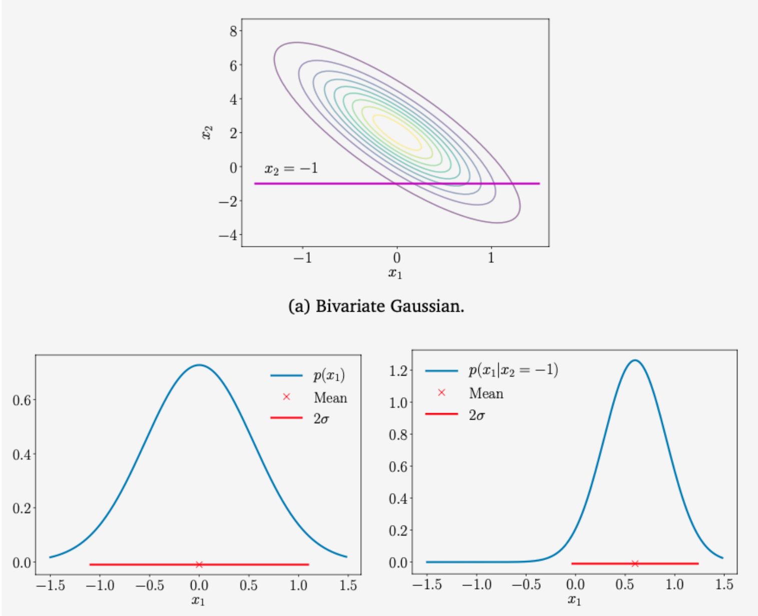
- $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Gaussians with mean vectors $\boldsymbol{\mu_X}$ and $\boldsymbol{\mu_Y}$, respectively.
- Gaussian random vector $\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}$ with $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{pmatrix}$ and the covarance matrix $\boldsymbol{\Sigma_Z} = \begin{pmatrix} \boldsymbol{\Sigma_X} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_Y} \end{pmatrix}$, where $\boldsymbol{\Sigma_{XY}} = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y})$.

\- Marginal.

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu_x}, \boldsymbol{\Sigma_X})$$

\- Conditional. $\boldsymbol{X} \mid \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu_{X|Y}}, \boldsymbol{\Sigma_{X|Y}})$,

$$\boldsymbol{\mu_{X|Y}} = \boldsymbol{\mu_X} + \boldsymbol{\Sigma_{XY}}\boldsymbol{\Sigma_Y}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu_Y})$$
$$\boldsymbol{\Sigma_{X|Y}} = \boldsymbol{\Sigma_X} - \boldsymbol{\Sigma_{XY}}\boldsymbol{\Sigma_Y}^{-1}\boldsymbol{\Sigma_{YX}}$$



(a) Bivariate Gaussian.
(b) Marginal distribution.
(c) Conditional distribution.

Product of Two Gaussian Densities

Note: this is not the density of the product of two Gaussian RVs (which does not have a closed-form expression).

- Lemma. Up to recaling, the pdf of the form $\exp(-\frac{1}{2}ax^2 - 2bx + c)$ is $\mathcal{N}(\frac{b}{a}, \frac{1}{a})$.
- Using the above Lemma, the product of two Gaussians $\mathcal{N}(\mu_0, \nu_0)$ and $\mathcal{N}(\mu_1, \nu_1)$ is Gaussian up to rescaling.

  Proof.

$$\exp\left(-(x-\mu_0)^2/2\nu_0\right) \times \exp\left(-(x-\mu_1)^2/2\nu_1\right)$$
$$= \exp\left[-\frac{1}{2}\left(\Big(\frac{1}{\nu_0}+\frac{1}{\nu_1}\Big)x^2 - 2\Big(\frac{\mu_0}{\nu_0}+\frac{\mu_1}{\nu_1}\Big)x + c\right)\right]$$
$$\implies \mathcal{N}\left(\overbrace{\frac{1}{\nu_0{}^{-1}+\nu_1{}^{-1}}}^{=\nu}, \nu\left(\frac{\mu_0}{\nu_0}+\frac{\mu_1}{\nu_1}\right)\right) = \mathcal{N}\left(\frac{\nu_1\mu_0+\nu_0\mu_1}{\nu_0+\nu_1}, \frac{\nu_0\nu_1}{\nu_0+\nu_1}\right)$$

Sum of Gaussians

Note: this is the vector form, and hence the scalar form holds trivially.

