Lecture 2
Random Variables and Stochastic Processes

- Probability theory
- Random Variables
- Stochastic processes theory
- Several Kinds of Stochastic Processes

Contents

- Probability theory
- Random Variables
- Stochastic processes theory
- Several Kinds of Stochastic Processes

In our attempts to filter a signal, we will be trying to extract meaningful information from a noisy signal. In order to accomplish this, we need to know something about what the noise is, some of its characteristics, and how it works.

Probability

- The probability of event A (see refs for formal definition)

$$P(A) = \frac{\text{Number of times } A \text{ occurs}}{\text{Total number of outcomes}}$$

- Example: what is the probability of getting the number 1 four times when rolling a six-sided die 6 times?)

$$P(A) = \frac{C_6^4 \cdot 5 \cdot 5}{6^6} = 0.0080$$

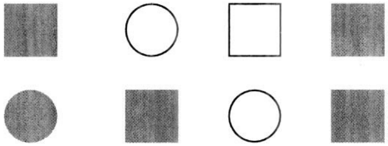
Probability

- The conditional probability of event A given event B: $(P(B) \neq 0)$

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

  - $P(A|B)$ is the conditional probability of A given B, i.e, the probability that A occurs given the fact that B occurred
  - $P(A,B)$ is the joint probability of A and B, i.e., the probability that event A and B both occur
  - $P(A)$ or $P(B)$ is called an *a priori* probability as it applies to the probability of an event apart from any previously known information
  - The conditional probability is called an *a posteriori* probability as it applies to a probability given the fact that some information about a possibly related event is already known

Example



$P(\text{circle}) = 3/8, P(\text{square}) = 5/8;$

$P(\text{gray, circle}) = 1/8, P(\text{gray}|\text{circle}) = 1/3;$

$$P(\text{white}|\text{square}) = \frac{1/8}{5/8} = 1/5.$$

Bayers' Rule

- $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ (statement of theorem)
- $P(\text{gray}|\text{circle}) = \frac{P(\text{circle}|\text{gray})P(\text{gray})}{P(\text{circle})} = \frac{(1/5)(5/8)}{3/8} = 1/3$

Independence

- We say that two events are independent if the occurrence of one event has no effect on the probability of the occurrence of the other event.
  - $P(A, B) = P(A)P(B)$
  - $P(A|B) = P(A)$
  - $P(B|A) = P(B)$

Contents

Random variables

- RV (random variable): a functional mapping from a set of experimental outcomes (the domain) to a set of real numbers (the range)
- the outcome of a particular experiment is not a RV
- the RV X exists independently of any of its realizations
- the RV X will always be random and will never be equal to a specific value

Random variables

- A RV can be either continuous or discrete (realizations belong to a discrete or continuous set of values)
- Probability distribution function (PDF):

$$F_X(x) = P(X \le x)$$

Properties:

- $F_X(x)$ is the PDF of the RV X
- x is a nonrandom independent variable or constant
- $F_X(x) \in [0, 1], F_X(-\infty) = 0, F_X(\infty) = 1$
- $F_X(a) \le F_X(b)$ if $a \le b$
- $P(a < X \le b) = F_X(b) - F_X(a)$

Probability density function (pdf)

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Properties:

- $F_X(x) = \int_{-\infty}^{x} f_X(z)dz$
- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- $P(a < x \leq b) = \int_a^b f_X(x)dx$

Example: uniformly-distributed RV

- Take a measurement with the set S of outcomes equal to any number between -1 and 1;
- Define the RV Z by $Z(\alpha) = \alpha$;
- the distribution function is given by

$$F_Z(z) = \begin{cases} 0, & z < -1 \\ 0.5(z+1), & -1 < z < 1; \\ 1, & z > 1; \end{cases}$$

- the density function is

$$f_Z(z) = \begin{cases} 0.5, & -1 < z < 1 \\ 0, & \text{otherwise.} \end{cases}$$

- The RV Z is a uniformly distributed continuous RV.

Example: Gaussian RV

- Take a measurement with the set S of outcomes equal to any number between -1 and 1;
- Define the RV Z by $Z(\alpha) = \alpha$;
- Assume the density function is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z-\eta)^2}{2\sigma^2}},$$

  where η is a real number and σ is a positive number;
- The RV Z is a Gaussian or normal RV, denoted as $Z \sim \mathcal{N}(\eta, \sigma^2)$.

Expected value

- The expected value (expectation, mean, average) of a RV X is defined as its average value over a large number of experiments.
- $E(X) = \lim\limits_{N \to \infty} \frac{1}{N} \sum\limits_{i=1}^{m} A_i n_i$
  - the outcome A_i occurs n_i times
- The expected value of any function $g(X)$:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Variance

- The variance of a RV is a measure of how much we expect the RV to vary from its mean.
- The variance is a measure of how much variability there is in a RV.
- $\sigma_X^2 = E[(X - EX)^2] = \int_{-\infty}^{\infty}(x - EX)^2 f_X(x)dx$, $\sigma_X^2 = E(X^2) - (EX)^2$.
- Standard deviation $\sigma(\sigma_X)$

Transformations of random variables

Suppose that we have two RVs, X and Y, related to one another by the monotonic functions $g(\cdot)$ and $h(\cdot)$:

$$Y = g(X)$$

$$X = g^{-1}(Y) = h(Y)$$

If we know the pdf of X, then we can compute the pdf of Y as follows:

$$P(X \in [x, x+dx]) = P(Y \in [y, y+dy])(dx > 0)$$

$$\int_x^{x+dx} f_X(z)dz = \begin{cases} \int_y^{y+dy} f_Y(z)dz & \text{if dy>0} \\ -\int_y^{y+dy} f_Y(z)dz & \text{if dy<0} \end{cases}$$

$$f_X(x)dx = f_Y(y)|dy|$$

$$f_Y(y) = |\frac{dx}{dy}| f_X[h(y)] = |h'(y)| f_X[h(y)]$$

Example: find the pdf of a linear function of a Gaussian RV

Suppose $X \sim N(\bar{x}, \sigma_x^2)$ and $Y = g(X) = aX + b, a, b \in \mathbb{R}$, Solve $f_Y(y)$.

