

(i) $0 \leq \Pr(A) \leq 1$
If one event occurs the other events do not

(ii) $\Pr(\emptyset) = 0$

(iii) If A_1, A_2, \dots is a sequence of mutually exclusive events, then $\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$

(i) $\Pr(A^c) = 1 - \Pr(A)$

Proof: $\Pr(A \cup A^c) = \Pr(A) + \Pr(A^c)$

$\Pr(A^c) = 1 - \Pr(A)$

(ii) $\Pr(\emptyset) = 0$, $\emptyset = \Omega^c$

(iii) If $A \subseteq B$ then $\Pr(A) \leq \Pr(B)$

Proof: $B = A \cup (A^c \cap B)$

$\Pr(B) = \Pr(A) + \Pr(A^c \cap B) \geq \Pr(A)$

(iv) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

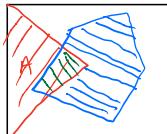
Union Bound

$$\Pr(\bigcup_i A_i) \leq \sum_i \Pr(A_i)$$

Conditional Probability

If A and B are two events, $\Pr(B) \neq 0$, then

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



"B becomes the new universe, $B = \Omega'$ "

Total Probability Theorem

If $\{E_1, E_2, \dots, E_k\}$ partition Ω , then

$\Pr(A) = \sum_{i=1}^k \Pr(A \cap E_i)$

This makes more sense since you know which event which occurred

$= \sum_{i=1}^k \Pr(A|E_i) \Pr(E_i)$

Example of Partition for k=4

$\Omega = E_1 \cup E_2 \cup E_3 \cup E_4$

No intersections
 $\Pr(E_i \cap E_j) = 0$

generative models

Bayes Rule

Want to find "ground truth" that gave this observation.

$$\Pr(E_i|A) = \frac{\Pr(A \cap E_i)}{\Pr(A)} = \frac{\Pr(E_i) \Pr(A|E_i)}{\sum_j \Pr(E_j) \Pr(A|E_j)} \xrightarrow{\text{prior knowledge / generative model}} \Pr(A) \text{ by TPT}$$

Independence

Two events, A_1, A_2 , are independent if $\Pr(A_1 \cap A_2) = \Pr(A_1)\Pr(A_2)$

$$\Rightarrow \Pr(A_1 | A_2) = \frac{\Pr(A_1 \cap A_2)}{\Pr(A_2)} = \frac{\Pr(A_1)\Pr(A_2)}{\Pr(A_2)} = \Pr(A_1)$$

Can be extended:

Events $\{A_1, A_2, \dots, A_n\}$ are independent if $\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \Pr(A_1)\Pr(A_2) \dots \Pr(A_n) \neq \prod_{i=1}^n \Pr(A_i)$

$$F_{X_1, \dots, X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n)$$

$$\Pr[X_1 \leq x_1, \dots, X_n \leq x_n] = \Pr[X_1 \leq x_1] \dots \Pr[X_n \leq x_n]$$

Random Variables

Given $(\Omega, \mathcal{F}, \Pr)$ a random variable is a function $X: \Omega \rightarrow \mathbb{R}$ such that $\forall z \in \mathbb{R}$, $\{\omega | X(\omega) \leq z\} \in \mathcal{F} \subseteq \Omega$

M^{th} Moment M^{th} Central Moment

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

$$\mathbb{E}[(X - \mathbb{E}[X])^n]$$

Cumulative Distribution Function

The CDF is defined as

$$F_X(z) = \Pr(X \leq z) \quad \forall z \in \mathbb{R}$$

$$= \Pr(\{\omega | X(\omega) \leq z\})$$

Properties of CDF

$$\lim_{z \rightarrow -\infty} F_X(z) = 0$$

$$\lim_{z \rightarrow \infty} F_X(z) = 1$$

$$\textcircled{2} \quad \forall z < y \quad F_X(z) \leq F_X(y)$$

$$\textcircled{3} \quad F \text{ is right continuous}$$

$$\lim_{x \rightarrow x_0^+} F_X(x) = F_X(x_0)$$

$$\textcircled{4} \quad \Pr[X \leq x \leq y] = F_X(y) - F_X(x)$$

PMF for Discrete Random Variables

$$p(x) = \Pr[X=x]$$

$$f_X(x) = \sum_{u \in x} p(u)$$

PDF for Continuous Random Variables

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \begin{cases} \text{measures how fast} \\ \text{we accumulate probability} \end{cases}$$

