1 Probability and Random Variables

1.1 Probability Space

The term *random experiment* is used to describe any situation which has a set of possible outcomes, each of which occurs with a particular probability. To mathematically describe a random experiment we must specify:

1. The sample space Ω, which is the set of all possible outcomes of the random experiment. We call any subset of $A \subseteq \Omega$ an event.
2. A mapping/function P from events to a number in the interval $[0,1]$, i.e. $\{P(A), A \subset \Omega\}$.

We call P the *probability* and (Ω, P) the *probability space*.

Axioms of probability: A probability P assigns each event E, $E \subset \Omega$, a number in $[0,1]$ and P must satisfy the following properties:

1. $P(\Omega) = 1$
2. For events A, B such that $A \cap B = \emptyset$ (i.e. disjoint) then $P(A \cup B) = P(A) + P(B)$.
3. If $A_1, A_2, \cdots$ are disjoint then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Define the *indicator* function for a set or event E,

$$\mathbb{I}_E(t) = \begin{cases} 1, & \text{if } t \in E \\ 0, & \text{otherwise} \end{cases}$$

When Ω is a discrete set $\{\omega_1, \omega_2, \cdots\}$, given any non-negative sequence of numbers $p_1, p_2, \cdots$ that add to 1, we can define a valid probability:

$$P(A) = \sum_{i=1}^{\infty} \mathbb{I}_A(\omega_i) p_i$$

When Ω is the real line, probability can be specified through a probability density function (pdf) $f(t)$. For a general event E, we can calculate the probability using:

$$P(E) = \int_{-\infty}^{\infty} \mathbb{I}_E(t) f(t) dt$$

1.2 Conditional Probability

The *conditional probability* of event A occurring given that event B has occurred is defined to be:

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

Two events A and B are independent if $P(AB) = P(A \cap B) = P(A)P(B)$.

Probability chain rule:

$$P(A_1 \cdots A_{n-1} A_n) = P(A_1) \left(\prod_{i=2}^{n} P(A_i | A_1 \cdots A_{i-1}) \right)$$

Bayes' theorem:

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

1.3 Random Variables

Given a probability space (Ω, P), a random variableis a function $X(\omega)$ which maps each element ω of the sample space Ω onto a point on the real line.

For a discrete random variable X with range $\{x_1, x_2, \cdots\}$, we define the *probability mass function*(pmf) of X to be the function $p_X : \{x1, x2, \cdots\} \to [0,1]$ where:

$$p_X(x_i) = \Pr(X = x_i)$$

For any set A:

$$\Pr(X \in A) = \sum_{i=1}^{\infty} \mathbb{I}_A(x_i) p_X(x_i)$$

Continuous random variables are defined as having a *probability density function* (pdf.) A random variable X is continuous if there exists a non-negative function $f_X(x) \geq 0$ such that $\int_{-\infty}^{\infty} f_X(x) dx = 1$ and for any set A:

$$\Pr(X \in A) = \int_{-\infty}^{\infty} \mathbb{I}_A(x) f_X(x) dx$$

The *cumulative distribution function* (cdf) can describe both discrete and continuous random variables and is defined to be:

$$F_X(x) = \Pr(X \leq x)$$

The cdf has the following properties:

1. $0 \leq F_X(x) \leq 1$.
2. $F_X(x)$ is non-decreasing as x increases.
3. $\Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$
4. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
5. If X is a continuous r.v. then $F_X(x)$ is continuous.
6. If X is discrete then F_X is *right-continuous*: $F_X(x) = \lim_{t \downarrow x} F(t)$ for all x.

For a random variable $Y = r(X)$ where r is strictly increasing or strictly decreasing, r has an inverse $r^{-1} = s$, we can derive a formula for f_Y:

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|$$

1.4 Bivariates

A *bivariate* are two jointly distributed random variables. For two discrete random variables X and Y where $X \in \{x_1, \cdots, x_m\}$, $Y \in \{y_1, \cdots, y_n\}$, we can define the joint pmf to be:

$$p_{X,Y}(x_i, y_j) = \Pr(X = x_i, Y = y_j)$$

The marginal pmfs are $p_X(x_k) = \sum_j p_{X,y}(x_k, y_j)$ and $p_Y(y_k) = \sum_i p_{X,y}(x_i, y_k)$

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).

