Basics of the Probability Theory

(COMS 4770/5770 Notes)

Yan-Bin Jia

Oct 31, 2024

1 Probability and Random Variable

Suppose we run an experiment like throwing a die a number of times. Sometimes six dots show up on the top face, but more often fewer than six dots show up. We refer to six dots appearing at the top face as event A. Common sense tells us that the probability of event A occurring is 1/6, because every face is equally likely to appear at the top after a throw. The *probability* of event A is defined as the ratio of the number of times A occurs to the total number of outcomes.

We call an experiment a procedure that yields one of a given set of possible outcomes. The *sample space*, denoted S, of the experiment is the set of possible outcomes. An *event* A is a subset of the sample space. Laplace's definition of probability of an event A is

$$\Pr(A) = \frac{|A|}{|S|}.$$

Let A and B be events with $\Pr(B) > 0$. The *conditional probability* of event A given event B with non-zero probability is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

Example 1. Consider throwing a die a number of times. Suppose A refers to the appearance of 4 on a die, and B refers to the appearance of an even number. We have $\Pr(A) = 1/6$. But if we know that the die has an even number showing up at the top, then the probability becomes 1/3. Formally, we know that $\Pr(B) = 1/2$, thus

$$\Pr(A|B) = \frac{1/6}{1/2} = \frac{1}{3}.$$

The *a priori* probability of A is 1/6. But the *a posteriori* probability of A given B is 1/3.

A *random variable* is a function from the sample space S of an experiment to the set of real numbers. Namely, it assigns a real number to each possible outcome. For example, the roll of a die can be viewed as a random variable that maps the appearance of one dot to 1, the appearance of the two dots to 2, and so on. Of course, after a throw, the value of the die is no longer a random variable; it is certain. So the outcome of a particular experiment is not a random variable.

A random variable can be either continuous or discrete. The throw of a die is a discrete random variable, whereas the high temperature tomorrow is a continuous random variable whose outcome

takes on a continuous set of values. Given a random variable X, its *cumulative distribution function* (CDF), also called its distribution function, is defined as

$$D(x) = \Pr(X \leq x). \tag{1}$$

In the case of a discrete probability $\Pr(x)$, that is, the probability that a discrete random variable X assumes value x, we have

$$D(x) = \sum_{X \leq x} \Pr(X).$$

Some trivial properties of the CDF are listed below:

$$\begin{aligned} D(x) &\in [0,1], \\ D(-\infty) &= 0, \\ D(\infty) &= 1, \\ D(a) &\leq D(b), \qquad \text{if } a < b, \\ \Pr(a < X \leq b) &= D(b) - D(a). \end{aligned}$$

The *probability density function* (PDF) $P(x)$ of a continuous random variable is defined as the derivative of the distribution function $D(x)$:

$$P(x) = \frac{d}{dx} D(x) \tag{2}$$

So

$$D(x) \equiv \int_{-\infty}^{x} P(\xi)\, d\xi.$$

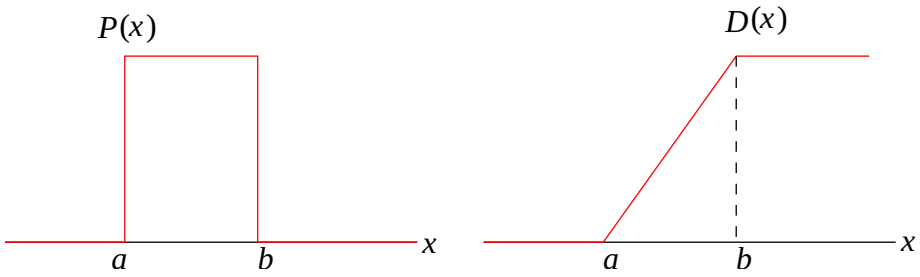
Some properties of the PDF can be obtained from the definition:

$$\begin{aligned} P(x) &\geq 0, \\ \int_{-\infty}^{\infty} P(x)\, dx &= 1, \\ D(a < x \leq b) &= \int_{a}^{b} P(x)\, dx. \end{aligned}$$