$$\text{thus } F_X(a) = \int_{-\infty}^a f_X(u) du$$

Must be non-negative

Properties of PDF

$$\textcircled{1} \quad f_X(x) \geq 0 \quad \forall x$$

$$\textcircled{2} \quad \int_{-\infty}^{\infty} f_X(x) dx = 1 = F_X(\infty)$$

$$\textcircled{3} \quad \Pr[x \leq X \leq y] = \int_x^y f_X(u) du$$

Variance

How much a r.v. varies from its expectation

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$= \mathbb{E}[X^2 - 2\mathbb{E}[X] + (\mathbb{E}[X])^2]$$

$$= \mathbb{E}[X^2] + (\mathbb{E}[X])^2$$

$$\text{Var}(X) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx = \int_{-\infty}^{+\infty} (X - \mathbb{E}[X])^2 f_X(x) dx$$

Correlation

Correlation between X and Y

$$\mathbb{E}[XY] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy f_{X,Y}(x,y) dx dy$$

$$\mathbb{E}[XY] = \sum_{x_i} \sum_{y_j} x_i y_j p_{X,Y}(x_i, y_j)$$

Covariance

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

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

If $\mathbb{E}[XY] = 0$, we say X is orthogonal to Y.

If $\text{Cov}(X, Y) = 0$, we say X is uncorrelated to Y.

Properties

Independence of r.v.s X, Y

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \Rightarrow \text{uncorrelatedness}$$

X, Y, Z r.v.'s

$$\text{Cov}(X, aY + bZ) = a\text{Cov}(X, Y) + b\text{Cov}(X, Z)$$

$$\text{Cov}(X - \mathbb{E}[X], Y - \mathbb{E}[Y]) = \text{Cov}(X, Y)$$

Expectation

$$\mathbb{E}[X] = \begin{cases} \sum_k k \Pr[X=k] & , X \text{ discrete} \\ \int_{-\infty}^{+\infty} x f_X(x) dx & , X \text{ continuous} \end{cases}$$

Properties

LOTUS Rule

If

$$Y = g(X)$$

then

$$\mathbb{E}[Y] = \int_{-\infty}^{+\infty} x f_Y(y) dy$$

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

Linearity of Expectation

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y], \quad \alpha, \beta \in \mathbb{R}$$

Preservation of Order

If

$$\Pr[X \geq Y] = 1$$

then

$$\mathbb{E}[X] \geq \mathbb{E}[Y]$$

Integration by Parts

A discrete, non-negative r.v. $X = 0, 1, 2, 3, \dots$

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i \Pr(X=i)$$

$$= \sum_{i=0}^{\infty} \Pr(X > i) \quad \begin{matrix} \text{Tail Probability} \\ 1 - F_X(i) \end{matrix}$$

To see this, observe

$$\sum_{i=0}^{\infty} \Pr(X=i) = 0 \cdot \Pr(X=0) + 1 \cdot \Pr(X=1) + 2 \cdot \Pr(X=2) + \dots$$

$$\sum_{i=0}^{\infty} \Pr(X > i) = \Pr(X > 0) + \Pr(X > 1) + \dots$$

etc

Thus

$$\mathbb{E}[X] = \int_0^{\infty} (1 - F_X(x)) dx = \int_{-\infty}^0 F_X(x) dx$$

Conditioning on Random Variables

Suppose X and Y have joint pmf $p_{X,Y}(x,y)$

The conditional pmf of X given $\{Y=y\}$ is

$$\Pr_{X|Y}(x|y) \triangleq \Pr[X=x | Y=y]$$

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

Similarly for continuous r.v.'s

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{if } f_Y(y) \neq 0 \quad \forall y$$

Conditional Expectation

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

$\mathbb{E}[X|Y]$ is a random variable which takes on value $\mathbb{E}[X|Y=y]$ w/ density $f_Y(y)$.

Since $\mathbb{E}[X|Y]$ is r.v. can take its expectation.