$$\begin{aligned}
X &= h(Y) \\
&= (Y-b)/a \\
h'(y) &= 1/a \\
f_Y(y) &= |h'(y)| f_X[h(y)] \\
&= |\tfrac{1}{a}| \tfrac{1}{\sigma_X\sqrt{2\pi}} \exp\left\{\tfrac{-[(y-b)/a-\bar{x}]^2}{2\sigma_X^2}\right\} \\
&= \tfrac{1}{|a|\sigma_X\sqrt{2\pi}} \exp\left\{\tfrac{-[y-(a\bar{x}+b)]^2}{2a^2\sigma_X^2}\right\}
\end{aligned}$$

i.e., $Y \sim \mathcal{N}(a\bar{x}+b, a^2\sigma_x^2)$.

Multiple random variables

Joint distribution function

- $F_{XY}(x, y) = P(X \leq x, Y \leq y)$ $(F(x, y))$
- $F(x, y) \in [0, 1]$, $F(x, -\infty) = F(-\infty, y) = 0$, $F(\infty, \infty) = 1$
- $F(a, c) \leq F(b, d)$ if $a \leq b$ and $c \leq d$
- $P(a < X \leq b, c < Y \leq d) = F(b, d) + F(a, c) - F(a, d) - F(b, c)$
- $F(x, \infty) = F(x)$, $F(\infty, y) = F(y)$ (marginal distribution function)

Joint probability density function

- $f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$ $(f(x,y))$
- $F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(z_1, z_2) dz_1 dz_2$
- $f(x,y) \geq 0$, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1$
- $P(a < X \leq b, c < Y \leq d) = \int_{c}^{d} \int_{a}^{b} f(x,y) dx dy$
- $f(x) = \int_{-\infty}^{\infty} f(x,y) dy$, $f(y) = \int_{-\infty}^{\infty} f(x,y) dx$ (marginal density function)

Mixed moments

- Expectation of functions of X and Y:

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

- Covariance of two scalar RVs X and Y: $C_{XY} = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$

Statistical independence

- The RVs X and Y are independent if they satisfy the following equality

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y), \ \forall x, y$$

- $F_{XY}(x,y) = F_X(x)F_Y(y)$, $f_{XY}(x,y) = f_X(x)f_Y(y)$

Statistical Uncorrelatedness

- Correlation coefficient of two scalar RVs X and Y: $\rho = \frac{C_{XY}}{\sigma_x \sigma_y}$.
- Correlation of two scalar RVs X and Y is defined as $R_{XY} = E(XY)$.
- The RVs X and Y are uncorrelated if

$$\rho = 0 \text{ or } R_{XY} = E(X)E(Y).$$

- Independent $\subsetneq$ Uncorrelated

Uncorrelatedness VS independence

- Two RVs X and Y have either a relationship or they don't have a relationship at all
- Now if there is a relationship, it's either linear or non-linear

Assume $Y = aX + b$, we have

$$E(XY) = aEX^2 + bEX$$

On the other hand,

$$EXEY = a(EX)^2 + bEX$$

Except for the case $EX^2 = (EX)^2$, i.e., $DX = 0$, we have

$$E(XY) \neq EXEY$$

Cases when there are no linear relationship between 2 RVs

Assume (X, Y) conforms to the uniform distribution on the boundary of a unit circle, and they satisfy,

$$X^2 + Y^2 = 1,$$

then we have $f(x, y) = \frac{1}{\pi}, \forall x \in [-1, 1], y = \pm\sqrt{1 - x^2}$.

We further have

$$f_X(x) = \frac{2\sqrt{1-x^2}}{\pi}, \forall x \in [-1, 1]$$
$$f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi}, \forall y \in [-1, 1]$$

Thus

$$E(XY) = 0, EX = 0, EY = 0$$

Statistical Orthogonality

- Two RVs are said to be orthogonal if $R_{XY} = 0$
- Two uncorrelated RVs are orthogonal only if at least one of them is zero-mean

Example

A slot machine is rigged so you get -1, 0, or 1 with equal probability for the first spin X. On the second spin Y you get 1 if $X = 0$, and 0 if $X \neq 0$.

$$\begin{aligned} E(X) &= \tfrac{-1+0+1}{3} = 0 \\ E(Y) &= \tfrac{0+1+0}{3} = 1/3 \\ E(XY) &= \tfrac{(-1)(0)+(0)(1)+(1)(0)}{3} = 0 \end{aligned}$$

- X and Y are uncorrelated because $E(XY) = E(X)E(Y)$
- X and Y are orthogonal because $E(XY) = 0$
- The two RVs are dependent because the realization of Y depends on the realization of X.

Conditional Density Functions

- Let X and Y be jointly distributed RVs;
- Define the conditional distribution function $F_Y(y|x_1 < X \leq x_2)$ as the conditional probability of the event $\{Y \leq y\}$ given that the event $\{x_1 < X \leq x_2\}$ occurred, i.e.,

$$F_Y(y|x_1 < X \leq x_2) = P(Y \leq y|x_1 < X \leq x_2);$$

- Define the conditional density function $f_Y(y|X = x)$ as

$$f_Y(y|X = x) = \lim_{\Delta x \to 0} f_Y(y|x < X \leq x + \Delta x);$$

- We have

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

Multivariate statistics

Given an n-element RV X and an m-element RV Y (assuming that both X and Y are column vectors), their correlation is defined as

$$
\begin{aligned}
R_{XY} &= E(XY^T) \\
&= \begin{bmatrix} E(X_1Y_1) & \cdots & E(X_1Y_m) \\ \vdots & & \vdots \\ E(X_nY_1) & \cdots & E(X_nY_m) \end{bmatrix}
\end{aligned}
$$

Their covariance is defined as

$$
\begin{aligned}
C_{XY} &= E[(X - E(X))(Y - E(Y))^T] \\
&= E(XY^T) - E(X)E(Y)^T
\end{aligned}
$$

The autocorrelation of the n-element RV X is defined as

$$
\begin{aligned}
R_X &= E[XX^T] \\
&= \begin{bmatrix} E(X_1^2) & \cdots & E(X_1X_n) \\ \vdots & & \vdots \\ E(X_nX_1) & \cdots & E(X_n^2) \end{bmatrix}
\end{aligned}
$$

We have $R_X = R_X^T$, i.e., an autocorrelation matrix is always symmetric.