For the discrete rvs X and Y, the conditional pmf of X given $Y = y$ is:

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

For continuous random variables X and Y, we call a non-negative function $f(x,y)$ their joint probability density function if $\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x,y) dx \right) dy = 1$ and for any sets (events) $A \in R$ and $B \in R$:

$$\Pr(X \in A, Y \in B) = \int_{-\infty}^{\infty} \mathbb{I}_B(y) \left(\int_{-\infty}^{\infty} \mathbb{I}_A(x) f(x,y) dx \right) dy$$

Two continuous rvs X and Y are *independent*if and only if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

For the continuous rvs X and Y, the conditional pdf of X given $Y = y$ is:

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

The pdf of the sum of two independent rvs is the convolution of their pdfs. Let X_1 and X_2 be two independent rvs and $Y = X_1 + X_2$.

$$f_Y(y) = \int_{-\infty}^{\infty} f_2(y - x_1) f_1(x_1)$$

The expected value or mean value or first moment of a function $r(X,Y)$ of the bivariate (X,Y) is:

$$\mathbb{E}\{r(X,Y)\} = \begin{cases} \sum_y \sum_x r(x,y) p_{X,Y}(x,y), & \text{disc.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x,y) f_{X,Y}(x,y) dx dy, & \text{cts.} \end{cases}$$

The conditional expectation is:

$$\mathbb{E}\{r(X,Y)|Y = y\} = \begin{cases} \sum_x r(x,y) p_{X|Y}(x|y), & \text{disc.} \\ \int_{-\infty}^{\infty} r(x,y) f_{X|Y}(x|y) dx, & \text{cts.} \end{cases}$$

Rule of iterated expectation:

$$\mathbb{E}\{r(X,Y)\} = \mathbb{E}\left(\mathbb{E}\{r(X,Y)|Y\}\right)$$

1.5 Multivariates

Let $X_1, X_2, \cdots, X_n$ be n continuous (or discrete) random variables. We call $X = (X_1, \cdots, X_n) \in \mathbb{R}^n$ a continuous (or discrete) random vector.

Let $f(x_1, \cdots, x_n)$ be a non-negative function that integrates to 1. Then f is called the pdf of the random vector X if for all events $A_1, \cdots, A_n$:

$$\Pr(X_1 \in A_1, \cdots, X_n \in A_n) = \int_{-\infty}^{\infty} \mathbb{I}_{A_n}(x_n) \cdots \int_{-\infty}^{\infty} \mathbb{I}_{A_1}(x_1) f(x_1, \cdots, x_n) \, dx_1 \cdots dx_n$$

The ith marginal of $f(x_1, \cdots, x_n)$ is obtained by:

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \cdots, x_n) \, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

The n random variables $X_1, \cdots, X_n$ are independent if and only if for every $A_1, \cdots, A_n$:

$$\Pr(X_1 \in A_1, \cdots, X_n \in A_n) = \Pr(X_1 \in A_1) \cdots \Pr(X_n \in A_n)$$

Independence is equivalent to checking that the joint pdf reduces to the *product of marginals*:

$$f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

Independence: If $X_1, \cdots, X_n$ are independent random variables then $\mathbb{E}\{\prod_{i=1}^{n} X_i\} = \prod_{i=1}^{n} \mathbb{E}\{X_i\}$, i.e. the expectation of the product is the product of the expectation.

Linearity: If $X_1, \cdots, X_n$ are random variables and if $a_1, \cdots, a_n$ are real constants then $\mathbb{E}\{\sum i = 1^n a_i X_i\} = \sum_{i=1}^{n} a_i \mathbb{E}\{X_i\}$

The change of variable formula can be applied to random vectors. Let $Y = G(X)$:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} g_1(X_1, \cdots, X_n) \\ \vdots \\ g_n(X_1, \cdots, X_n) \end{bmatrix}$$

If G is invertible then $X = G^{-1}(Y)$. Let $H(Y) = G^{-1}(Y)$:

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} h_1(Y_1, \cdots, Y_n) \\ \vdots \\ h_n(Y_1, \cdots, Y_n) \end{bmatrix}$$