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]] = \mathbb{E}[X]$$

Cauchy-Schwartz Inequality

$$\text{IE}[XY] \leq \sqrt{\text{IE}[X^2]} \sqrt{\text{IE}[Y^2]}$$

Correlation Coefficient from Cauchy-Schwartz inequality

$$P_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad (|P_{XY}| \leq 1)$$

If X_1, \dots, X_m are pairwise uncorrelated, then

$$\text{Var}(X_1 + X_2 + \dots + X_m) = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_m)$$

$$\text{Var}\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m \text{Var}(X_i)$$

Linear Minimum Mean Squared Error

Now our estimator $g(Y)$ takes the form

$$g(Y) = aY + b$$

The MSE becomes

$$\text{MSE} = \text{IE}[(X - g(Y))^2]$$

Want

$$\{a^*, b^*\} = \underset{a,b}{\operatorname{argmin}} \text{IE}[(X - aY - b)^2]$$

Use orthogonality principle

$$W \perp g(Y) \Rightarrow X - aY - b \perp aY + b \quad \forall a, b$$

Consider the following cases

$$\begin{aligned} \text{① } a=0, b=1 \\ X - a^*Y - b^* \perp Y \Rightarrow \text{IE}[(X - a^*Y - b^*)Y] = 0 \\ \text{IE}[X] - \text{IE}[a^*Y] - \text{IE}[b^*] = 0 \quad \xrightarrow{\text{W is UNBIASED}} \\ \Rightarrow b^* = \text{IE}[X] - a^* \text{IE}[Y] \end{aligned}$$

$$\text{② } a=1, b=0$$

$$X - a^*Y - b^* \perp Y \Rightarrow \text{IE}[(X - a^*Y - b^*)Y] = 0$$

$$\text{IE}[(X - a^*Y - b^*)Y]$$

$$= \text{IE}[(X - a^*Y - \text{IE}[Y]) - a^*(Y - \text{IE}[Y])]Y$$

$$\xrightarrow{\text{Frob Norm}} \text{IE}[\text{Cov}(X - a^*Y - \text{IE}[Y], Y)]$$

$$\Rightarrow \text{Cov}(X - a^* \text{IE}[Y] - a^*(Y - \text{IE}[Y]), Y) = 0$$

$$\text{Cov}(X, Y) - a^* \text{Cov}(Y, Y) = 0$$

$$a^* = \text{Cov}(X, Y) / \text{Var}(Y) \xleftarrow{\text{W is UNBIASED}}$$

Aside

$$\text{Cov}(X, aY + bZ) = a \text{Cov}(X, Y) + b \text{Cov}(X, Z)$$

$$\text{Cov}(X - \text{IE}[X], Y - \text{IE}[Y]) = \text{Cov}(X, Y)$$

So

$$\begin{aligned} \hat{x}_{\text{MSE}} &= a^*Y + b^* \\ &= \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} Y + \text{IE}[X] - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \text{IE}[Y] \\ &= \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{IE}[Y]) + \text{IE}[X] \end{aligned}$$

and the MSE is

$$\text{MSE} = \text{IE}[W^2] = \text{Cov}(W, W)$$

$$= \text{Cov}((X - \text{IE}[X]) - a^*(Y - \text{IE}[Y]), (X - \text{IE}[X]) - a^*(Y - \text{IE}[Y]))$$

$$= \text{Cov}(X, X) + 2a^* \text{Cov}(X, Y) - 2a^* \text{Cov}(Y, X)$$

$$= \text{Var}(X) + \frac{(\text{Cov}(X, Y))^2}{\text{Var}(Y)} \text{Var}(Y) - 2 \frac{(\text{Cov}(X, Y)) \text{Cov}(Y, X)}{\text{Var}(Y)}$$

$$= \text{Var}(X) - \frac{(\text{Cov}(X, Y))^2}{\text{Var}(Y)}$$

$$W = X - \hat{x}_{\text{MSE}}$$

$$= X - (\text{IE}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{IE}[Y]))$$

$$= X - \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} Y - (\text{IE}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \text{IE}[Y])$$

Know

$W \perp Y$ are Jointly Gaussian

$$\text{IE}[W] = 0$$

$$\text{IE}[WY] = 0 \quad (\text{orthogonality})$$

so $\text{Cov}(W, Y) = 0 \Rightarrow$ independence by property 2

Need W is error given by best MSE

$$\text{IE}[Wg(Y)] = 0 \quad \text{if } g(Y) \Rightarrow$$

Trivially true since W independent of Y .