- $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu_X}, \boldsymbol{\Sigma_X})$ and $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu_Y}, \boldsymbol{\Sigma_Y})$

$$\implies a\boldsymbol{X} + b\boldsymbol{Y} \sim \mathcal{N}(a\boldsymbol{\mu_X} + b\boldsymbol{\mu_Y}, a^2\boldsymbol{\Sigma_X} + b^2\boldsymbol{\Sigma_Y})$$

Mixture of Two Gaussian Densities

- $f_1(x)$ is the density of $\mathcal{N}(\mu_1, \sigma_1^2)$ and $f_2(x)$ is the density of $\mathcal{N}(\mu_2, \sigma_2^2)$
- Question. What are the mean and the variance of the random variable Z which has the following density $f(x)$?

$$f(x) = \alpha f_1(x) + (1-\alpha) f_2(x)$$

Answer:

$$\mathbb{E}(Z) = \alpha\mu_1 + (1-\alpha)\mu_2$$

$$\text{var}(Z) = \left(\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right) + \left([\alpha\mu_1^2 + (1-\alpha)\mu_2^2] - [\alpha\mu_1 + (1-\alpha)\mu_2]^2\right)$$

Linear Transformation

- Linear transformation[3] preserves normality

Linear transformation of Normal

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then for $a \neq 0$ and b, $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

- Thus, every normal rv can be standardized:

  If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

- Thus, we can make the table which records the following CDF values:

  $$\Phi(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(Y < y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-t^2/2} dt$$

[3] Strictly speaking, this is affine transformation.

Roadmap

(1) Construction of a Probability Space

(2) Discrete and Continuous Probabilities

(3) Sum Rule, Product Rule, and Bayes' Theorem

(4) Change of Variables/Inverse Transform

(5) Entropy and KL Divergence

Knowing Distributions of Functions of RVs

- If $X \sim \mathcal{N}(0,1)$, what is the distribution of $Y = X^2$?
- If $X_1, X_2 \sim \mathcal{N}(0,1)$, what is the distribution of $Y = \frac{1}{2}(X_1 + X_2)$?
- Two techniques
  - CDF-based technique
  - Change-of-Variable technique
- In this lecture note, we focus on the case of univarate random variables for simplicity.

CDF-based Technique

S1. Find the CDF: $F_Y(y) = \mathbb{P}(Y \leq y)$

S2. Differentiate the CDF to get the pdf $f_Y(y)$: $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y)$

- Example. $f_X(x) = 3x^2,\ 0 \leq x \leq 1$. What is the pdf of $Y = X^2$?

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(X \leq \sqrt{y}) = F_X(\sqrt{y})$$

$$= \int_0^{\sqrt{y}} 3t^2 \mathrm{d}t = y^{\frac{3}{2}}, \quad 0 \leq y \leq 1$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{3}{2}\sqrt{y}, \quad 0 \leq y \leq 1$$

How to Get Random Samples of a Given Distribution? (1)

- Assume that $X \sim \exp(1)$, i.e., $f_X(x) = e^{-x}$ and $F_X(x) = 1 - e^{-x}$. How to make a programming code that gives random samples following the distribution X?
- Theorem. Probability Integral Theorem. Let X be a continuous rv with a strictly monotonic CDF $F(\cdot)$. Then, if we define a new rv U as $\boxed{U := F(X)}$, then U follows the uniform distribution over [0.1].
- Proof. Will show that $F_U(u) = u$, which is the CDF of a standard uniform rv.

  $$F_U(u) = \mathbb{P}(U \le u) = \mathbb{P}(F(X) \le u) \stackrel{(*)}{=} \mathbb{P}(X \le F^{-1}(u)) = F(F^{-1}(u)) = u,$$

  where $(*)$ is due to the strict monotonicity of $F(\cdot)$.

