Probabilistic Systems Analysis

by Fabián Kozynski

Probability

Probability models and axioms

Definition (Sample space) A sample space Ω is the set of all possible outcomes. The set's elements must be mutually exclusive, collectively exhaustive and at the right granularity.

Definition (Event) An event is a subset of the sample space. Probability is assigned to events.

Definition (Probability axioms) A probability law $\mathbb{P}$ assigns probabilities to events and satisfies the following axioms:

Nonnegativity $\mathbb{P}(A) \geq 0$ for all events A.

Normalization $\mathbb{P}(\Omega) = 1$.

(Countable) additivity For every sequence of events $A_1, A_2, \ldots$ such that $A_i \cap A_j = \varnothing$: $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

Corollaries (Consequences of the axioms)

- $\mathbb{P}(\varnothing) = 0$.
- For any finite collection of disjoint events $A_1, \ldots, A_n$, $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i)$.
- $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.
- $\mathbb{P}(A) \leq 1$.
- If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Example (Discrete uniform law) Assume Ω is finite and consists of n equally likely elements. Also, assume that $A \subset \Omega$ with k elements. Then $\mathbb{P}(A) = \frac{k}{n}$.

Conditioning and Bayes' rule

Definition (Conditional probability) Given that event B has occurred and that $P(B) > 0$, the probability that A occurs is

$$\mathbb{P}(A|B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark (Conditional probabilities properties) They are the same as ordinary probabilities. Assuming $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A|B) \geq 0$.
- $\mathbb{P}(\Omega|B) = 1$
- $\mathbb{P}(B|B) = 1$.
- If $A \cap C = \varnothing$, $\mathbb{P}(A \cup C|B) = \mathbb{P}(A|B) + \mathbb{P}(C|B)$.

Proposition (Multiplication rule)

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Theorem (Total probability theorem) Given a partition $\{A_1, A_2, \ldots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and for every event B, we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$

Theorem (Bayes' rule) Given a partition $\{A_1, A_2, \ldots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and if $\mathbb{P}(A_i) > 0$ for all i, then for every event B, the conditional probabilities $\mathbb{P}(A_i|B)$ can be obtained from the conditional probabilities $\mathbb{P}(B|A_i)$ and the initial probabilities $\mathbb{P}(A_i)$ as follows:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

Independence

Definition (Independence of events) Two events are independent if occurrence of one provides no information about the other. We say that A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Equivalently, as long as $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

$$\mathbb{P}(B|A) = \mathbb{P}(B) \qquad \mathbb{P}(A|B) = \mathbb{P}(A).$$

Remarks

- The definition of independence is symmetric with respect to A and B.
- The product definition applies even if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Corollary If A and B are independent, then A and B^c are independent. Similarly for A^c and B, or for A^c and B^c.

Definition (Conditional independence) We say that A and B are independent conditioned on C, where $\mathbb{P}(C) > 0$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Definition (Independence of a collection of events) We say that events $A_1, A_2, \ldots, A_n$ are independent if for every collection of distinct indices $i_1, i_2, \ldots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Counting

This section deals with finite sets with uniform probability law. In this case, to calculate $\mathbb{P}(A)$, we need to count the number of elements in A and in Ω.

Remark (Basic counting principle) For a selection that can be done in r stages, with n_i choices at each stage i, the number of possible selections is $n_1 \cdot n_2 \cdots n_r$.

Definition (Permutations) The number of permutations (orderings) of n different elements is

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Definition (Combinations) Given a set of n elements, the number of subsets with exactly k elements is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Definition (Partitions) We are given an n–element set and nonnegative integers $n_1, n_2, \ldots, n_r$, whose sum is equal to n. The number of partitions of the set into r disjoint subsets, with the i^{th} subset containing exactly n_i elements, is equal to

$$\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

Remark This is the same as counting how to assign n distinct elements to r people, giving each person i exactly n_i elements.

Discrete random variables

Probability mass function and expectation

Definition (Random variable) A random variable X is a function of the sample space Ω into the real numbers (or $\mathbb{R}^n$). Its range can be discrete or continuous.