$$\text{IE}[Wg(Y)] = \text{IE}[W] \text{IE}[g(Y)] = 0$$

Proof of ④

Hilbert Space

A Hilbert space is a vector space that

① Has an inner product $\langle \cdot, \cdot \rangle$ defined

② Is complete w.r.t. respect to the norm

$\|X\| = \sqrt{\text{IE}[X^2]}$ induced by $\langle \cdot, \cdot \rangle$

→ complete in the sense every Cauchy sequence converges

Hilbert Space of Random Variables

All random variables w/ finite 2nd moments form a Hilbert space w/ inner product rvs not vectors

$$\langle X, Y \rangle \triangleq \text{IE}[XY]$$

Inner Product \Rightarrow norm/length \Rightarrow distance metric
length/norm of $X = \sqrt{\langle X, X \rangle} = \sqrt{\text{IE}[X^2]}$ ≠ this ≤ 1

$$\text{Angle b/w } X, Y ; \theta_{XY} : \cos(\theta_{XY}) = \frac{\langle X, Y \rangle}{\|X\| \|Y\|} = \frac{\text{IE}[XY]}{\sqrt{\text{IE}[X^2]} \sqrt{\text{IE}[Y^2]}}$$

Projection of X onto Y a random variable

$$\Pi_Y(X) = \frac{\langle X, Y \rangle}{\|Y\|^2} Y = \frac{\text{IE}[XY]}{\text{IE}[Y^2]}$$

Orthogonality $\langle X, Y \rangle \triangleq \text{IE}[XY] = 0$

$$\Pi_Y(X) = 0$$

$$\text{Pythagorean} \quad \text{all } X_i \text{ pairwise orthogonal}$$

$$\text{IE}\left[\left(\sum_{i=1}^m X_i\right)^2\right] = \sum_{i=1}^m \text{IE}[X_i^2]$$

Jointly Gaussian Random Variables

If

$$X \sim N(\mu, \sigma^2)$$

then

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

Note: K is our covariance matrix.

Here, $K = \begin{pmatrix} \text{Cov}(X,X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Cov}(Y,Y) \end{pmatrix}$

$$= \text{IE}\left[\begin{pmatrix} X & Y \end{pmatrix} \begin{pmatrix} X & Y \end{pmatrix}^\top\right]$$

$$= \begin{pmatrix} \text{Cov}(X,X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Cov}(Y,Y) \end{pmatrix}$$

Definition: X, Y are jointly Gaussian if $aX + bY$ is Gaussian if $a, b \in \mathbb{R}$.

Definition: Joint pdf of such a distribution is

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{|K|}} \exp\left(-\frac{1}{2} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}^\top K^{-1} \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}\right)$$

Properties of Jointly Gaussian Random Variables

$$\begin{aligned} \text{① J.G.} &\Rightarrow \text{M.G.} \\ \text{② Uncorrelated JG r.v.'s are independent} &\xrightarrow{\text{special!}} \text{independence} \Rightarrow \text{uncorrelated} \end{aligned}$$

in general

③ If X, Y jointly Gaussian, then

$a_1X + b_1Y + c_1, a_2X + b_2Y + c_2$, are jointly Gaussian

④ For JG X, Y

$$\hat{x}_{\text{MSE}} = \text{IE}[X|Y] = \hat{x}_{\text{MSE}} = \text{IE}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{IE}[Y])$$

$$\text{MSE} = \text{Var}(X)(1 - j_{\hat{x}_{\text{MSE}}}) = \text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}$$

$$\text{⑤ For J.G. } X, Y \quad \text{IE}[X|Y] = \text{IE}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{IE}[Y])$$

$$\text{IE}[X|Y] = \text{IE}[X] + \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} (Y - \text{IE}[Y])$$

$$\text{Var}(X|Y) = \text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}$$

$$\text{Var}(X|Y) = \text{Var}(X) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(Y)}$$

⑥ The conditional distribution of X given $Y=y$ is Gaussian

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\text{Var}(X|Y)}} e^{-\frac{(x-\mu_{X|Y})^2}{2\text{Var}(X|Y)}} \quad \mu_{X|Y} = \text{IE}[X|Y] \quad \sigma^2 = \text{Var}(X|Y)$$

Mean Squared Error (MSE)

Have r.v. X which can't be directly observed. Want to estimate \hat{X}

Note: we are in a Hilbert space of random variables

The estimation error

$$W = X - \hat{X}$$

The MSE of \hat{X} is

$$\text{IE}[(X - \hat{X})^2] = \text{IE}[W^2] = \langle W, W \rangle = \|W\|^2 = (\text{d}(X, \hat{X}))^2$$