Besides, an autocorrelation matrix is always positive semidefinite.

$$z^T R_X z = z^T E[XX^T] z = E[z^T X X^T z] = E[(z^T X)^2] \geq 0$$

The autocovariance of n-element RV X is defined as

$$\begin{aligned}
C_X &= E[(X - E(X))(X - E(X)^T)] \\
&= \begin{bmatrix} E(X_1 - E(X_1))^2 & \cdots & E[(X_1 - E(X_1))(X_n - E(X_n))] \\ \vdots & & \vdots \\ E[(X_n - E(X_n))(X_1 - E(X_1))] & \cdots & E[(X_n - E(X_n))^2] \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix}
\end{aligned}$$

An auto covariance matrix is always symmetric and positive semidefinite.

$$z^T C_X z = z^T E[(X - \bar{X})(X - \bar{X})^T] z = E[(z^T (X - \bar{X}))^2] \geq 0$$

Linear transformation of Gaussian RV

- An n-element RV X is Gaussian (normal) if

$$\text{pdf}(X) = \frac{1}{(2\pi)^{n/2}|\det(C_X)|^{1/2}} \exp\left[-\frac{1}{2}(x - E(X))^T C_X^{-1}(x - E(X))\right]$$

- Consider a Gaussian RV X that undergoes a linear transformation $Y = g(X) = AX + b$, where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$.

- If A is invertible, we have

$$\begin{aligned} f_Y(y) &= |h'(y)| f_X[h(y)] \\ &= \tfrac{1}{(2\pi)^{n/2}|\det(AC_XA^T)|^{1/2}} \exp\left[-\tfrac{1}{2}(y - E(Y))^T (AC_XA^T)^{-1}(y - E(Y))\right], \end{aligned}$$

i.e., $Y \sim \mathcal{N}(AE(X)+b, AC_XA^T)$. The normality is preserved in linear transformations of random vectors (just as in scalar case).

Matrix derivative

Usually, for vector derivative, the vector is defined as a column vector.

For $f(x) : \mathbb{R}^n \to \mathbb{R}$, the Jacobian of $f(x)$ is an $n \times 1$ vector and the Hessian of $f(x)$ is an $n \times n$ matrix.

$$\nabla_x f = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \nabla_x^2 f = \frac{\partial^2 f}{\partial x \partial x^T} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix}$$

Matrix derivative

Vector by vector derivative, $f(x) : \mathbb{R}^n \to \mathbb{R}^m (m > 1)$, where $f = [f_1, \ldots, f_m]^T$, $x = [x_1, \ldots, x_n]^T$, the Jacobian matrix,

$$\nabla_x f = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Matrix derivative

Scalar by matrix derivative, $f(X) : \mathbb{R}^{n \times m} \to \mathbb{R}$, where $n, m > 1$ and

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

we have

$$\nabla_x f = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1m}} \\ \frac{\partial f}{\partial x_{21}} & \cdots & \frac{\partial f}{\partial x_{2m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \cdots & \frac{\partial f}{\partial x_{nm}} \end{bmatrix}$$

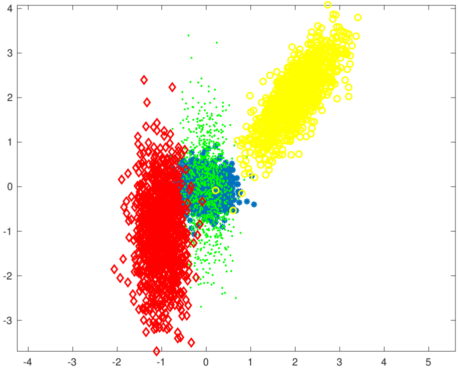
Properties of the determinant

- $\det\left(I_n\right) = 1$ where I_n is the $n \times n$ identity matrix.
- $\det\left(A^T\right) = \det(A)$,
- $\det\left(A^{-1}\right) = \dfrac{1}{\det(A)} = \det(A)^{-1}$
- For square matrices A and B of equal size, $\det(AB) = \det(A)\det(B)$.
- $\det(cA) = c^n \det(A)$ for an $n \times n$ matrix A.

Linear transformation of Gaussian RV

$$
\begin{aligned}
f_Y(y) &= |h'(y)| f_X[h(y)] \\
&= |\det(A^{-1})| f_X[h(y)] \\
&= |\det(A^{-1})| \frac{1}{(2\pi)^{n/2} |\det(C_X)|^{1/2}} \cdot \\
&\quad \exp\left\{ -\frac{1}{2} \left[A^{-1}(y-b) - E(X) \right]^T C_X^{-1} [*] \right\} \\
&= |\det(A^{-1})| \frac{1}{(2\pi)^{n/2} |\det(C_X)|^{1/2}} \cdot \\
&\quad \exp\left\{ -\frac{1}{2} \left[A^{-1}y - A^{-1}b - \bar{x} \right]^T C_X^{-1} [*] \right\} \\
&= \frac{1}{(2\pi)^{n/2} |\det(A)| |\det(C_X)|^{1/2}} \cdot \\
&\quad \exp\left\{ -\frac{1}{2} \left[A^{-1}y - A^{-1}b - A^{-1}\bar{y} + A^{-1}b \right]^T C_X^{-1} [*] \right\} \\
&= \frac{1}{(2\pi)^{n/2} |\det(A)|^{1/2} |\det(C_X)|^{1/2} |\det(A^T)|^{1/2}} \exp\left[-\frac{1}{2} (y-\bar{y})^T (A^{-1})^T C_X^{-1} A^{-1} (y-\bar{y}) \right] \\
&= \frac{1}{(2\pi)^{n/2} |\det(A C_X A^T)|^{1/2}} \exp\left[-\frac{1}{2} (y-\bar{y})^T (A C_X A^T)^{-1} (y-\bar{y}) \right]
\end{aligned}
$$