Definition (Probability mass funtion (PMF)) The probability law of a discrete random variable X is called its PMF. It is defined as

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) = x\}\right).$$

Properties

$p_X(x) \geq 0$, $\forall\, x$.

$\sum_x p_X(x) = 1$.

Example (Bernoulli random variable) A Bernoulli random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Ber}(p)$) takes the following values:

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

An indicator random variable of an event ($I_A = 1$ if A occurs) is an example of a Bernoulli random variable.

Example (Discrete uniform random variable) A Discrete uniform random variable X between a and b with $a \leq b$ ($X \sim \text{Uni}[a, b]$) takes any of the values in $\{a, a+1, \ldots, b\}$ with probability $\frac{1}{b-a+1}$.

Example (Binomial random variable) A Binomial random variable X with parameters n (natural number) and $0 \leq p \leq 1$ ($X \sim \text{Bin}(n, p)$) takes values in the set $\{0, 1, \ldots, n\}$ with probabilities $p_X(i) = \binom{n}{i} p^i (1-p)^{n-i}$.
It represents the number of successes in n independent trials where each trial has a probability of success p. Therefore, it can also be seen as the sum of n independent Bernoulli random variables, each with parameter p.

Example (Geometric random variable) A Geometric random variable X with parameter $0 \leq p \leq 1$ ($X \sim \text{Geo}(p)$) takes values in the set $\{1, 2, \ldots\}$ with probabilities $p_X(i) = (1-p)^{i-1} p$.
It represents the number of independent trials until (and including) the first success, when the probability of success in each trial is p.

Definition (Expectation/mean of a random variable) The expectation of a discrete random variable is defined as

$$\mathbb{E}[X] \triangleq \sum_x x p_X(x).$$

assuming $\sum_x |x| p_X(x) < \infty$.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- If $X = c$ then $\mathbb{E}[X] = c$.

Example Expected value of know r.v.

- If $X \sim \text{Ber}(p)$ then $\mathbb{E}[X] = p$.
- If $X = I_A$ then $\mathbb{E}[X] = \mathbb{P}(A)$.
- If $X \sim \text{Uni}[a, b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$.
- If $X \sim \text{Bin}(n, p)$ then $\mathbb{E}[X] = np$.
- If $X \sim \text{Geo}(p)$ then $\mathbb{E}[X] = \frac{1}{p}$.

Theorem (Expected value rule) Given a random variable X and a function $g:\mathbb{R}\to\mathbb{R}$, we construct the random variable $Y=g(X)$. Then

$$\sum_y yp_Y(y)=\mathbb{E}[Y]=\mathbb{E}[g(X)]=\sum_x g(x)p_X(x).$$

Remark (PMF of $Y=g(X)$) The PMF of $Y=g(X)$ is $p_Y(y)=\sum_{x:g(x)=y}p_X(x)$.

Remark In general $g(\mathbb{E}[X])\neq\mathbb{E}[g(X)]$. They are equal if $g(x)=ax+b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu=\mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\mathrm{Var}(X)\triangleq\mathbb{E}\left[(X-\mu)^2\right]=\sum_x(x-\mu)^2p_X(x).$$

Definition (Standard deviation)

$$\sigma_X=\sqrt{\mathrm{Var}(X)}.$$

Properties (Properties of the variance)

- $\mathrm{Var}(aX)=a^2\,\mathrm{Var}(X)$, for all $a\in\mathbb{R}$.
- $\mathrm{Var}(X+b)=\mathrm{Var}(X)$, for all $b\in\mathbb{R}$.
- $\mathrm{Var}(aX+b)=a^2\,\mathrm{Var}(X)$.
- $\mathrm{Var}(X)=\mathbb{E}[X^2]-(\mathbb{E}[X])^2$.

Example (Variance of known r.v.)