How to Get Random Samples of a Given Distribution? (2)

Pseudo Code of getting a random sample with the distribution $F(\cdot)$.

Step 1. Get a random sample u over $[0, 1]$ (most of software packages include this capability of generating a random number generation)

Step 2. Get a value $x = F^{-1}(u)$.

Change-of-Variables Technique: Univariate

- Chain rule of calculus: $\int f(g(x))g'(x)\mathrm{d}x = \int f(u)\mathrm{d}u$, where $u = g(x)$.
- Consider a rv $X \in [a, b]$ and an invertible, strictly increasing function U.

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(U(X) \leq y) = \mathbb{P}(X \leq U^{-1}(y)) = \int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(y)} f_X(x)\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}y}\int_a^{U^{-1}(y)} f_X(U^{-1}(y))U^{-1'}(y)\mathrm{d}y$$

$$= f_X(U^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y}U^{-1}(y)$$

- Including the case when U is strcitly decreasing,

$$f_Y(y) = f_X(U^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y}U^{-1}(y)\right|$$

Change-of-Variables Technique: Multivariate (Optional)

- Theorem. Let $f_{\boldsymbol{X}}(\boldsymbol{x})$ is the pdf of multivariate continuous random vector $\boldsymbol{X}$. If $\boldsymbol{Y} = U(\boldsymbol{X})$ is differentiable and invertible, the pdf of $\boldsymbol{Y}$ is given as:

$$f(\boldsymbol{y}) = f_{\boldsymbol{X}}(U^{-1}(\boldsymbol{y})) \cdot \left| \det \left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{y}} U^{-1}(\boldsymbol{y}) \right) \right|$$

- Example. For a bivariate rv $\boldsymbol{X}$ with its pdf $f(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = \frac{1}{2\pi} \exp \left(-\frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right)$,

  consider $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then, we have the following pdf of $\boldsymbol{Y}$:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{2\pi} \exp \left(-\frac{1}{2} \boldsymbol{y}^{\mathsf{T}} (\boldsymbol{A}^{-1})^{\mathsf{T}} \boldsymbol{A}^{-1} \boldsymbol{y} \right) |ad - bc|^{-1}$$

Sums of Independent RVs

- (Pictorial) Meaning of $Z = X + Y$
- Example: Roll 2 dices
- Use convolution: $(f * g)$

Find Z's PMF:

- $p_Z(z) = \sum_{y \in Y} p_X(z-y) p_Y(y)$

Find Z's PDF:

- $f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \mathrm{d}y$

Visit `https://en.wikipedia.org/wiki/List_of_convolutions_of_probability_distributions` for some known convolution results.

Statistics of Sums of Independent RVs

Nonetheless, finding the expectation and variance are easier

- Linearity of Expectation: $\mathbb{E}[x+y] = \mathbb{E}[x] + \mathbb{E}[y]$. Note: True even if they are not independent RVs.
- $\text{var}[X+Y] = \text{var}[X] + \text{var}[Y]$. Note: variance exhibits linearity only for independent RVs, as there is no covariance

Other common cases:

- $\mathbb{E}[aX+b] = a\mathbb{E}[X] + b$
- $\text{var}[aX+b] = a^2\text{var}[X]$
- $\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i]$

Further down the road

Law of Large Numbers

Let $X_1, X_2 \dots X_n$ be independent and identically distributed random variables. The average of these random variables (sample mean) converges to the expected value μ (population mean):

$$\sum_{i=1}^{n} X_i \to \mu$$

The Central Limit Thorem (Average Version)

Let $X_1, X_2 \dots X_n$ be independent and identically distributed random variables. The average of these random variables approaches a normal as $n \to \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Where $\mu = \mathrm{E}\left[X_i\right]$ and $\sigma^2 = \mathrm{Var}\left(X_i\right)$.