Linear transformation of Gaussian RV: Understanding the covariance



Points after linear transformation. The blue denotes the original points conforming to normal distribution, $C_X = \mathrm{diag}(0.3^2, 0.3^2)$, the green points $A = \begin{bmatrix} 0 & 1 \\ 3.1623 & 0 \end{bmatrix}$, $b = 0$, the red $A = \begin{bmatrix} 0 & 1 \\ 3.1623 & 0 \end{bmatrix}$, $b = [-1, -1]^T$, the yellow $A = \begin{bmatrix} -1.5648 & -0.7425 \\ -2.1711 & 0.5352 \end{bmatrix}$, $b = [2, 2]^T$

Ellipsoid

- If v is a point and A is a real, symmetric, positive-definite matrix, then the set of points $\mathbf{x}$ that satisfy the equation

$$(\mathbf{x} - \mathbf{v})^T A(\mathbf{x} - \mathbf{v}) = 1$$

  is an ellipsoid centered at v.

- The eigenvectors of A are the principal axes of the ellipsoid, and the eigenvalues of A are the reciprocals of the squares of the semi-axes: a^{-2}, b^{-2} and c^{-2}.

- An invertible linear transformation applied to a sphere produces an ellipsoid.

- If the linear transformation is represented by a symmetric 3×3 matrix, then the eigenvectors of the matrix are orthogonal and represent the directions of the axes of the ellipsoid

Eigen decomposition of the covariance matrix

- An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it. It can be expressed as

$$Av = \lambda v$$

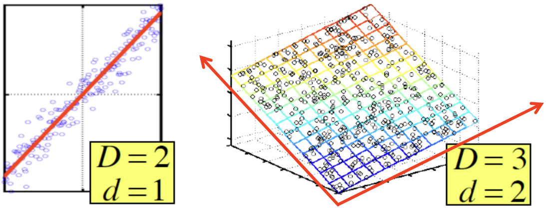
- For a covariance matrix Σ, assume the SVD decomposition is,

$$\Sigma = U \Lambda U^{-1}$$

  then we have $\Sigma U = U \lambda$, meaning that U and Λ represents the eigenvectors and eigenvalues of Σ, respectively.

- The eigenvectors are unit vectors representing the direction of the largest variance of the data, while the eigenvalues represent the magnitude of this variance in the corresponding directions.

Principle component analysis



- In case where data lies on or near a low d--dimensional linear subspace, axes of this subspace are an effective representation of the data.
- Identifying the axes is known as Principal Components Analysis, and can be obtained by using classic matrix computation tools (Eigen or Singular Value Decomposition).

PCA algorithm

- Given data $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, compute the covariance matrix Σ,

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

  where $\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}_i$.
- PCA basis vectors = the eigenvectors of Σ
- Larger eigenvalue $\Rightarrow$ more important eigenvectors

PCA: eigenvalue & eigenvector

- For symmetric matrices, eigenvectors for distinct eigenvalues are orthogonal,

$$\Sigma v_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1^T v_2 = 0.$$

- All eigenvalues of a real symmetric matrix are real.
- All eigenvalues of a positive semidefinite matrix are nonnegative.

Let $z = v_1^T \Sigma v_2$, as z is a scalar, we have $z^T = z$, i.e.,

$$v_2^T \Sigma^T v_1 = v_1^T \Sigma v_2$$

As Σ is symmetric, we then have

$$v_2^T \Sigma v_1 = v_2^T \lambda_1 v_1 = v_1^T \lambda_2 v_2^T$$

that is

$$\lambda_1 v_1^T v_2 = \lambda_2 v_1^T v_2$$

as $\lambda_1 \neq \lambda_2$, we have

$$v_1^T v_2 = 0.$$

PCA algorithm

- Eigenvalue decomposition:

$$\Sigma = U\Lambda U^{-1}$$

  - Columns of U are eigenvectors of Σ
  - Diagonal elements of Λ are eigenvalues of Σ

$$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m), \lambda_i \geq \lambda_{i+1}.$$

- Select

$$U_k = [u_1, \dots, u_k], \Lambda_k = \mathrm{diag}(\lambda_1, \dots, \lambda_k),$$

  Let

$$\mathbf{z}_i = U_k^T \mathbf{x}_i$$

- PCA learns the above linear transformation and construct the dataset

$$Z = \{\mathbf{z}_1, \dots, \mathbf{z}_m\}.$$

  with $\mathrm{cov}(Z, Z) = \Lambda_k$. (dimensionality reduction)

Contents

A stochastic process, also called a random process, is a very simple generalization of the concept of a RV. A stochastic process $X(t)$ is a RV X that changes with time.

- continuous random process: the RV at each time is continuous and time is continuous (the temperature at each moment of the day)
- discrete random process: the RV at each time is discrete and time is continuous (the number of people in a given building at each moment of the day)
- continuous random sequence: the RV at each time is continuous and time is discrete (the high temperature each day)
- discrete random sequence: the RV at each time is discrete and time is discrete (the number of people in a given building each day)

Distribution and density

Since a stochastic process is a RV that changes with time, it has a distribution and density function that are functions of time.