The matrix of partial derivatives of $H(y)$ forms the Jacobian:

$$J(y) = \begin{bmatrix} \frac{\partial}{\partial y_1} h_1 \cdots \frac{\partial}{\partial y_n} h_1 \\ \vdots \\ \frac{\partial}{\partial y_1} h_n \cdots \frac{\partial}{\partial y_n} h_n \end{bmatrix}$$

$$f_Y(y) = f_X(H(y)) \left| \det J(y) \right|$$

The *characteristic function* of a (discrete or continuous) random variable X is $\varphi_X(t) = \mathbb{E}\{\exp(itX)\}$, $t \in \mathbb{R}$. For a random vector $X = (X_1, X_2, \cdots, X_n)$, the characteristic function is $\varphi_X(t) = \mathbb{E}\{\exp(it^T X)\}$, $t \in \mathbb{R}^n$. Similarly to the Fourier transform, the characteristic function uniquely describes a pdf. Suppose that X and Y are random vectors with $\varphi_X(t) = \varphi_Y(t)$ for all $t \in \mathbb{R}^n$, then X and Y have the same probability distribution.

$$i^n \mathbb{E}(X^n) = \frac{d^n}{dt^n} \varphi_X(t = 0)$$

2 Random Processes

2.1 Random Process

A *discrete time* random (or stochastic) process is one of the following infinite collection of random variables: $\{X_n\}_{n=-\infty}^{\infty}$ or $\{X_n\}_{n=0}^{\infty}$ or $\{X_n\}_{n=1}^{\infty}$ *Random walk:*

$$X_n = \begin{cases} X_{n-1}+1, & \text{w.p. } q \\ X_{n-1}-1, & \text{w.p. } 1-q \end{cases}$$

$$X_0 = 0$$

2.2 Finite Dimensional Distributions

To completely specify a discrete time random process $X_0, X_1, \cdots$, we must specify their joint probability density function $f_{X_0,X_1,\cdots,X_n}(x_0,x_1,\cdots,x_n)$ for all integers $n \geq 0$ when $X_0, X_1, \cdots$ is a collection of continuous random variables.

If $X_0, X_1, \cdots$ is a collection of discrete random variables then we must specify their joint probability mass function $p_{X_0,X_1,\cdots,X_n}(x_0,x_1,\cdots,x_n)$ for all integers $n \geq 0$.

Markov chain: Let $\{X_n\}_{n\geq 0}$ be discrete random variables taking values in $S = \{1,\cdots,L\}$. The transition probability matrix Q is a non-negative matrix and each row sums to one.

$$Q = \begin{bmatrix} Q_{1,1} & Q_{1,2} & \cdots & Q_{1,L} \\ Q_{2,1} & Q_{2,2} & \cdots & Q_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{L,1} & Q_{L,2} & \cdots & Q_{L,L} \end{bmatrix}$$

The conditional pmf of X_n given $X_0 = i_0, \cdots, X_{n-1} = i_{n-1}$ is determined by Q:

$$\Pr(X_n = i_n | X_0 = i_0, \cdots, X_{n-1} = i_{n-1}) = Q_{i_{n-1}, i_n}$$

Assume the pmf of X_0 is $p_{X_0}(i) = \lambda_i$ where $\lambda = (\lambda_1, \cdots, \lambda_L)$ is given. The pair (λ, Q) completely defines the Markov chain. We call Q the *transition probability matrix* of the MC and λ the *initial distribution* of the chain. Only the most recent value $X_{n-1} = i_{n-1}$ is needed to generate X_n. This *limited memory* property is known as the Markov property.

Marginals of a Markov chain:

$$p(i_n) = (\lambda Q^n)_{i_n}$$

A discrete time random process $X_0, X_1, \cdots$ is *strictly stationary* if for all (section size) k and displacement $m > 0$:

$$f_{X_0,\cdots,X_k}(x_0,\cdots,x_k) = f_{X_m,\cdots,X_{k+m}}(x_m,\cdots,x_{k+m})$$

Strict stationarity means any two 'sections' of the process $(X_0,\cdots,X_k)$ and $(X_m,\cdots,X_{m+k})$ are statistically indistinguishable for any displacement m.

Invariant distribution of a Markov chain: Consider the transition probability matrix Q with state-space S. The pmf $\pi = (\pi_i : i \in S)$ is invariant for Q if $\pi Q = \pi$ for all $j \in S$:

$$\sum_{i \in S} \pi_i Q_{i,j} = \pi_j$$

The Markov chain (π, Q) is strictly stationary. The pmf of $(X_m, \cdots, X_{m+k})$, for any $m \in \{0, 1, \cdots\}$, can be written as:

$$p(i_m, \cdots, i_{m+k}) = \pi_{i_m} Q_{i_m, i_{m+1}} \cdots Q_{i_{m+k-1}, i_{m+k}}$$

Ergodic theorem: When the MC is irreducible then for any initial distribution λ, the sample (or empirical) average converges to the ensemble average:

$$\frac{1}{n+1} \sum_{k=0}^{n} r(X_k) \to \sum_{i \in S} \pi_i r(i)$$

An irreducible Markov chain refers to a chain where all state values in S communicate with each other. This means for any pair of states (i,j), the Markov chain starting in i will eventually visit j and vice versa.