Shannon's Information Theory

Claude Shannon (1948): A Mathematical Theory of Communication

Shannon's measure of information is the number of bits to represent the amount of uncertainty (randomness) in a data source, and is defined as entropy

$$H = -\sum_{i=1}^{n} p_i \log (p_i)$$

Where there are n symbols $1, 2, \ldots \quad n$, each with probability of occurrence of p_i

Justification of Shannon's Entropy

- A set of possible events with probabilities p_i $(1 \leq i \leq n)$.
- Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome? Denote it as $H(p_1, p_2, \ldots, p_n)$.
- (Axiomatic approach) $H()$ should satisfy the following properties:
    - H should be continuous in each p_i.
    - H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
    - If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

1a A vs $\{B, C\}$: $\frac{1}{2}$ vs $\frac{1}{2}$

1b B vs C: $\frac{2}{3}$ vs $\frac{1}{3}$

2 A vs B vs C: $\frac{1}{2}$ vs $\frac{1}{3}$ vs $\frac{1}{6}$

$H(\frac{1}{2}, \frac{1}{2}) + H(\frac{2}{3}, \frac{2}{3})$

$H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$

Roadmap

(1) Construction of a Probability Space

(2) Discrete and Continuous Probabilities

(3) Sum Rule, Product Rule, and Bayes' Theorem

(4) Change of Variables/Inverse Transform

(5) Entropy and KL Divergence

Measure Distance Between Two Distributions

Applications in Computer Science:

- **Machine Learning:**
  - **Model Evaluation:** Comparing predicted vs. true distributions.
  - **Generative Models:** Ensuring generated data resembles real data.
- **Information Theory:**
  - **Encoding Efficiency:** Measuring information loss.

Challenge: How to measure the distance between $B(n, p_1$ and $B(n_{p_2}$, where $B(n, p)$ is the Binomial distribution?

- $\|p_1 - p_2\|$? $B(n = 10, 0.2)$ and $B(n = 10, 0.1)$ vs $B(n = 10, 0.4)$ and $B(n = 10, 0.5)$
- Lesson learned: need to compare the PMFs (PDFs), not the parameters

Common Metrics

- Euclidean (L_2) Distance:
  - $\|u, v\|_p = (\sum_i |u_i - v_i|^p)^{\frac{1}{p}}$, $p = 2$
  - Not suitable for probability distributions.
- Manhattan (L_1) Distance:
  - $p = 1$
  - Ignores underlying distribution properties.
- **Kullback-Leibler (KL) Divergence** and its generalization:
  - Asymmetric and be careful of its interpretation.
- Wasserstein Distance (Earth Mover's Distance):
  - Measures the minimum "cost" required to transform one distribution into another, based on moving "mass" in a metric space. It's particularly useful for distributions defined on continuous spaces.
  - Related to Optimal Transport.

Kullback-Leibler (KL) Divergence

Notation:

- P: True distribution
- Q: Approximate distribution
- $D_{\mathrm{KL}}(P \parallel Q)$: KL divergence from Q to P

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \qquad \text{(discrete case)}$$

or

$$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx \qquad \text{(continuous case)}$$

Intuition behind KL

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right) = - \left(\sum_x P(x) \log Q(x) - \sum_x P(x) \log P(x) \right)$$
$$= \sum_x P(x) \log \frac{1}{Q(x)} - \sum_x P(x) \log \frac{1}{P(x)} = H(P, Q) - H(P)$$

1. $H(P, Q)$: Average code length of a source P with estimated distribution Q
   - $\log \frac{1}{Q(x)}$: Use $\log \frac{1}{Q(x)}$ bits (assuming base $= 2$) to encode the message x.
   - **Expectation Over P:** the average code length.
2. $D_{\mathrm{KL}}(P \parallel Q)$ describes the excessive number of bits needed to encode the true distribution P using an estimated distribution Q.