- The PDF of $X(t)$ is $F_X(x,t) = P(X(t) \leq x)$ (If $X(t)$ is a random vector, then the inequality above is an element-by-element inequality, i.e., $F_X(x,t) = P[X_1(t) \leq x_1, \cdots, X_n(t) \leq x_n]$)
- The pdf of $X(t)$ is $f_X(x,t) = \frac{dF_X(x,t)}{dx}$ (If $X(t)$ is a random vector, then the derivative is taken once with respect to each element of x, i.e., $f_X(x,t) = \frac{\partial^n F_X(x,t)}{\partial x_1 \cdots \partial x_n}$)

Mean and covariance (over x)

The mean and covariance of $X(t)$ are also functions of time:

- Mean: $\bar{x}(t) = \int_{-\infty}^{\infty} x f(x,t) dx$ (changes with time)
- Covariance: $C_X(t) = E\{[X(t) - \bar{x}(t)][X(t) - \bar{x}(t)]^T\} = \int_{-\infty}^{\infty} [x - \bar{x}(t)][x - \bar{x}(t)]^T f(x,t) dx$ (changes with time)

Stochastic process at two different times

Different random variables: $X(t_1)$ and $X(t_2)$

- joint distribution (second-order distribution) function:

$$F(x_1, x_2, t_1, t_2) = P(X(t_1) \leq x_1, X(t_2) \leq x_2)$$

- joint density (second-order density) function:

$$f(x_1, x_2, t_1, t_2) = \frac{\partial^2 F(x_1, x_2, t_1, t_2)}{\partial x_1 \partial x_2}$$

If $X(t)$ is an n-element random vector, then the inequality that defines $F(x_1, x_2, t_1, t_2)$ actually consists of $2n$ inequalities, and the derivative that defines $f(x_1, x_2, t_1, t_2)$ actually consists of $2n$ derivatives.

Autocorrelation and Autocovariance

- Autocorrelation of the stochastic process $X(t)$: the correlation between the two RVs $X(t_1)$ and $X(t_2)$

$$R_X(t_1, t_2) = E[X(t_1)X^T(t_2)]$$

- Autocovariance of a stochastic process:

$$C_X(t_1, t_2) = E\{[X(t_1) - \bar{x}(t_1)][X(t_2) - \bar{x}(t_2)]^T\}$$

Stationary stochastic process

- Strict-sense stationary: the stochastic process $\{X(t)\}$ is said to be strictly stationary, strongly stationary or strict-sense stationary if

$$F_X(x(t_1+\tau),\ldots,x(t_n+\tau)) = F_X(x(t_1),\ldots,x(t_n))$$

$$\text{for all } \tau, t_1, \ldots, t_n \in \mathbb{R} \text{ and for all } n \in \mathbb{N}$$

  e.g., flipping a coin ten times.

- Wide-sense stationary: the mean of the stochastic process is constant with respect to time, and the autocorrelation is a function of the time difference $t_2 - t_1$ (not a function of the absolute times):

$$E[X(t)] = \bar{x}, \quad E[X(t_1)X^T(t_2)] = R_X(t_2 - t_1)$$

- Stationary implies wide-sense stationary; wide-sense stationary does not implies stationary

Examples of stationary and non stationary stochastic process

- The high temperature each day. Not stationary.
- Electrical noise. If the statistics of the noise are the same every day, then the electrical noise is a stationary process. For practical purposes, if the statistics of a random process do not change over the time interval of interest, then we consider the process to be stationary.
- tomorrow's closing price of the Dow Jones Industrial Average. Nonstationary stochastic process.
- More examples?

Properties of wide-sense stationary stochastic process

- $R_X(0) = E[X(t)X^T(t)]$
- $R_X(-\tau) = R_X^T(\tau)$
- For scalar stochastic processes, we have $|R_X(\tau)| \leq R_X(0)$

Time average and autocorrelation

Suppose that the process has a realization $x(t)$. For continuous-time random processes, we define:

- Time average (sample average):

$$A[X(t)] = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) dt$$

- Time autocorrelation:

$$R[X(t), \tau] = A[X(t)X^T(t+\tau)]$$

Ergodic process

An ergodic process is a stationary random process for which

$$A[X(t)] = E(X)$$

$$R[X(t), \tau] = R_X(\tau)$$

In the real world, we are often limited to only a few realizations of a stochastic process. We can compute the time average, time autocorrelation, and other time-based statistics of the realization. If the random process is ergodic, then we can use those time averages to estimate the statistics of the stochastic process.

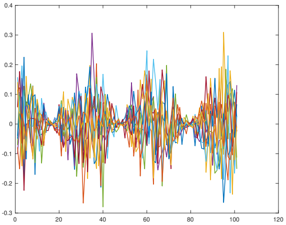
Example: Waves coming up on a beach

- If you look from side-to-side, you get an idea of the distribution of heights at different spots at any one time
- If you measure at one spot, you get an idea of the distribution of heights at one spot over time.
- assume the process is ergodic, you would look up and down at a specific spot of the beach and infer the time series behavior of waves
- You will fail if the waves are not ergodic over the relevant time scale (we can assume a time scale for the ergodicity to be valid)

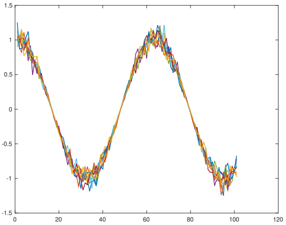
Example

Suppose X is a random variable, and $Y(t) = X \cos t$ is a stochastic process.

1. Find the expected value of $Y(t)$.
2. Find $A[Y(t)]$, the time average of $Y(t)$.
3. Under what condition is $E[Y(t)] = A[Y(t)]$?



(a) Plot of $y(t)$ when $EX = 0$



(b) Plot of $y(t)$ when $EX = 1$

Two stochastic processes

- The cross correlation of $X(t)$ and $Y(t)$:

$$R_{XY}(t_1, t_2) = E[X(t_1)Y^T(t_2)]$$

- Two random processes $X(t)$ and $Y(t)$ are said to be uncorrelated if $R_{XY}(t_1, t_2) = E[X(t_1)]E[Y(t_2)]^T$ for all t_1 and t_2.