2.3 Time-Series Analysis

A time series is a set of observations y_n, $n = 0, 1, \cdots$, arranged in increasing time.

White noise: Let $\{W_n\}_{n=-\infty}^{\infty}$ be a sequence of random variables such that $\mathbb{E}(W_n) = 0$ for all n,

$$\mathbb{E}(W_i W_j) = \begin{cases} \sigma^2, & \text{for } i = j \\ 0, & \text{for } i \neq j \end{cases}$$

Auto-regressive (AR) process: The $\mathrm{AR}(p)$ process $\{X_n\}_{n=-\infty}^{\infty}$ of the *order* p is:

$$X_n = \left(\sum_{i=1}^{p} a_i X_{n-i} \right) + W_n$$

For the $\mathrm{AR}(1)$ case:

$$X_n = a X_{n-1} + W_n = \sum_{k=0}^{\infty} W_{n-k} a^k$$

$\mathrm{AR}(1)$ is causal with impulse response $\{a^k\}_{k\geq 0}$.

$$\mathbb{E}\{X_n\} = \sum_{k=0}^{\infty} \mathbb{E}\left\{W_{n-k} a^k\right\} = 0$$

$$\mathbb{E}\{X_n^2\} = \sum_{k=0}^{\infty} \mathbb{E}\left\{W_{n-k}^2 a^{2k}\right\} = \frac{\sigma^2}{1-a^2}$$

Wide sense stationary (WSS): $\{X_n\}_{n=-\infty}^{\infty}$ is wide-sense stationary if:

1. $\mathbb{E}\{X_n\} = \mu$ for all n (has constant mean)
2. $\mathbb{E}\{X_n^2\} < \infty$ for all n (has finite variance)
3. $\mathbb{E}\{X_{n_1} X_{n_2}\} = \mathbb{E}\{X_{n_1+k} X_{n_2+k}\}$ for any n_1, n_2, k.

The *correlation function* of a WSS process is defined as $R_X(k) = \mathbb{E}\{X_0 X_k\}$. The $\mathrm{AR}(1)$ process is WSS and $R_X(k) = a^k \sigma_X^2$.

Moving average (MA) process: The $\mathrm{MA}(q)$ process $\{X_n\}_{n=-\infty}^{\infty}$ of the *order* q is:

$$X_n = \sum_{i=1}^{q} b_i W_{n-i} + W_n$$

$$\mathbb{E}\{X_n^2\} = \sum_{i=0}^{q} b_i^2 \mathbb{E}\left\{W_{n-i}^2\right\} + \mathbb{E}\left\{W_n^2\right\} = \sigma^2(1 + b_1^2 + \cdots + b_q^2)$$

If the input $\{W_n\}_{n=-\infty}^{\infty}$ of a discrete time LTI system with impulse response $\{h_n\}_{n=-\infty}^{\infty}$ is WSS then its output $\{Y_n\}_{n=-\infty}^{\infty}$ is also WSS.

$$\mathbb{E}\{Y_n\} = \mathbb{E}\left\{ \sum_{k=-\infty}^{\infty} h_{n-k} W_k \right\} = \mathbb{E}\{W_0\} \sum_{k=-\infty}^{\infty} h_{n-k}$$

The correlation of the output is

$$\mathbb{E}\{Y_{n_1} Y_{n_2}\} = \mathbb{E}\left\{ \sum_{l=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} h_k h_l W_{n_1-k} W_{n_2-l} \right\} = \sum_{l=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} h_k h_l R_W(n_2 - n_1 + k - l)$$

Thus the MA process is WSS.