A *uniform distribution* has constant PDF. The probability density function and distribution function for a continuous uniform distribution on the interval $[a, b]$ are

$$P(x) = \begin{cases} 0, & \text{for } x < a, \\ \dfrac{1}{b-a}, & \text{for } a \leq x \leq b, \\ 0, & \text{for } x > b; \end{cases}$$

$$D(x) = \begin{cases} 0, & \text{for } x < a, \\ \dfrac{x-a}{b-a}, & \text{for } a \leq x \leq b, \\ 1, & \text{for } x > b. \end{cases}$$



2 Generating a Continuous Distribution

A programming language often has some built-in function for generating uniformly distributed pseudo-random numbers. For instance, in C++, we can use the function `rand()` to generate a pseudo-random number between 0 and 65535 after seeding the built-in random number generator `srand(`n`)`, where n is an integer.

The utility of pseudo-random numbers can be wielded for simulating a given distribution function $D(x)$. Here we present a method for generating continuous distributions as offered in [3, p. 121]. By its definition (1), the function $D(x)$ increases monotonically from zero to one. Suppose $D(x)$ is continuous and strictly increasing, there exists an inverse function $D^{-1}(y)$ such that, for $0 < y < 1$,

$$y = D(x) \qquad \text{if and only if} \qquad x = D^{-1}(y).$$

We can compute a random variable X with distribution $D(x)$ by setting

$$X = D^{-1}(Y),$$

where Y is a random variable with uniform distribution over $[0, 1]$. The reasoning is as follows:

$$\begin{aligned} \Pr(X \le x) &= \Pr(D^{-1}(Y) \le x) \\ &= \Pr(Y \le D(x)) \\ &= D(x), \end{aligned}$$

where the last equation above follows from the uniform distribution of Y.

Example 2 (uniformly distributed random points). Given a rectangle with length l and width w, we can easily generate random points uniformly distributed inside the rectangle. Simply plot a point at the position (X, Y), where X and Y are random variables uniformly distributed within the intervals $[0, l]$ and $[0, w]$, respectively.

Suppose we want to generate uniform random points inside a circle of radius ρ and centered at the origin. We can represent the location of such a point as $R(\cos\Theta, \sin\Theta)$, where R and Θ are two random variables with ranges $[0, \rho]$ and $[0, 2\pi]$, respectively. The value of Θ follows the uniform distribution since the polar angle of a random point is equally likely to take on any value in $[0, 2\pi]$. But the radius variable R does not. If we were to generate its values according to a uniform distribution over $[0, \rho]$, then the resulting points would be more concentrated near the origin than far from it.

Hence, we need to find a distribution for the radius variable R. Here, the uniform distribution is over the area. The probability of a randomly generated point falling with a circle centered at the origin with the radius r, $r \le \rho$, is $\pi r^2/(\pi\rho^2)$. Namely, we have the distribution function:

$$D(r) = \Pr(R \le r) = \frac{\pi r^2}{\pi\rho^2} = \frac{r^2}{\rho^2}.$$

Clearly, $0 \le D(r) \le 1$. We let $s = D(r)$ and obtain $r = \rho\sqrt{s}$. Introduce a random variable S with uniform distribution over $[0, 1]$. So R can be computed as $R = \rho\sqrt{S}$.

In summary, random points should be generated at positions $\rho\sqrt{S}(\cos\Theta, \sin\Theta)$, where S and Θ are random variables with uniform distributions over $[0, 1]$ and $[0, 2\pi]$.