- If $X\sim\mathrm{Ber}(p)$, then $\mathrm{Var}(X)=p(1-p)$.
- If $X\sim\mathrm{Uni}[a,b]$, then $\mathrm{Var}(X)=\frac{(b-a)(b-a+2)}{12}$.
- If $X\sim\mathrm{Bin}(n,p)$, then $\mathrm{Var}(X)=np(1-p)$.
- If $X\sim\mathrm{Geo}(p)$, then $\mathrm{Var}(X)=\frac{1-p}{p^2}$

Proposition (Conditional PMF and expectation, given an event) Given the event A, with $\mathbb{P}(A)>0$, we have the following

- $p_{X|A}(x)=\mathbb{P}(X=x|A)$.
- If A is a subset of the range of X, then:
  $p_{X|A}(x)\triangleq p_{X|\{X\in A\}}(x)=\begin{cases}\frac{1}{\mathbb{P}(A)}p_X(x), & \text{if } x\in A,\\ 0, & \text{otherwise.}\end{cases}$
- $\sum_x p_{X|A}(x)=1$.
- $\mathbb{E}[X|A]=\sum_x xp_{X|A}(x)$.
- $\mathbb{E}[g(X)|A]=\sum_x g(x)p_{X|A}(x)$.

Proposition (Total expectation rule) Given a partition of disjoint events $A_1,\dots,A_n$ such that $\sum_i\mathbb{P}(A_i)=1$, and $\mathbb{P}(A_i)>0$,

$$\mathbb{E}[X]=\mathbb{P}(A_1)\mathbb{E}[X|A_1]+\cdots+\mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Memorylessness of the geometric random variable) When we condition a geometric random variable X on the event $X>n$ we have memorylessness, meaning that the "remaining time" $X-n$, given that $X>n$, is also geometric with the same parameter. Formally,

$$p_{X-n|X>n}(i)=p_X(i).$$

Definition (Joint PMF) The joint PMF of random variables $X_1,X_2,\dots,X_n$ is $p_{X_1,X_2,\dots,X_n}(x_1,\dots,x_n)=\mathbb{P}(X_1=x_1,\dots,X_n=x_n)$.

Properties (Properties of joint PMF)

- $\sum_{x_1}\cdots\sum_{x_n}p_{X_1,\dots,X_n}(x_1,\dots,x_n)=1$.
- $p_{X_1}(x_1)=\sum_{x_2}\cdots\sum_{x_n}p_{X_1,\dots,X_n}(x_1,x_2,\dots,x_n)$.
- $p_{X_2,\dots,X_n}(x_2,\dots,x_n)=\sum_{x_1}p_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)$.

Definition (Functions of multiple r.v.) If $Z=g(X_1,\dots,X_n)$, where $g:\mathbb{R}^n\to\mathbb{R}$, then $p_Z(z)=\mathbb{P}(g(X_1,\dots,X_n)=z)$.

Proposition (Expected value rule for multiple r.v.) Given $g:\mathbb{R}^n\to\mathbb{R}$,

$$\mathbb{E}[g(X_1,\dots,X_n)]=\sum_{x_1,\dots,x_n}g(x_1,\dots,x_n)p_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

Properties (Linearity of expectations)

- $\mathbb{E}[aX+b]=a\mathbb{E}[X]+b$.
- $\mathbb{E}[X_1+\cdots+X_n]=\mathbb{E}[X_1]+\cdots+\mathbb{E}[X_n]$.

Conditioning on a random variable, independence

Definition (Conditional PMF given another random variable) Given discrete random variables X,Y and y such that $p_Y(y)>0$ we define

$$p_{X|Y}(x|y)\triangleq\frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

Proposition (Multiplication rule) Given jointly discrete random variables X,Y, and whenever the conditional probabilities are defined,

$$p_{X,Y}(x,y)=p_X(x)p_{Y|X}(y|x)=p_Y(y)p_{X|Y}(x|y).$$

Definition (Conditional expectation) Given discrete random variables X,Y and y such that $p_Y(y)>0$ we define

$$\mathbb{E}[X|Y=y]=\sum_x xp_{X|Y}(x|y).$$

Additionally we have

$$\mathbb{E}[g(X)|Y=y]=\sum_x g(x)p_{X|Y}(x|y).$$

Theorem (Total probability and expectation theorems) If $p_Y(y)>0$, then

$$p_X(x)=\sum_y p_Y(y)p_{X|Y}(x|y),$$
$$\mathbb{E}[X]=\sum_y p_Y(y)\mathbb{E}[X|Y=y].$$

Definition (Independence of a random variable and an event) A discrete random variable X and an event A are independent if $\mathbb{P}(X=x\text{ and }A)=p_X(x)\mathbb{P}(A)$, for all x.