2.4 Power Spectrum

Let $R_X(k)$ be the correlation function of a discrete time WSS process. The power spectrum density $S_X(f)$ is:

$$S_X(f) = \sum_{k=-\infty}^{\infty} R_X(k) e^{-j2\pi f k}$$

The inversion formula is:

$$R_X(n) = \int_{-1/2}^{1/2} S_X(f) e^{j2\pi f n} df$$

Power spectrum shows how the variance of X_n is spread across frequency. If $R_X(k)$ is the correlation function of a discrete time WSS process then the power spectrum density $S_X(f)$ is an even, real valued and nonnegative function of f. Moreover, $S_X(f)$ is a continuous function if $\sum_{k=-\infty}^{\infty} |R_X(k)| < \infty$

The area under $S_X(f)$ in the frequency band gives the power of the bandlimited output Y_n.

*The ARMA model:*The $\mathrm{ARMA}(p,q)$ process $\{X_n\}_{n=-\infty}^{\infty}$ is the discrete time process satisfying:

$$X_n = \sum_{i=1}^{p} a_i X_{n-i} + W_n + \sum_{i=1}^{q} b_i W_{n-i}$$

The ARMA process is WSS and it can be expressed as a causal filter applied to $\{W_n\}_{n=-\infty}^{\infty}$.

For a random process $\{X_n\}$, the expected instantaneous power is $\mathbb{E}\left\{X_n^2\right\}$ while the *expected average power* is:

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \mathbb{E}\left\{X_n^2\right\}$$

If the process is WSS, then $\mathbb{E}\left\{X_n^2\right\} = R_X(0)$ for all n.

$$R_X(0) = \int_{-1/2}^{1/2} S_X(f) df$$

The area under the (non-negative) function $S_X(f)$ gives the total power of the WSS process $\{X_n\}_{n=-\infty}^{\infty}$

The area under $S_X(f)$ in the frequency band gives the power of the bandlimited output Y_n.

$$R_Y(0) = \int_{-f_2}^{-f_1} S_X(f) df + \int_{f_1}^{f_2} S_X(f) df = 2 \int_{f_1}^{f_2} S_X(f) df$$

3 Detection, Estimation and Inference

3.1 Discrete-time Random Processes

We define a discrete-time random process as an ensemble of functions of ω which is a random variable having a probability density function $f(\omega)$.

$$\{X_n(\omega)\}, n = -\infty, \cdots, -1, 0, 1, \cdots, \infty$$

The mean of a random process $\{X_n\}$ is defined as $\mathbb{E}[X_n]$ and the *autocorrelation* function as:

$$r_{XX}[n,m] = \mathbb{E}[X_n X_m]$$

The *cross-correlation* function between two processes $\{X_n\}$ and $\{Y_n\}$ is:

$$r_{XY}[n,m] = \mathbb{E}[X_n Y_m]$$

A stationary process has the same statistical characteristics irrespective of shifts along the time axis.

For a wide-sense stationary random process $\{X_n\}$, the power spectrum is defined as the discrete-time Fourier transform (DTFT) of the discrete autocorrelation function:

$$S_X\left(e^{j\Omega}\right) = \sum_{m=-\infty}^{\infty} r_{XX}[m] e^{-jm\Omega}$$

The normalised frequency is $\Omega = \omega T$ where T is the sampling interval of the discrete time process and ω is the sampling frequency. The power spectrum can be interpreted as a density spectrum in the sense that the mean-squared signal value at the output of an ideal band-pass filter with lower and upper cut-off frequencies of ω_l and ω_u is given by:

$$E\left[Y_n^2\right] = \frac{1}{\pi} \int_{\omega_l T}^{\omega_u T} S_X\left(e^{j\Omega}\right) d\Omega$$

White noise is defined in terms of its *auto-covariance* function. A wide sense stationary process is termed white noise if:

$$c_{XX}[m] = \mathbb{E}[(X_n - \mu)(X_{n+m} - \mu)] \\ = \sigma_X^2 \delta[m]$$

$\sigma_X^2 = \mathbb{E}\left[(X_n - \mu)^2\right]$ is the variance of the process. The power spectrum of zero mean white noise is σ_X^2.

The N-th order pdf for the Gaussian white noise process is:

$$f_{X_{n_1},X_{n_2},\cdots,X_{n_N}}(\alpha_1,\alpha_2,\cdots,\alpha_N) = \prod_{i=1}^{N} \mathcal{N}\left(\alpha_i|0,\sigma_X^2\right)$$

The Gaussian white noise process is Strict sense stationary.