3 Expected Value and Standard Deviation

The *expected value*, or *mean*, of a random variable X is its average value over a large number of experiments. Suppose we run the experiment N times and observe a total of m different outcomes. Among them, the outcome x_1 occurs n_1 times, x_2 occurs n_2 times, ..., and x_m occurs n_m times. Then the expected value is computed as

$$E(X) = \frac{1}{N}\sum_{i=1}^{m} x_i n_i.$$

EXAMPLE 3. Suppose we roll a die an infinite number of times. We expect that each number appears 1/6 of the time. So the expected value is

$$\begin{aligned} E(X) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{6} i \cdot \frac{n}{6} \\ &= \frac{7}{2}. \end{aligned}$$

For a continuous random variable X with probability density function $P(x)$, its expected value is given as

$$E(X) = \int_{-\infty}^{\infty} xP(x)\,dx. \tag{3}$$

A discrete random variable with N possible values $x_1, \ldots, x_N$ and mean μ has *variance*

$$\begin{aligned} \mathrm{var}(X) &= E\Big(\big(X - E(X)\big)^2\Big) \\ &= \sum_{i=1}^{N} \Pr(x_i)(x_i - \mu)^2, \end{aligned}$$

where $\Pr(x_k)$ is probability of the value x_k, for $1 \le k \le N$. The variance measures the dispersion of these values taken by X around its mean value. We often write $\mathrm{var}(X) = \sigma^2$ and call σ the *standard deviation*. For a continuous distribution, the variance is given by

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} P(x)(x - \mu)^2\,dx. \tag{4}$$

Note that the variance can be written as

$$\begin{aligned} \sigma^2 &= E((X - \mu)(X - \mu)) \\ &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2. \end{aligned}$$

If c is any constant, from the definition of variance we can easily establish that

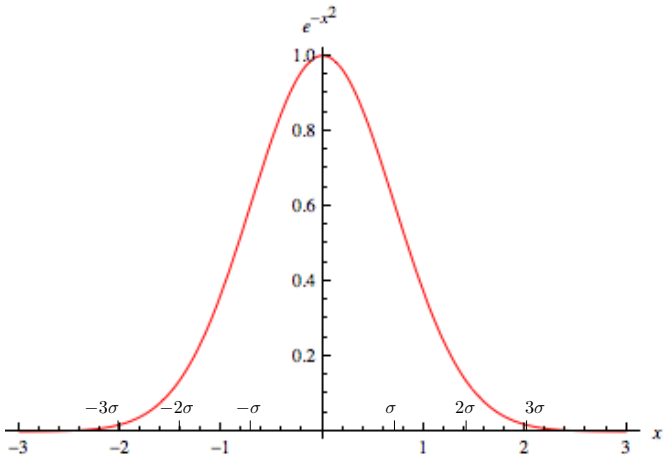
$$\begin{aligned} \mathrm{var}(X+c) &= \mathrm{var}(X), \\ \mathrm{var}(cX) &= c^2\mathrm{var}(X). \end{aligned}$$

4 Normal Distribution

A random variable X with mean μ and variance σ^2 has *Gaussian distribution* or *normal distribution* if its probability density function is given by

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{5}$$

The variable X is said to be *normal* or *Gaussian random*. This next figure[1] plots the function with $\mu = 0$ and $\sigma = \frac{\sqrt{2}}{2}$, after being scaled by $\sqrt{\pi}$.



It is easy to see from (5) that $P(x)$ has the maximum $1/(\sigma\sqrt{2\pi})$. It is possible to show that

$$E(X) = \mu \qquad \text{and} \qquad \mathrm{Var}(X) = \sigma^2,$$

namely, μ and σ are the mean and standard deviation of X, respectively. We will use the notation $N(\mu, \sigma^2)$ for a Gaussian distribution. When $\mu = 0$ and $\sigma = 1$, X is called a *standard normal* or *standard Gaussian* variable.

For the distribution (5) the following approximate values can be obtained:

$$\begin{aligned} \Pr(\mu - \sigma \le X \le \mu + \sigma) &\approx 0.6826895, \\ \Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) &\approx 0.9544997, \\ \Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) &\approx 0.9973002., \qquad (6) \\ \Pr(\mu - 4\sigma \le X \le \mu + 4\sigma) &\approx 0.9999366., \\ \Pr(\mu - 5\sigma \le X \le \mu + 5\sigma) &\approx 0.9999994.. \qquad (7) \end{aligned}$$

[1] modified over a figure from the Wolfram MathWorld at `http://mathworld.wolfram.com/GaussianFunction.html`.