Definition (Independence of two random variables) Two discrete random variables X and Y are independent if $p_{X,Y}(x,y)=p_X(x)p_Y(y)$ for all x,y.

Remark (Independence of a collection of random variables) A collection $X_1,X_2,\dots,X_n$ of random variables are independent if

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n)=p_{X_1}(x_1)\cdots p_{X_n}(x_n),\ \forall\,x_1,\dots,x_n.$$

Remark (Independence and expectation) In general, $\mathbb{E}[g(X,Y)]\neq g(\mathbb{E}[X],\mathbb{E}[Y])$. An exception is for linear functions: $\mathbb{E}[aX+bY]=a\mathbb{E}[X]+b\mathbb{E}[Y]$.

Proposition (Expectation of product of independent r.v.) If X and Y are discrete independent random variables,

$$\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)]=\mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) IF X and Y are discrete independent random variables,

$$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y).$$

Continuous random variables

PDF, Expectation, Variance, CDF

Definition (Probability density function (PDF)) A probability density function of a r.v. X is a non-negative real valued function f_X that satisfies the following

- $\int_{-\infty}^{\infty}f_X(x)\mathrm{d}x=1$.
- $\mathbb{P}(a\le X\le b)=\int_a^b f_X(x)\mathrm{d}x$ for some random variable X.

Definition (Continuous random variable) A random variable X is continuous if its probability law can be described by a PDF f_X.

Remark Continuous random variables satisfy:

- For small $\delta>0$, $\mathbb{P}(a\le X\le a+\delta)\approx f_X(a)\delta$.
- $\mathbb{P}(X=a)=0$, $\forall a\in\mathbb{R}$.

Definition (Expectation of a continuous random variable) The expectation of a continuous random variable is

$$\mathbb{E}[X]\triangleq\int_{-\infty}^{\infty}xf_X(x)\mathrm{d}x.$$

assuming $\int_{-\infty}^{\infty}|x|f_X(x)\mathrm{d}x<\infty$.

Properties (Properties of expectation)

- If $X\ge0$ then $\mathbb{E}[X]\ge0$.
- If $a\le X\le b$ then $a\le\mathbb{E}[X]\le b$.
- $\mathbb{E}[g(X)]=\int_{-\infty}^{\infty}g(x)f_X(x)\mathrm{d}x$.
- $\mathbb{E}[aX+b]=a\mathbb{E}[X]+b$.

Definition (Variance of a continuous random variable) Given a continuous random variable X with $\mu=\mathbb{E}[X]$, its variance is

$$\mathrm{Var}(X)=\mathbb{E}\left[(X-\mu)^2\right]=\int_{-\infty}^{\infty}(x-\mu)^2f_X(x)\mathrm{d}x.$$

It has the same properties as the variance of a discrete random variable.

Example (Uniform continuous random variable) A Uniform continuous random variable X between a and b, with $a<b$, $(X\sim\mathrm{Uni}(a,b))$ has PDF

$$f_X(x)=\begin{cases}\frac{1}{b-a}, & \text{if } a<x<b,\\ 0, & \text{otherwise.}\end{cases}$$

We have $\mathbb{E}[X]=\frac{a+b}{2}$ and $\mathrm{Var}(X)=\frac{(b-a)^2}{12}$.

Example (Exponential random variable) An Exponential random variable X with parameter $\lambda > 0$ ($X \sim Exp(\lambda)$) has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We have $E[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

Definition (Cumulative Distribution Function (CDF)) The CDF of a random variable X is $F_X(x) = \mathbb{P}(X \leq x)$.
In particular, for a continuous random variable, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\mathrm{d}x,$$

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}.$$

Properties (Properties of CDF)

- If $y \geq x$, then $F_X(y) \geq F_X(x)$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.

Definition (Normal/Gaussian random variable) A Normal random variable X with mean μ and variance $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) has PDF

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

We have $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Remark (Standard Normal) The standard Normal is $\mathcal{N}(0, 1)$.