When a wide-sense stationary discrete random process $\{X_n\}$ is passed through a stable, linear time invariant (LTI) system with digital impulse response $\{h_n\}$, the output process $\{Y_n\}$ is also WSS.

$$y_n = \sum_{k=-\infty}^{\infty} h_k x_{n-k} = x_n \star h_n$$

The output correlation functions and power spectra can be expressed in terms of the input statistics and the LTI system:

$$r_{XY}[k] = \sum_{l=-\infty}^{\infty} h_l r_{XX}[k-l] = h_k * r_{XX}[k]$$

$$r_{YY}[l] = \sum_{k=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} h_k h_i r_{XX}[l+i-k] \\ = h_l * h_{-l} * r_{XX}[l]$$

$$S_Y\left(e^{j\omega T}\right) = \left|H\left(e^{j\omega T}\right)\right|^2 S_X\left(e^{j\omega T}\right)$$

For an *Ergodic* random process we can estimate expectations by performing time-averaging on a single sample function:

$$\mu = \mathbb{E}[X_n] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n$$

$$r_{XX}[k] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} x_n x_{n+k}$$

A necessary and sufficient condition for *mean ergodicity* is given by:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} c_{XX}[k] = 0$$

3.2 Optimal Filtering Theory

Wiener filter is a linear filter which would optimally estimate d_n given just the noisy observations x_n and some assumptions about the statistics of the random signal and noise processes.

$$x_n = d_n + v_n$$

In general, we can filter the observed signal x_n with an infinite dimensional filter, having a non-causal impulse response h_p, to obtain an estimate $\hat{d}_n$ of the desired signal:

$$\hat{d}_n = \sum_{p=-\infty}^{\infty} h_p x_{n-p}$$

We can measure performance of the filter in terms of expectations using the mean-squared error (MSE) criterion:

$$\epsilon_n = d_n - \hat{d}_n = d_n - \sum_{p=-\infty}^{\infty} h_p x_{n-p}$$

$$J = \mathbb{E}\left[\epsilon_n^2\right]$$

The Wiener filter assumes that $\{x_n\}$ and $\{d_n\}$ are jointly wide-sense stationary while $\{d_n\}$ and $\{v_n\}$ have zero mean.

The expected error may be minimised with respect to the impulse response values h_q:

$$\frac{\partial J}{\partial h_q} = \mathbb{E}\left[\frac{\partial \epsilon_n^2}{\partial h_q}\right] = \mathbb{E}\left[2\epsilon_n \frac{\partial \epsilon_n}{\partial h_q}\right] = 0$$

The orthogonality principle:

$$\mathbb{E}[\epsilon_n x_{n-q}] = r_{xd}[q] - \sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = 0$$

The Wiener-Hopf equations:

$$\sum_{p=-\infty}^{\infty} h_p r_{xx}[q-p] = h_q * r_{xx}[q] = r_{xd}[q]$$

The *cross-power spectrum* of d and x is:

$$S_{xd}\left(e^{j\Omega}\right) = H\left(e^{j\Omega}\right) S_x\left(e^{j\Omega}\right)$$

The cross-power spectrum is in general complex valued and measures the coherence between two process at a particular frequency.

Frequency domain Wiener filter:

$$H\left(e^{j\Omega}\right) = \frac{S_{xd}\left(e^{j\Omega}\right)}{S_x\left(e^{j\Omega}\right)}$$

The minimum mean-squared error value of the optimal filter:

$$J_{\min} = \mathbb{E}[\epsilon_n d_n] = r_{dd}[0] - \sum_{p=-\infty}^{\infty} h_p r_{xd}[p]$$

The minimum error in frequency domain:

$$\frac{1}{2\pi} \int_{-\pi}^{+\pi} S_d\left(e^{j\Omega}\right) - H\left(e^{j\Omega}\right) S_{xd}^*\left(e^{j\Omega}\right) d\Omega$$

When the desired signal process d_n is uncorrelated with the noise process v_n:

$$r_{dv}[k] = \mathbb{E}[d_n v_{n+k}] = 0$$

$$r_{xd}[q] = r_{dd}[q]$$

$$r_{xx}[q] = r_{dd}[q] + r_{vv}[q]$$

Thus the Wiener filter becomes:

$$H\left(e^{j\Omega}\right) = \frac{S_d\left(e^{j\Omega}\right)}{S_d\left(e^{j\Omega}\right) + S_v\left(e^{j\Omega}\right)} \\ = \frac{1}{1 + 1/\rho(\Omega)}$$