The above formulae tells, for a normal distribution, the population are concentrated within intervals conveniently described in terms of multiples of the standard deviation. Among them, the "three-sigma rule" (6) has particular importance in practice because it is useful to treat a probability above 99.7% as near certainty.[2]

5 Two Random Variables

Let X and Y be random variables with distribution functions

$$G(x) = \Pr(X \leq x) \qquad \text{and} \qquad H(y) = \Pr(Y \leq y),$$

respectively. The probability for both $X \leq x$ and $Y \leq y$ is defined as the joint distribution function

$$D(x, y) = \Pr(X \leq x \wedge Y \leq y).$$

Below are some properties of this function:

$$\begin{aligned}
D(x, y) &\in [0, 1], \\
D(x, -\infty) = D(-\infty, y) &= 0, \\
D(\infty, \infty) &= 1, \\
D(a, c) &\leq D(b, d), \qquad \text{if } a \leq b \text{ and } c \leq d, \\
\Pr(a < x \leq b \wedge c < y \leq d) &= D(b, d) + D(a, c) - D(a, d) - D(b, c), \\
D(x, \infty) &= G(x), \\
D(\infty, y) &= H(y).
\end{aligned}$$

The joint probability density function is defined as the following second order partial derivative:

$$P(x, y) = \frac{\partial^2}{\partial x \partial y} D(x, y).$$

It possesses some properties which can be obtained from the definition:

$$\begin{aligned}
D(x, y) &= \int_{-\infty}^{x} \int_{-\infty}^{y} P(w, z)\, dw\, dz, \\
\Pr(a < x \leq b \wedge c < y \leq d) &= \int_{c}^{d} \int_{a}^{b} f(x, y)\, dx\, dy, \\
P_X(x) &= \int_{-\infty}^{\infty} P(x, y)\, dy, \\
P_Y(y) &= \int_{-\infty}^{\infty} P(x, y)\, dx.
\end{aligned}$$

where $P_X(x)$ and $P_Y(y)$ are the probability density functions of X and Y, respectively.

[2] Different fields use different confidence levels to consider a result to be "significant". In the social sciences, the level is of the order of a two-sigma effect (95.45%), whereas in particle physics, it is of the order of a five-sigma effect (99.99994%).

Two random variables X and Y are statistically *independent* if they satisfy the following relation:

$$\Pr(X \leq x \wedge Y \leq y) = \Pr(X \leq x) \cdot \Pr(Y \leq y), \tag{8}$$

for all $x, y \in \mathbb{R}$. Hence, the following hold for the joint distribution and probability density functions:

$$\begin{aligned} D(x,y) &= G(x)H(y), \\ P(x,y) &= P_X(x)P_Y(y). \end{aligned}$$

Let X and Y be two random variables with means μ_X and μ_Y, standard deviations σ_X and σ_Y, respectively. Then,

$$\begin{aligned} E(X+Y) &= E(X) + E(Y) \\ &= \mu_X + \mu_Y. \end{aligned} \tag{9}$$

The *covariance* of X and Y is defined as

$$\begin{aligned} \operatorname{cov}(X,Y) &= E\Big((X-\mu_X)(Y-\mu_Y)\Big) \\ &= E(XY) - \mu_X\mu_Y. \end{aligned} \tag{10}$$

When X and Y are the same variable, we see that $\operatorname{cov}(X,X) = \sigma_X^2$ is the variance of X. The *correlation coefficient* of the two variables is

$$\operatorname{cor}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{11}$$

where σ_Y is the standard deviation of Y. The correlation coefficient gives the strength of the relationship between the two random variables.