Proposition (Linearity of Gaussians) Given $X \sim \mathcal{N}(\mu, \sigma^2)$, and if $a \neq 0$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
Using this $Y = \frac{X-\mu}{\sigma}$ is a standard gaussian.

Conditioning on an event, and multiple continuous r.v.

Definition (Conditional PDF given an event) Given a continuous random variable X and event A with $P(A) > 0$, we define the conditional PDF as the function that satisfies

$$\mathbb{P}(X \in B|A) = \int_B f_{X|A}(x)\mathrm{d}x.$$

Definition (Conditional PDF given $X \in A$) Given a continuous random variable X and an $A \subset \mathbb{R}$, with $P(A) > 0$:

$$f_{X|X\in A}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} f_X(x), & x \in A, \\ 0, & x \notin A. \end{cases}$$

Definition (Conditional expectation) Given a continuous random variable X and an event A, with $P(A) > 0$:

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} f_{X|A}(x)\mathrm{d}x.$$

Definition (Memorylessness of the exponential random variable) When we condition an exponential random variable X on the event $X > t$ we have memorylessness, meaning that the "remaining time" $X - t$ given that $X > t$ is also geometric with the same parameter i.e.,

$$\mathbb{P}(X - t > x|X > t) = \mathbb{P}(X > x).$$

Theorem (Total probability and expectation theorems) Given a partition of the space into disjoint events $A_1, A_2, \ldots, A_n$ such that $\sum_i \mathbb{P}(A_i) = 1$ we have the following:

$$F_X(x) = \mathbb{P}(A_1)F_{X|A_1}(x) + \cdots + \mathbb{P}(A_n)F_{X|A_n}(x),$$
$$f_X(x) = \mathbb{P}(A_1)f_{X|A_1}(x) + \cdots + \mathbb{P}(A_n)f_{X|A_n}(x),$$
$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \cdots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Jointly continuous random variables) A pair (collection) of random variables is jointly continuous if there exists a joint PDF $f_{X,Y}$ that describes them, that is, for every set $B \subset \mathbb{R}^n$

$$\mathbb{P}((X, Y) \in B) = \iint_B f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y.$$

Properties (Properties of joint PDFs)

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y$.
- $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{X,Y}(u, v)\mathrm{d}v \right] \mathrm{d}u$.
- $f_{X,Y}(x) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x\, \partial y}$.

Example (Uniform joint PDF on a set S) Let $S \subset \mathbb{R}^2$ with area $s > 0$, then the random variable (X, Y) is uniform over S if it has PDF

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{s}, & (x, y) \in S, \\ 0, & (x, y) \notin S. \end{cases}$$

Conditioning on a random variable, independence, Bayes' rule

Definition (Conditional PDF given another random variable) Given jointly continuous random variables X, Y and a value y such that $f_Y(y) > 0$, we define the conditional PDF as

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

Additionally we define $\mathbb{P}(X \in A|Y = y) \int_A f_{X|Y}(x|y)\mathrm{d}x$.

Proposition (Multiplication rule) Given jointly continuous random variables X, Y, whenever possible we have

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y).$$

Definition (Conditional expectation) Given jointly continuous random variables X, Y, and y such that $f_Y(y) > 0$, we define the conditional expected value as

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\mathrm{d}x.$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y)\mathrm{d}x.$$

Theorem (Total probability and total expectation theorems)

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y)\mathrm{d}y,$$
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} f_Y(y)\mathbb{E}[X|Y = y]\mathrm{d}y.$$

Definition (Independence) Jointly continuous random variables X, Y are independent if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all x, y.

Proposition (Expectation of product of independent r.v.) If X and Y are independent continuous random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables) If X and Y are independent continuous random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Proposition (Bayes' rule summary)

- For X, Y discrete: $p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.
- For X, Y continuous: $f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.
- For X discrete, Y continuous: $p_{X|Y}(x|y) = \frac{p_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.
- For X continuous, Y discrete: $f_{X|Y}(x|y) = \frac{f_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.