$\rho(\Omega) = S_d\left(e^{j\Omega}\right)/S_v\left(e^{j\Omega}\right)$ is the (frequency dependent) signal-to-noise (SNR) power ratio. At those frequencies where the SNR is large, the gain of the filter tends to unity; whereas the gain tends to a small value at those frequencies where the SNR is small. The minimum expected error in this case reduces, in the frequency domain to:

$$J_{\min} = \frac{1}{2\pi} \int_{-\pi}^{+\pi} S_d\left(e^{j\Omega}\right) \left(\frac{1}{1 + \rho(\Omega)}\right) d\Omega$$

In general, the Wiener filter is non-causal, and hence physically unrealisable. In the a causal P-th order Finite Impulse Response (FIR) Wiener filter, the signal estimate is formed as:

$$\hat{d}_n = \sum_{p=0}^{P} h_p x_{n-p}$$

The filter derivation proceeds much as before, leading to Wiener-Hopf equations as follows:

$$\sum_{p=-0}^{P} h_p r_{xx}[q-p] = r_{xd}[q]$$

The equations may be written in matrix form as $R_x h = r_{xd}$:

$$R_x \begin{bmatrix} h_0 \\ \vdots \\ h_P \end{bmatrix} = \begin{bmatrix} r_{xd}[0] \\ \vdots \\ r_{xd}[P] \end{bmatrix}$$

The *correlation matrix* R_x is given by:

$$R_x = \begin{bmatrix} r_{xx}[0] & \cdots & r_{xx}[P] \\ \vdots & \ddots & \vdots \\ r_{xx}[P] & \cdots & r_{xx}[0] \end{bmatrix}$$

The correlation matrix is symmetric and has constant diagonals (a symmetric Toeplitz matrix) since $r_{xx}[k] = r_{xx}[-k]$. The coefficient vector of the FIR Wiener filter:

$$h = R_x^{-1} r_{xd}$$

The minimum mean-squared error is given by:

$$J_{\min} = r_{dd}[0] - r_{xd}^T h = r_{dd}[0] - r_{xd}^T R_x^{-1} r_{xd}$$

3.3 Optimal Detection

The *matched filter* detects a known deterministic signal s_n buried in random noise v_n.

$$x_n = s_n + v_n$$

The output of an FIR filter at time $N-1$ is:

$$y_{N-1} = \sum_{m=0}^{N-1} h_m x_{N-1} = h^T \tilde{x} \\ = h^T \tilde{s} + h^T \tilde{v} = y_{N-1}^s + y_{N-1}^n$$

$\tilde{x} = [x_{N-1}, x_{N-2}, \cdots, x_0]^T$ is the time-reversed vector and $\tilde{s}$ is defined similarly.

Define the signal-to-noise ratio (SNR) at the output of the filter as:

$$\frac{\mathbb{E}\left[|y_{N-1}^s|^2\right]}{\mathbb{E}\left[|y_{N-1}^n|^2\right]} = \frac{|h^T \tilde{s}|^2}{\mathbb{E}\left[|h^T \tilde{v}|^2\right]}$$

We can represent the filter coefficient vector h as a linear combination of the eigenvectors of $\tilde{s}\tilde{s}^T$:

$$h = \alpha e_0 + \beta e_1 + \gamma e_2 + \cdots$$

The unit length vector $e_0 = \tilde{s}/|\tilde{s}|$ is an eigenvector and $\alpha = (\tilde{s}^T \tilde{s})$ is the corresponding eigenvalue. The set of $N-1$ orthonormal vectors $e_1, e_2, \cdots$ are orthogonal to $\tilde{s}$ with eigenvalue $\beta = \gamma = \cdots = 0$. The SNR may be expressed as:

$$\frac{|h^T \tilde{s}|^2}{\mathbb{E}\left[|h^T \tilde{v}|^2\right]} = \frac{\alpha^2 \tilde{s}^T \tilde{s}}{\sigma_v^2 \left(\alpha^2 + \beta^2 + \gamma^2 + \cdots\right)}$$

The largest possible value of α given that $|h| = 1$ corresponds to $\alpha = 1$.

$$h^{\text{opt}} = e_0 = \frac{\tilde{s}}{|\tilde{s}|}$$

The optimal filter coefficients are just the (normalised) time-reversed signal. The maximum SNR at the optimal filter setting is given by:

$$\text{SNR}^{\text{opt}} = \frac{\tilde{s}^T \tilde{s}}{\sigma_v^2}$$

3.4 Estimation Theory and Inference

In estimation theory, we start off with a vector of signal measurements x and some unknown quantities, or parameters, θ that we wish to infer. The probability distribution of data x can be expressed in terms of a joint probability density function (or probability mass function if the data are discrete-valued), or *likelihood* function:

$$p(x|\theta)$$

The *prior* probability density function can be formulated for θ from physical or other modelling considerations:

$$p(\theta)$$

In the Linear Model it is assumed that the data x are generated as a linear function of the parameters θ with an additive random modelling error term e.

$$x = G\theta + e$$

G is the design matrix.