If X and Y are independent, then we have

$$\begin{aligned} E(XY) &= \iint xyP(x,y)\,dx\,dy \\ &= \iint xyP_X(x)P_Y(y)\,dx\,dy \\ &= \int xP_X(x)\,dx \int yP_Y(y)\,dy \\ &= E(X)E(Y). \end{aligned} \tag{12}$$

It then follows from (10) and (11) that

$$\operatorname{cov}(X,Y) = \operatorname{cor}(X,Y) = 0.$$

Also, the variance of their sum is the sum of individual variances:

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y). \tag{13}$$

Two random variables X and Y are *uncorrelated* if

$$\operatorname{cov}(X,Y) = 0,$$

or equivalently, if

$$E(XY) = E(X)E(Y). \tag{14}$$

Independence implies uncorrelatedness, but not necessarily vice versa.

Two variables X and Y are *orthogonal* if

$$E(XY) = 0. \tag{15}$$

If they are orthogonal, they may or may not be uncorrelated. If they are uncorrelated, then from (12) they are orthogonal only if $E(X) = 0$ or $E(Y) = 0$.

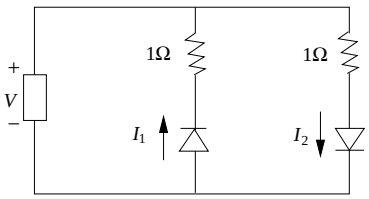
EXAMPLE 4. Let X and Y represent two throws of the dice. The two variables are independent because the outcome of one throw does not affect that of the other. We have that

$$E(X) = E(Y) = \frac{1+2+3+4+5+6}{6} = 3.5.$$

All possible outcomes of the two throws are (i, j), $1 \le i, j \le 6$. Each combination has probability $\frac{1}{36}$. The mean of XY is

$$\begin{aligned} E(XY) &= \frac{1}{36}\sum_{i=1}^{6}\sum_{j=1}^{6} ij \\ &= 12.25 \\ &= E(X)E(Y). \end{aligned}$$

Thus X and Y are uncorrelated. However, they are not orthogonal since $E(XY) \neq 0$.



EXAMPLE 5.[3] Consider the circuit in the left figure. The input voltage V is uniformly distributed on $[-1, 1]$ in volts. The two currents, in amps, are

$$\begin{aligned} I_1 &= \begin{cases} 0 & \text{if } V > 0, \\ V & \text{if } V \le 0; \end{cases} \\ I_2 &= \begin{cases} V & \text{if } V \ge 0, \\ 0 & \text{if } V < 0. \end{cases} \end{aligned}$$

So I_1 is uniformly distributed on $[-1, 0]$ and I_2 is uniformly distributed on $[0, 1]$. The expected values of the three random variables are as follows:

$$\begin{aligned} E(V) &= 0, \\ E(I_1) &= -\frac{1}{2}, \\ E(I_2) &= \frac{1}{2}. \end{aligned}$$

The random variables I_1 and I_2 are related in that exactly one of them is zero at any time instant. Since $I_1 I_2 = 0$ always holds, $E(I_1 I_2) = 0$. So the two variables are orthogonal. Because $E(I_1)E(I_2) = -\frac{1}{4} \neq E(I_1 I_2) = 0$, I_1 and I_2 are correlated.

[3] from Example 2.11 in [4, p. 65].

6 Multivariate Statistics and the Central Limit Theorem

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\top$ be a vector of n random variables such that, for $1 \leq i \leq n$,

$$\mu_i = E(X_i).$$

The *covariance matrix* is Σ whose (i,j)-th entry is as

$$\begin{aligned}(\Sigma)_{ij} &= \operatorname{cov}(X_i, X_j) \\ &= E\Big((X_i - \mu_i)(X_j - \mu_j)\Big) \\ &= E(X_i X_j) - \mu_i \mu_j.\end{aligned}$$

Denoting $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$, we can simply write the covariance matrix as

$$\Sigma = E\Big((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\Big) = E(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top.$$

The *correlation matrix* has in the (i,j)-th entry the correlation coefficient

$$\operatorname{cor}(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sigma_i \sigma_j}.$$

When the random variables are normalized, i.e., with unit standard variations, the correlation matrix is the same as the covariance matrix.