Derived distributions

Proposition (Discrete case) Given a discrete random variable X and a function g, the r.v. $Y = g(X)$ has PMF

$$p_Y(y) = \sum_{x:g(x)=y} p_X(x).$$

Remark (Linear function of discrete random variable) If $g(x) = ax + b$, then $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

Proposition (Linear function of continuous r.v.) Given a continuous random variable X and $Y = aX + b$, with $a \neq 0$, we have

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Corollary (Linear function of normal r.v.) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a \neq 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Example (General function of a continuous r.v.) If X is a continuous random variable and g is any function, to obtain the pdf of $Y = g(X)$ we follow the two-step procedure:

1. Find the CDF of Y: $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.
2. Differentiate the CDF of Y to obtain the PDF: $f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y}$.

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{\mathrm{d}h}{\mathrm{d}y}(y) \right|.$$

where $h = g^{-1}$ in the interval where g is monotonic.

Sums of independent r.v., covariance and correlation

Proposition (Discrete case) Let X, Y be discrete independent random variables and $Z = X + Y$, then the PMF of Z is

$$p_Z(z) = \sum_x p_X(x) p_Y(z - x).$$

Proposition (Continuous case) Let X, Y be continuous independent random variables and $Z = X + Y$, then the PDF of Z is

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

Proposition (Sum of independent normal r.v.) Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ independent. Then $Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Definition (Covariance) We define the covariance of random variables X, Y as

$$\mathrm{Cov}(X, Y) \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

Properties (Properties of covariance)

- If X, Y are independent, then $\mathrm{Cov}(X, Y) = 0$.
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- $\mathrm{Cov}(aX + b, Y) = a \mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$.
- $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Proposition (Variance of a sum of r.v.)

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j).$$

Definition (Correlation coefficient) We define the correlation coefficient of random variables X, Y, with $\sigma_X, \sigma_Y > 0$, as

$$\rho(X, Y) \triangleq \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Properties (Properties of the correlation coefficient)

- $-1 \leq \rho \leq 1$.
- If X, Y are independent, then $\rho = 0$.
- $|\rho| = 1$ if and only if $X - \mathbb{E}[X] = c(Y - \mathbb{E}[Y])$.
- $\rho(aX + b, Y) = \mathrm{sign}(a)\rho(X, Y)$.

Conditional expectation and variance, sum of random number of r.v.

Definition (Conditional expectation as a random variable) Given random variables X, Y the conditional expectation $\mathbb{E}[X|Y]$ is the random variable that takes the value $\mathbb{E}[X|Y = y]$ whenever $Y = y$.

Theorem (Law of iterated expectations)

$$\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X].$$

Definition (Conditional variance as a random variable) Given random variables X, Y the conditional variance $\mathrm{Var}(X|Y)$ is the random variable that takes the value $\mathrm{Var}(X|Y = y)$ whenever $Y = y$.

Theorem (Law of total variance)

$$\mathrm{Var}(X) = \mathbb{E}\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left(\mathbb{E}[X|Y]\right).$$

Proposition (Sum of a random number of independent r.v.)
Let N be a nonnegative integer random variable.
Let $X, X_1, X_2, \ldots, X_N$ be i.i.d. random variables.
Let $Y = \sum_i X_i$. Then

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X],$$

$$\mathrm{Var}(Y) = \mathbb{E}[N]\mathrm{Var}(X) + (\mathbb{E}[X])^2 \mathrm{Var}(N).$$

Inference

Introduction to Bayesian Inference

In this framework, we have an unknown parameter Θ which has a prior distribution (p_Θ or f_Θ). Additionally we have an observation X with some observation model ($p_{X|\Theta}$ or $f_{X|\Theta}$). To do inference, we use the appropriate version of Bayes' rule. The complete answer is a posterior distribution $p_{\Theta|X}$ or $f_{\Theta|X}$.

Definition (Estimator) An estimator $\hat{\Theta} = g(X)$ is a random variable that takes a value for each realization of the observation.

Definition (Estimate) An estimate $\hat{\theta} = g(x)$ is a number which results after applying our estimator function to the realization of the observation.