- The cross covariance of $X(t)$ and $Y(t)$ is defined as

$$C_{XY}(t_1, t_2) = E\{[X(t_1) - \bar{X}(t_1)][Y(t_2) - \bar{Y}(t_2)]^T\}$$

Contents

Markov model

- In probability theory, a Markov model is a stochastic model used to model randomly changing systems
- It is assumed that future states depend only on the current state, not on the events that occurred before it (that is, it assumes the *Markov property*)

Markov models

	System state is fully observable	System state is partially observable
System is autonomous	Markov chain	Hidden Markov model
System is controlled	Markov decision process	Partially observable Markov decision process

Markov Chain

For a discrete random sequence, the outcome of the n-th trial is the random variable X_n, X_0 is the initial position of the process.

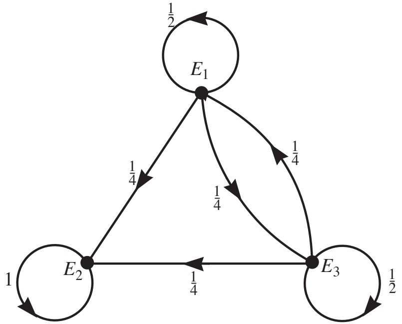
The discrete random sequence is called a **Markov Chain**, if we have

$$P\{X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}$$
$$= P\{X_{n+1} = i_{n+1} | X_n = i_n\}$$

for all $n \in \mathbb{N}_0$, $i_0, \ldots, i_n, i_{n+1} \in S$ (S is the state set).

Markov property: the memoryless property of a stochastic process.

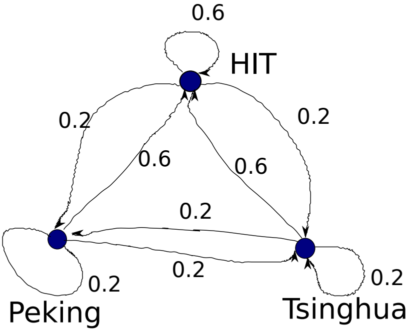
Illustration



Probability transition matrix:

$$\begin{array}{cc} & \text{To} \\ \text{From} \quad T = & \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & 1 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \end{array}$$

Example: Where shall we go for lunch?



$$\begin{array}{cc} & \text{To} \\ \text{From} \quad T = & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.6 & 0.2 & 0.2 \\ 0.6 & 0.2 & 0.2 \end{bmatrix} \end{array}$$

Example: Where shall we go for lunch?

Predict the preference for the restaurant:

$$x_0 = [1\ 0\ 0], X_n = ?$$

Steady state of preference for the restaurant?

$$q = \lim_{n \to \infty} X_n$$

What will happen if we change the transition matrix T?

$$x_1 = x_0 \cdot T,$$

each element indicates the corresponding probability

Ergodic Markov Chain

- A Markov chain is called an ergodic chain if it is possible to go from every state to every state (not necessarily in one move).
- A transition matrix is regular where there is power of T that contains all positive no zeros entries.
- Any transition matrix that has no zeros determines a regular Markov chain. It is possible for a regular Markov chain to have a transition matrix that has zeros.
- Every regular chain is ergodic.
- Is it stationary? (the Markov chain stationary with stationary distribution π if $\pi = \pi \cdot T$) If a Markov chain is regular , then it will have a unique stationary matrix and successive state matrices will always approach this stationary matrix

Hidden Markov Model

- Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process-call it X-with unobservable ("hidden") states.
- HMM assumes that there is another process Y whose behavior "depends" on X
- HMM stipulates that, for each time instance n_0, the conditional probability distribution of Y_{n_0} given the history $\{X_n = x_n\}_{n=n_0}$ must NOT depend on $\{x_n\}_{n<n_0}$
- The goal is to learn about X by observing Y

Definition and application

- Definition: Let X_n and Y_n be discrete-time stochastic processes and $n \geq 1$. The pair (X_n, Y_n) is a hidden markov model if
  - X_n is a Markov process and is not directly observable ("hidden");
  - $\mathbf{P}\left(Y_n \in A \;\middle|\; X_1 = x_1, \ldots, X_n = x_n\right) = \mathbf{P}\left(Y_n \in A \;\middle|\; X_n = x_n\right)$, for every $n \geq 1$, $x_1, \ldots, x_n$, and an arbitrary (measurable) set A.
- The states of the process X_n are called hidden states, and $\mathbf{P}\left(Y_n \in A \;\middle|\; X_n = x_n\right)$ is called emission probability or output probability.
- Application: reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, and bioinformatics.

Example: a hypothetical dishonest casino

- The casino uses a fair die most of the time,
- Occasionally the casino secretly switches to a loaded die, and later the casino switches back to the fair die.
- A probabilistic process determines the switching back-and-forth from loaded die to fair die and back again after each toss of the die, with the switch from fair-to-loaded occurring with probability 0.05 and from loaded-to-fair with probability 0.1.
- Assume that the loaded die will come up “six" with probability 0.5 and the remaining five numbers with probability 0.1 each.

Example: a hypothetical dishonest casino

The transition matrix is

$$A = \begin{array}{c} \\ F \\ L \end{array} \begin{array}{c} \begin{array}{cc} F & L \end{array} \\ \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix} \end{array}$$

and the emission probability matrix is

$$B = \begin{array}{c} \\ F \\ L \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{bmatrix} \end{array}$$

If you can see only the sequence of rolls (the sequence of observations or signals) you do not know which rolls used a loaded die and which used a fair die, because the casino hides the state.

Question # 1 – Evaluation

GIVEN

A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

$$\underbrace{}_{\text{Prob} = 1.3 \times 10^{-35}}$$

QUESTION

How likely is this sequence, given our model of how the casino works?