Whereas theoreticians are primarily interested in the covariance matrix, practitioners prefer the correlation matrix, because a correlation coefficient is more intuitive than a covariance. Other than that, both matrices have essentially the same properties

An n-element $\boldsymbol{X}$ is Gaussian (or normal) if its probability density function is given by

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})\right).$$

Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and have identical probability density functions. Then,

$$\begin{aligned}E(X_1) = E(X_2) = \cdots = E(X_n) = \mu, \\ \operatorname{Var}(X_1) = \operatorname{Var}(X_2) = \cdots = \operatorname{Var}(X_n) = \sigma^2.\end{aligned}$$

Denote by S_N the sum of these variables, namely, $S_N = X_1 + X_2 + \cdots + X_N$. From (9) and (13) it follows that

$$\begin{aligned}E(S_N) &= N\mu, \\ \operatorname{Var}(S_N) &= N\sigma^2.\end{aligned}$$

Now, consider a random variable Z_N following a normal distribution with the same mean $N\mu$ and variance $N\sigma^2$.

Theorem 1 (Central Limit Theorem) *The probability density of the sum S_N approaches that of the normal variable Z_N in the way that, for every x and all large enough N,*

$$\Pr\left(\frac{S_N - N\mu}{\sigma\sqrt{N}} < x\right) \approx \Pr\left(\frac{Z_N - N\mu}{\sigma\sqrt{N}} < x\right).$$

The sum of a large number of identical random variables is approximately normal.

Intuitively, the central limit theorem says that the distribution of data influenced by many small and unrelated random effects is approximately normal. It is valid even when not all of $X_1, X_2, \ldots, X_n$ are identical or independent, as long as single variables do not play a dominating role in the sum. Whenever an influence has to sum up a large number of independent random factors, it can be viewed as following some normal distribution.

The central limit theorem explains why normal random variables are often encountered in nature. A random variable in nature often appears to follow a Gaussian distribution because it is the sum of many individual and independent random variables. For instance, the high temperature on any given day in any given location is affected by clouds, precipitation, wind, air pressure, humidity, etc. It has a Gaussian probability density function.

7 Stochastic Process

A *stochastic process* $X(t)$ is a random variable that changes with time. Time may be continuous or discrete. The value of the random variable may be continuous at every time instant or discrete at every time instant. So a stochastic process can be one of four types.

The distribution and density functions of a stochastic process are functions of time:

$$\begin{aligned} D(x,t) &= \Pr(X(t) \le x), \\ P(x,t) &= \frac{d}{dx} D(x,t). \end{aligned}$$

For a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, these functions are defined as

$$\begin{aligned} D(\boldsymbol{x},t) &= \Pr(X_1(t) \le x_1 \wedge \cdots \wedge X_n(t) \le x_n), \\ P(\boldsymbol{x},t) &= \frac{\partial^n}{\partial x_1 \cdots \partial x_n} D(\boldsymbol{x},t). \end{aligned}$$

The mean and covariance of a stochastic process $X(t)$, originally defined in (3) and (4) for a time-independent random variable, respectively, are also functions of time. The integrations are carried out over the values of the random variable.

A stochastic process $X(t)$ at two different times t_1 and t_2 comprises two different random variables $X(t_1)$ and $X(t_2)$. So, we can talk about their joint distribution and joint density functions, defined as follows:

$$\begin{aligned} D(x_1, x_2, t_1, t_2) &= \Pr(X(t_1) \le x_1 \wedge X(t_2) \le x_2), \\ P(x_1, x_2, t_1, t_2) &= \frac{\partial^2}{\partial x_1 \partial x_2} D(x_1, x_2, t_1, t_2). \end{aligned}$$

A stochastic process is called *stationary* if its probability density does not change with time.

EXAMPLE 6.[4] Tomorrow's closing price of the Dow Jones Industrial Average might be a random variable with a certain mean and variance. However, a century ago the mean was much lower. The closing price is a random variable with generally increasing mean with time. It is not stationary.

[4] Example 2.12 in [4, p. 70].