Einstein-Wiener-Khinchin Theorem: Take a time-windowed version of the signal x_n, having duration $2N+1$ samples and zero elsewhere:

$$x_n^N = w_n^N x_n$$

$$w_n^N = \begin{cases} 1, & -N \le n \le N \\ 0, & \text{otherwise} \end{cases}$$

We have the following DTFT relationship:

$$\text{DTFT}\{r_{xx}[m]t[m]\} = \mathbb{E}\left[\frac{1}{2N+1}|X^N\left(e^{j\Omega}\right)|^2\right]$$

$t[m]$ is the deterministic autocorrelation function of the window function w_n. In the limit we can prove that the power spectrum is proportional to the expected value of the DTFT-squared of the data.

$$S_x\left(e^{j\Omega}\right) = \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{2N+1}|X^N\left(e^{j\Omega}\right)|^2\right]$$

An estimator is termed *unbiased* if $\mathbb{E}[\hat{\mu}] = \mu$. An estimator is termed *consistent* if it is unbiased and its variance tends to zero as $N \to \infty$.

The 'best fit' model that matches the data can be found by minimising the error $J = e^T e$. For invertible $G^T G$, the classical Ordinary Least Squares (OLS) estimator of θ is:

$$\theta^{\text{OLS}} = \left(G^T G\right)^{-1} G^T x$$

$$\mathbb{E}\left[\theta^{\text{OLS}}\right] = \left(G^T G\right)^{-1} G^T G\theta = \theta$$

The OLS estimator is the minimum variance unbiased estimator of θ. Such an estimator is termed a Best Linear Unbiased Estimator (BLUE).

The Maximum Likelihood (ML) estimate for θ is the value of θ which maximises the likelihood for given observations x:

$$\theta^{ML} = \underset{\theta}{\text{argmax}}\{p(x|\theta)\}$$

The ML solution is identical to the OLS solution for the linear Gaussian model. The noise variance is the just mean-squared error at the ML parameter solution.

$$\sigma_e^{2ML} = J^{ML}/N$$

The posterior or *a posteriori* probability for the parameter is given by Bayes' Theorem:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta)$$

The denominator $p(x)$, referred to as the marginal likelihood, is constant for any given observation x.

$$p(x) = \int p(x|\theta)p(\theta)dx$$

The maximum a posteriori (MAP) estimate is the value of θ which maximises the posterior distribution:

$$\theta^{MAP} = \underset{\theta}{\text{argmax}}\{p(\theta|x)\}$$

Suppose that the prior on parameter vector θ is the multivariate Gaussian $p(\theta) = \mathcal{N}(m_\theta, C_\theta)$. The MAP estimator for Linear Gaussian model is then given by:

$$\theta^{\text{MAP}} = \Phi^{-1}\Theta = \left(G^T G + \sigma_e^2 C_\theta^{-1}\right)^{-1}\left(G^T x + \sigma_e^2 C_\theta^{-1} m_\theta\right)$$

The posterior distribution is itself a multivariate Gaussian:

$$p(\theta|x) = \mathcal{N}\left(\theta^{\text{MAP}}, \sigma_e^2\Phi^{-1}\right)$$

We can write the expected cost over all of the unknown parameters, conditional upon the observed data x:

$$\mathbb{E}\left[C(\hat{\theta},\theta)\right] = \int C(\hat{\theta},\theta)p(\theta|x)d\theta$$

A classic estimation technique related to the Wiener filtering objective function is the Minimum mean-squared error (MMSE) estimation method.

$$\theta^{\text{MMSE}} = \underset{\hat{\theta}}{\text{argmin}}\,\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}[\theta|x] = \int \theta p(\theta|x)d\theta$$

The MMSE estimator for the Linear Gaussian model is identical to the MAP solution.

(The End)