This is the **EVALUATION** problem in HMMs

Question # 2 – Decoding

GIVEN

A sequence of rolls by the casino player

124552646214614613613	666166466163661636616361 6	5156151151461235623 44
FAIR	LOADED	FAIR

QUESTION

What portion of the sequence was generated with the fair die, and what portion with the loaded die?

This is the **DECODING** question in HMMs

Question # 3 – Learning

GIVEN

A sequence of rolls by the casino player

1245526462146146136136	661664661636616366163616	515615115146123562344

Prob(6) = 64%

QUESTION

How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?

This is the **LEARNING** question in HMMs

Random Walk

- A random walk is a stochastic or random process, that describes a path that consists of a succession of random steps on some mathematical space such as the integers.
- Examples
  - The random walk on the integer number line, $\mathbb{Z}$, which starts at 0 and at each step moves $+1$ or -1 with equal probability
  - the path traced by a molecule as it travels in a liquid or a gas
  - the search path of a foraging animal
  - the price of a fluctuating stock
  - the financial status of a gambler
- The term random walk was first introduced by Karl Pearson in 1905

Random Walk

- The term random walk most often refers to a special category of Markov chains or Markov processes
- Random walks can also take place on a variety of spaces
  - graphs
  - on the integers or the real line
  - in the plane or higher-dimensional vector spaces
  - on curved surfaces or higher-dimensional Riemannian manifolds
  - on finite groups, or Lie

1-dimensional Random Walk

- Take independent random variables $Z_1, Z_2, \ldots$, where each variable is either 1 or -1, with a probability of p and $1-p$, respectively. Set $S_0 = 0$ and $S_n = \sum_{j=1}^n Z_j$. The series $\{S_n\}$ is called the simple random walk on $\mathbb{Z}$.
- If $p = 0.5$, we have

$$E(S_n) = \sum_{j=1}^n E(Z_j) = 0$$

$$E(S_n^2) = \sum_{i=1}^n E(Z_i^2) + 2\sum_{1\le i<j\le n} E(Z_i Z_j) = n.$$

- A one-dimensional random walk can also be looked at as a **Markov chain**, whose state space is given by the integers $i = 0, \pm 1, \pm 2, \ldots$, the transition probablity

$$P_{i,i+1} = p = 1 - P_{i,i-1}.$$

Wiener Process

A standard (one-dimensional) Wiener process (depicts Brownian motion) is a stochastic process $\{W_t\}_{t \geq 0_+}$ indexed by nonnegative real numbers t with the following properties:

- $W_0 = 0$
- W has independent increments, i.e., for every $t > 0$, the future increments $W_{t+u} - W_t$,$u \geq 0$, are independent of the past values W_s, $s \leq t$.
- W has Gaussian increments: $W_{t+u} - W_t$ is normally distributed with mean 0 and variance u, $W_{t+u} - W_t \sim \mathcal{N}(0, u)$.
- W has continuous paths: W_t is continuous in t.

Wiener Process as a Limit of Random Walks

- One of the many reasons that Brownian motion is important in probability theory is that it is, in a certain sense, a limit of rescaled simple random walks.
- Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with mean 0 and variance 1. For each n, define a continuous time stochastic process

$$W_n(t) = \frac{1}{\sqrt{n}} \sum_{1 \le k \le \lfloor nt \rfloor} \xi_k, \qquad t \in [0, 1]$$

- Increments of W_n are independent because that ξ_k are independent.
- For large n, $W_n(t) - W_n(s)$ is close to $\mathcal{N}(0, t-s)$ by the central limit theorem.

Markov property of Wiener process

- For all $t_1 < t_2 \cdots < t_n$, given $W(t_1), \ldots, W(t_{n-1})$, the conditional probability density function of $P(W(t_n)|W(t_1), \ldots, W(t_{n-1}))$ is the same as $P(W(t_n)|W(t_{n-1}))$.
- For all $t_1 > t_2 \cdots > t_n$, given $W(t_1), \ldots, W(t_{n-1})$, we have

$$P(W(t_n)|W(t_1), \ldots, W(t_{n-1})) = P(W(t_n)|W(t_{n-1})).$$

- For all $t_1 < t_2 \cdots < t_n$, given $W(t_1), \ldots, W(t_{i-1}), W(t_{i+1}), W(t_n)$, then we have

$$\begin{aligned} P(W(t_i)|W(t_1), \ldots, W(t_{i-1}), W(t_{i+1}), W(t_n)) = \\ P(W(t_i)|W(t_{i-1}), W(t_{i+1})). \end{aligned}$$

Application of Wiener Process

- The Wiener process plays an important role in both pure and applied mathematics.
- In pure mathematics, the Wiener process gave rise to the study of continuous time martingales, it plays a vital role in stochastic calculus, diffusion processes and even potential theory.
- In applied mathematics, the Wiener process is used to represent the integral of a white noise Gaussian process
- It is useful as a model of noise in electronics engineering (see **Brownian noise**), instrument errors in filtering theory
- It is used to describe unknown forces in control theory

Poisson Processes

Let $N(t)$ be a stochastic process. It is called a homogeneous Poisson counting process with rate $\lambda > 0$ if

- $P\{N(0) = 0\} = 1$
- $\forall n \in N, 0 < t_0 < t_1 < ... < t_n$: The increments $N(t_0), N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$ are independent
- $\forall 0 < s < t : N(t) - N(s) \sim \text{Pois}(\lambda(t-s))$

It is clear that

$$P(N(t) = n) = P(N(t) - N(0) = n | N(0) = 0) = P(N(t) - N(0) = n)$$
$$= \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$
$$\sum_{n=0}^{\infty} p_n(t) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = 1, \forall t$$