Definition (MAP estimator) A MAP estimator is one that maximizes the posterior probability given an observation x, that is

$$\hat{\theta}_{MAP} = \begin{cases} \text{argmax}_\theta \, p_{\Theta|X}(\theta|x) & \theta \text{ discrete,} \\ \text{argmax}_\theta \, f_{\Theta|X}(\theta|x) & \theta \text{ continuous.} \end{cases}$$

Definition (Probability of error) If Θ and X are discrete, the posterior probability of error is defined as

$$\mathbb{P}(\Theta \neq \hat{\theta}|X = x),$$

and the overall probability of error is

$$\mathbb{P}(\Theta \neq \hat{\Theta}) = \sum_x \mathbb{P}(\Theta \neq \hat{\Theta}|X = x)p_X(x) = \sum_\theta \mathbb{P}(\Theta \neq \hat{\Theta}|\Theta = \theta)p_\Theta(\theta).$$

There are analogous formulas when X is continuous.

Remark The MAP estimator minimizes the conditional probability of error as well as the overall probability of error.

Definition (LMS estimator) The LMS estimate is the conditional expected value given an observation x, that is

$$\hat{\theta}_{LMS} = \mathbb{E}[\Theta|X = x].$$

Linear models with normal noise

Proposition (Recognizing normal PDFs) If a random variable X has PDF

$$f_X(x) = c \cdot \mathrm{e}^{-(\alpha x^2 + \beta x + \gamma)},$$

with $\alpha > 0$, then $X \sim \mathcal{N}\left(-\frac{\beta}{2\alpha}, \frac{1}{2\alpha}\right)$.

Proposition (Estimating a normal parameter with WAGN) Let $\Theta \sim \mathcal{N}(0,1)$, $W \sim \mathcal{N}(0,1)$ independent of Θ, and $X = \Theta + W$. Then $\hat{\Theta}_{LMS} = \hat{\Theta}_{MAP} = \frac{X}{2}$.

Proposition (The case of multiple observations) Let $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$ and $W_i \sim \mathcal{N}(0, \sigma_i^2)$ independent and independent of Θ, for i from 1 to n. Let $X_i = \Theta + W_i$ be the observations. Then

$$f_{\Theta|X}(\theta|\boldsymbol{x}) \propto \prod_{i=0}^{n} \mathrm{e}^{-\frac{1}{2\sigma_i^2}(x_i - \theta)^2},$$

and then

$$\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = \frac{\sum_{i=0}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}.$$

Proposition (The mean squared error) In the normal case

$$\mathbb{E}\left[(\Theta - \hat{\Theta})^2|X = x\right] = \mathbb{E}\left[(\Theta - \hat{\Theta})^2\right] = \text{Var}(\Theta|X = x) = \left(\sum_{i=0}^n \frac{1}{\sigma_i^2}\right)^{-1}.$$

Least mean squares (LMS) estimation

Definition (Mean squared error) We define the mean squared error between a random variable Θ and an estimator $\hat{\Theta}$ as $\mathbb{E}\left[(\Theta - \hat{\Theta})^2\right]$.

Definition (LMS estimator) The LMS estimator minimizes the mean squared error of the estimator, given the observations.

Proposition (LMS estimation in the absence of observations) We want to minimize $\mathbb{E}\left[(\Theta - \hat{\Theta})^2\right]$. Then $\hat{\theta} = \mathbb{E}[\Theta]$.
The optimal mean squared error is $\text{Var}(\Theta)$.

Proposition (LMS estimation based on X) In this case, we minimize the conditional mean squared error $\mathbb{E}\left[(\Theta - \hat{\Theta})^2|X = x\right]$. This yields $\hat{\theta} = \mathbb{E}[\Theta|X = x]$.

Theorem (LMS estimator minimizes mean squared error) $\hat{\Theta}_{LMS} = \mathbb{E}[\Theta|X]$ minimizes $\mathbb{E}\left[(\Theta - g(X))^2\right]$ among all estimators $\hat{\Theta} = g(X)$.

Remark (Performance of the LMS estimator) Given a measurement x, the mean squared error of the LMS estimator is $\text{Var}(\Theta|X = x)$. The expected performance of this estimator is $\mathbb{E}\left[\text{Var}(\Theta|X)\right]$.

Definition (Estimation error) Given the LMS estimator $\hat{\Theta}$, we define the error as $\tilde{\Theta} = \hat{\Theta} - \Theta$.