8 Noise Simulation

If the random variables $X(t_1)$ and $X(t_2)$ are independent for all $t_1 \neq t_2$, then the stochastic process $X(t)$ is called *white noise*. Otherwise, it is called *colored noise*. In optimal filtering research and experiments, we often have to simulate correlated white noise. Phrased in the discrete sense, we need to create random vectors whose elements are correlated with each other according to some predefined covariance matrix.

Suppose we want to generate an n-element random vector $\boldsymbol{X}$ which has zero mean and covariance matrix:

$$\Sigma = E((\boldsymbol{X} - \boldsymbol{0})(\boldsymbol{X} - \boldsymbol{0})^\top) = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1n}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{1n}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix},$$

where $\sigma_{ii} = \sigma_i$ is the standard deviation of X_i, for $1 \leq i \leq n$. As a covariance matrix, Σ must be positive semi-definite. Its eigenvalues are real and nonnegative, and denoted as $\mu_1^2, \ldots, \mu_n^2$. Let $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n$ be the corresponding unit eigenvectors which can be chosen orthogonal since Σ is symmetric. By the Spectral Theorem [5, p. 273] in linear algebra, the matrix Σ has a decomposition

$$\Sigma = Q \Lambda Q^\top,$$

where $Q = (\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n)$ is orthogonal, and

$$\Lambda = \mathrm{diag}(\mu_1^2, \ldots, \mu_n^2).$$

We introduce a new random vector $\boldsymbol{Y} = Q^{-1}\boldsymbol{X} = Q^\top \boldsymbol{X}$ so that $\boldsymbol{X} = Q\boldsymbol{Y}$. Therefore,

$$\begin{aligned} E(\boldsymbol{Y}\boldsymbol{Y}^\top) &= E(Q^\top \boldsymbol{X}\boldsymbol{X}^\top Q) \\ &= Q^\top E(\boldsymbol{X}\boldsymbol{X}^\top) Q \\ &= Q^\top \Sigma Q \\ &= \Lambda. \end{aligned}$$

The covariance matrix of $\boldsymbol{X}$ is as given:

$$\begin{aligned} E(\boldsymbol{X}\boldsymbol{X}^\top) &= E((Q\boldsymbol{Y})(Q\boldsymbol{Y})^\top) \\ &= E(Q\boldsymbol{Y}\boldsymbol{Y}^\top Q^\top) \\ &= Q E(\boldsymbol{Y}\boldsymbol{Y}^\top) Q^\top \\ &= Q \Lambda Q^\top \\ &= \Sigma. \end{aligned}$$

The following steps summarize the algorithm of generating $\boldsymbol{X}$ with zero mean and covariance matrix Σ.

1. Compute the eigenvalues $\mu_1^2, \ldots, \mu_n^2$ of Σ.
2. Find the corresponding unit eigenvectors $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n$ and let $Q = (\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n)$.
3. For $i = 1, \ldots, n$, compute the random variable $Y_i = \mu_i Z_i$, where Z_i is an independent random number with zero mean and unit variance. You may assume Gaussian distribution given in (5) with $\mu = 0$ and $\sigma = 1$, and then apply the procedure from Section 2.
4. Let $\boldsymbol{X} = Q(Y_1, \ldots, Y_n)^\top$.

We refer to [2, pp. 391–448] for eigenvalue computation for symmetric matrices.

References

[1] W. Feller. The fundamental limit theorems in probability. *Bulletin of the Americal Mathematical Society*, 51:800–832, 1945.

[2] G. H. Golub and C. F. Van Loan. *Matrix Computations*, 3rd edition. The Johns Hopkins University Press, Baltimore, Maryland, 1996.

[3] D. E. Knuth. *Seminumerical Algorithms*, vol. 2 of *The Art of Computer Programming*, 3rd edition. Addison-Wesley, Reading, Massachusetts, 1997.

[4] D. Simon. *Optimal State Estimations*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[5] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, Massachusetts, 1993.

[6] Wolfram MathWorld. `http://mathworld.wolfram.com/`