Thinking

- Markov Property: For all $k \in \mathbb{N}$ and events $\{(X_r)_{r\leq t} \in A\}$and $\{(X_{t+s})_{s\geq 0} \in B\}$, we have: if $P(X_t = k, (X_r)_{r\leq t} \in A) > 0$, then

  $$P((X_{t+s})_{s\geq 0} \in B | X_t = k, (X_r)_{r\leq t} \in A) = P((X_{t+s})_{s\geq 0} \in B | X_t = k)$$

- Poisson process can be used for activity forecasting. "A Poisson Process Model for Activity Forecasting"

Examples

- the number of telephone calls at an office logged up to time t
- the number of vehicles which pass a roadside speed camera within a specified hour
- the number of students in Teaching Building 6 at time t
- $\cdots\cdots$

Example

The number of failures $N(t)$, which occur in a computer network over the time interval $[0, t)$, can be described by a homogeneous Poisson process $\{N(t), t \geq 0\}$. On an average, there is a failure after every 4 hours, i.e. the intensity of the process is equal to $\lambda = 0.25[h^{-1}]$. Derive the probability of at most 1 failure in $[0, 8)$.

Hints: $E[N(t)] = \lambda t, N(0) = 0$.

White noise

- In signal processing, white noise is a random signal having equal intensity at different frequencies, giving it a constant power spectral density.
- In discrete time, white noise is a discrete signal whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance.
- In particular, if each sample has a normal distribution with zero mean, the signal is said to be **Gaussian white noise**.

Power spectral density (Power spectrum)

- The power spectral density (PSD) refers to the measure of signal's power content versus frequency
- Parseval's theorem: Summation or integration of the spectral components yields the total power (for a physical process) or variance (in a statistical process), identical to what would be obtained by calculating the time average of $x^2(t)$, i.e.,

$$P = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} x^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)d\omega$$

PSD for continuous time random process

- The power spectrum $S_X(\omega)$ of a wide-sense stationary stochastic process $X(t)$ is defined as the Fourier transform of the autocorrelation.

$$S_X(\omega) = \int_{-\infty}^{\infty} R_X(\tau) e^{-j\omega\tau} d\tau$$

- The autocorrelation is the inverse Fourier transform of the power spectrum

$$R_X(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) e^{j\omega\tau} d\omega$$

- The power of a wide-sense stationary stochastic process (ergodic):

$$P_X = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} x^2(t) dt = E[X^2(t)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) d\omega$$

Cross power spectral density

- The cross power spectral density (CPSD) or cross spectral density (CSD) of two wide-sense stationary stochastic processes $X(t)$ and $Y(t)$ is Fourier transform of the cross correlation:

$$S_{XY}(\omega) = \int_{-\infty}^{\infty} R_{XY}(\tau) e^{-j\omega\tau} d\tau$$
$$R_{XY}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{XY}(\omega) e^{j\omega\tau} d\omega$$

Power spectral density for discrete-time random processes

The power spectral density of a discrete-time random process:

$$S_X(\omega) = \sum_{k=-\infty}^{\infty} R_X(k)e^{-j\omega k}, \omega \in [-\pi, \pi]$$

$$R_X(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_X(\omega)e^{j\omega k}d\omega$$

Discrete-time white noise

A discrete-time stochastic process $X(k)$ is called white noise if

$$\begin{aligned} R_X(k) &= \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \\ &= \sigma^2 \delta_k \end{aligned}$$

where δ_k is the Kronecker delta function, defined as

$$\delta_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}$$

Interpretation of discrete-time white noise

- If $X(k)$ is a discrete-time white noise process, then the RV $X(n)$ is uncorrelated with $X(m)$ unless $n = m$.
- The power spectral density of a discrete-time white noise process is equal at all frequencies:

$$S_X(\omega) = R_X(0), \forall \omega \in [-\pi, \pi]$$

Continuous-time white noise

- For a continuous-time random process, white noise has equal power at all frequencies (like white light):

$$S_X(\omega) = R_{X,0}, \forall\omega$$

- For continuous-time white noise, we have

$$R_X(\tau) = R_{X,0}\delta(\tau)$$

  where $\delta(\tau)$ is the continuous-time impulse function.

- Continuous-time white noise is not something that occurs in the real world because it has infinite power
- Many continuous-time processes approximate white noise and are useful in mathematical analysis of signals and systems

Continuous-time white noise

- An infinite-bandwidth white noise signal is a purely theoretical construction.
- The bandwidth of white noise is limited in practice by the mechanism of noise generation, by the transmission medium and by finite observation capabilities.
- Thus, a random signal is considered "white noise" if it is observed to have a flat spectrum over the range of frequencies that is relevant to the context.

Example

Suppose that a zero-mean stationary stochastic process has the autocorrelation function

$$R_X(\tau) = \sigma^2 e^{-\beta|\tau|}, \beta \in \mathbb{R}_+$$

Calculate the power spectrum as well as the power of the stochastic process.

Example

The power spectrum

$$
\begin{aligned}
S_X(\omega) &= \int_{-\infty}^{\infty} \sigma^2 e^{-\beta|\tau|e^{-j\omega\tau}} d\tau \\
&= \int_{-\infty}^{0} \sigma^2 e^{(\beta - j\omega)\tau} d\tau + \int_{0}^{\infty} \sigma^2 e^{-(\beta + j\omega)\tau} d\tau \\
&= \frac{\sigma^2}{\beta - j\omega} + \frac{\sigma^2}{\beta + j\omega} \\
&= \frac{2\sigma^2\beta}{\omega^2 + \beta^2}
\end{aligned}
$$

The variance (also power) of the stochastic process is computed as

$$
\begin{aligned}
E[X^2(t)] &= R_X(0) \\
&= P_X = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(\omega) d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2\sigma^2\beta}{\omega^2 + \beta^2} d\omega \\
&= \frac{\sigma^2}{\pi} \arctan \frac{\omega}{\beta} \Big|_{-\infty}^{\infty} \\
&= \sigma^2
\end{aligned}
$$