Properties

- **Non-Negativity:** $D_{\mathrm{KL}}(P \parallel Q) \geq 0$
- **Zero Divergence:** $D_{\mathrm{KL}}(P \parallel Q) = 0 \iff P = Q$ almost everywhere
- **Asymmetric:** $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$
  - Implications: Changing the order of distributions changes the divergence value.
  - $D_{\mathrm{KL}}(P \parallel Q)$ measures the expected information loss when Q is used to approximate P, weighted by P.
  - Different Emphasis: the asymmetry arises because P and Q place different weights on outcomes.

Example

Let P and Q Be Simple Distributions

- **Distribution P:**
  - $P(0) = 0.9$
  - $P(1) = 0.1$
- **Distribution Q:**
  - $Q(0) = 0.5$
  - $Q(1) = 0.5$

Calculate $D_{\mathbf{KL}}(P \parallel Q)$:

$$= 0.9 \log\left(\frac{0.9}{0.5}\right) + 0.1 \log\left(\frac{0.1}{0.5}\right) \approx 0.9 \times 0.847 + 0.1 \times (-1.609) \approx 0.762 - 0.161 = 0.601 \text{ bits}$$

Calculate $D_{\mathbf{KL}}(Q \parallel P)$:

$$= 0.5 \log\left(\frac{0.5}{0.9}\right) + 0.5 \log\left(\frac{0.5}{0.1}\right) \approx 0.5 \times (-0.847) + 0.5 \times 1.609 \approx -0.423 + 0.805 = 0.382 \text{ bits}$$

Observation:

$D_{\mathrm{KL}}(P \parallel Q) > D_{\mathrm{KL}}(Q \parallel P)$

3. Why KL Divergence is Asymmetric

Expectation Basis

- $D_{\mathrm{KL}}(P \parallel Q)$ measures the expected information loss when Q is used to approximate P, weighted by P.

Different Emphasis

- The asymmetry arises because P and Q place different weights on outcomes.

Impact of Asymmetry on "Nearest" Distribution

- **Task:** Find a distribution Q that is "closest" to P based on a chosen divergence measure.

Asymmetric Implications

1. **Direction Matters:**
   - $D_{\mathrm{KL}}(P \parallel Q)$ aims to minimize information loss when approximating P with Q.
   - Minimizing $D_{\mathrm{KL}}(Q \parallel P)$ focuses on different aspects, potentially highlighting different "closeness."
2. **Mode Seeking vs. Mean Covering:**
   - $D_{\mathrm{KL}}(P \parallel Q)$ tends to be **mode-seeking**:
     - Q covers the modes of P but might miss some support, because ...
   - $D_{\mathrm{KL}}(Q \parallel P)$ tends to be **mean-covering**:
     - Q covers all support of P, potentially assigning probability to regions where P has low probability, because ...

An Example

Scenario: Approximating a Distribution

- **True Distribution** P**:** Highly concentrated around several values.
- **Candidate Distribution** Q**:** More spread out.

Using $D_{\mathbf{KL}}(P \parallel Q)$**:**

- Q adjusts to cover the peaks of P, potentially ignoring low-probability regions.

Using $D_{\mathbf{KL}}(Q \parallel P)$**:**

- Q must cover all regions where P has support, avoiding assigning probability mass where P is zero or near-zero.

Choosing the Direction

Task Dependent:

- **Information Loss Minimization:** Use $D_{\mathrm{KL}}(P \parallel Q)$.
- **Support Coverage:** Use $D_{\mathrm{KL}}(Q \parallel P)$.

Model Selection:

- The asymmetry influences which aspects of the distribution are prioritized in modeling.