Properties (Estimation error)

- $\mathbb{E}[\tilde{\Theta}|X = x] = 0$.
- $\text{Cov}(\tilde{\Theta}, \Theta) = 0$.
- $\text{Var}(\Theta) = \text{Var}(\hat{\Theta}) + \text{Var}(\tilde{\Theta})$.

Linear least mean squares (LLMS) estimation

Definition (Linear least mean squares estimator) The linear least mean squares estimator is the linear estimator that minimizes the mean squared, that is $\hat{\Theta}_{LLMS} = aX + b$ with a and b such that

$$\mathbb{E}\left[(\Theta - aX - b)^2\right]$$

is minimized.

Corollary If $\hat{\Theta}_{LMS}$ is linear, then $\hat{\Theta}_{LLMS} = \hat{\Theta}_{LMS}$.

Proposition (Solution to the LLMS problem) If $\sigma_X > 0$,

$$\begin{aligned} \hat{\Theta}_{LLMS} &= \mathbb{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{Var}(X)}(X - \mathbb{E}[X]), \\ &= \mathbb{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - \mathbb{E}[X]). \end{aligned}$$

Remarks

- If $\rho = 0$, then $\hat{\Theta}_{LLMS} = \mathbb{E}[\Theta]$.
- $\mathbb{E}\left[(\hat{\Theta}_{LLMS} - \Theta)^2\right] = (1 - \rho^2)\,\text{Var}(\Theta)$.
- Solution only depends on means, variances and covariances

Remark (Multiple observations) Given that for the normal case the LMS estimator is linear, we know the form of the LLMS. As the LLMS only depends on the means and variances, the same formula is valid without assuming normality, as long as $X_i = \Theta + W_i$, with $\Theta, W_1, \ldots, W_n$ independent:

$$\hat{\Theta}_{LLMS} = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^n \frac{X_i}{\sigma_i^2}}{\sum_{i=0}^n \frac{1}{\sigma_i^2}}.$$

Convergence of random variables

Inequalities, convergence, and the Weak Law of Large Numbers

Theorem (Markov inequality) Given a random variable $X \geq 0$ and, for every $a > 0$ we have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Theorem (Chebyshev inequality) Given a random variable X with $\mathbb{E}[X] = \mu$ and $\text{Var}(X) = \sigma^2$, for every $\epsilon > 0$ we have

$$\mathbb{P}\left(|X - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{\epsilon^2}.$$

Theorem (Weak Law of Large Number (WLLN)) Given a sequence of i.i.d. random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

$$M_n = \frac{1}{n}\sum_{i=1}^n X_i,$$

for every $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left(|M_n - \mu| \geq \epsilon\right) = 0.$$

Definition (Convergence in probability) A sequence of random variables $\{Y_i\}$ converges in probability to the random variable Y if

$$\lim_{n\to\infty} \mathbb{P}\left(|Y_i - Y| \geq \epsilon\right) = 0,$$

for every $\epsilon > 0$.

Properties (Properties of convergence in probability) If $X_n \to a$ and $Y_n \to b$ in probability, then

- $X_n + Y_n \to a + b$.
- If g is a continuous function, then $g(X_n) \to g(a)$.
- $\mathbb{E}[X_n]$ does not always converge to a.

The Central Limit Theorem

Theorem (Central Limit Theorem (CLT)) Given a sequence of independent random variables $\{X_1, X_2, \ldots\}$ with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

$$Z_n = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n X_i.$$

Then, for every z, we have

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \mathbb{P}(Z \leq z),$$

where $Z \sim \mathcal{N}(0,1)$.

Corollary (Normal approximation of a binomial) Let $X \sim Bin(n,p)$ with n large. Then S_n can be approximated by $Z \sim \mathcal{N}\left(np, np(1-p)\right)$.

Remark (De Moivre-Laplace 1/2 approximation) Let $X \sim Bin$, then $\mathbb{P}(X = i) = \mathbb{P}\left(i - \frac{1}{2} \leq X \leq i + \frac{1}{2}\right)$ and we can use the CLT to approximate the PMF of X.