Minimum Mean Squared Error

$$\hat{x}_{\text{MSE}} = \underset{X}{\operatorname{argmin}} \text{IE}[(X - \hat{X})^2]$$

MSE of X using a constant

$$\hat{x}_{\text{MSE}} = a^*$$

where

$$a^* = \underset{a}{\operatorname{argmin}} \text{IE}[X - a]^2$$

it's $\text{IE}[X]$!

$$\text{MSE} = \text{IE}[X - a]^2$$

$$= \text{IE}[(X - \text{IE}[X])^2 + (\text{IE}[X] - a)^2]$$

$$= \text{IE}[(X - \text{IE}[X])^2] + 2\text{IE}[(X - \text{IE}[X])(\text{IE}[X] - a)] + \text{IE}[(\text{IE}[X] - a)^2]$$

Want to minimize

$$\text{IE}[(X - \text{IE}[X])^2] + (\text{IE}[X] - a)^2$$

$$\rightarrow \text{Var}(X) + (\text{IE}[X] - a)^2$$

$$\text{Take } a = \text{IE}[X]$$

MSE of X given $Y=y$

similar proof

$$\hat{x}_{\text{MSE}} = \text{IE}[X|Y]$$

$$\text{MSE}(\hat{x}_{\text{MSE}}) = \text{IE}[\text{Var}(X|Y)]$$

The error

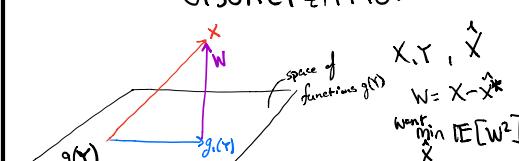
$$W = X - \text{IE}[X|Y]$$

has

$$\text{IE}[W] = \text{IE}[X - \text{IE}[X|Y]] = \text{IE}[X] - \text{IE}[X] = 0$$

we call this type of estimator UNBIASED

VISUALIZATION



- Orthogonality principle:

$$W \perp g(Y) \Leftrightarrow g(Y) \perp g(Y) \quad \text{necessary AND sufficient condition}$$

W induced by \hat{x}^* (best estimator)

$$\text{claim: } X - \text{IE}[X|Y] \perp g(Y) \perp g(Y) \quad \text{inner product } \langle X - \text{IE}[X|Y], g(Y) \rangle = 0$$

$$\text{prof: } \text{IE}[(X - \text{IE}[X|Y]) g(Y)] = 0$$

LHS

$$\text{IE}[X g(Y)] - \text{IE}[(\text{IE}[X|Y]) g(Y)]$$

$$\text{IE}_Y [\text{IE}_{X|Y} [X g(Y) | Y]] - \text{IE}_Y [\text{IE}_{X|Y} [\text{IE}_Y [X|Y] g(Y) | Y]]$$

$$\text{IE}_Y [g(Y) \text{IE}_{X|Y} [X|Y]] - \text{IE}_Y [g(Y) \text{IE}_{X|Y} [X|Y]] = 0 \quad \checkmark$$

MSE Using Pythagorean Theory

$$\|W\|^2 = \langle W, W \rangle \text{ in Hilbert Space}$$

$$\|W\|^2 = \text{IE}[W^2] = \text{IE}[X^2] - \text{IE}[X]^2$$

$$\text{IE}[X] = \text{IE}[X] = \text{Var}(X) - \text{Var}(\hat{x}_{\text{MSE}})$$

$$\text{IE}[X] = \text{IE}[X] = \text{Var}(X) - \text{Var}(\hat{x}_{\text{MSE}})$$

$$\text{IE}[X] = \text{IE}[X] = \text{Var}(X) - \text{Var}(\hat{x}_{\text{MSE}})$$

- \hat{x}_{MSE} is a constant

Thus $X \sim N(\mu_W + \hat{x}_{\text{MSE}}, \sigma^2)$

- W maintains its distribution

Random Vectors
 $\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}$, $E[\mathbf{X}] = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_m] \end{bmatrix}$

Correlation (b/t a random vector w/ itself)

Notice Symmetry
 $E[\mathbf{X} \mathbf{X}^T] = \begin{bmatrix} E[X_1^2] & E[X_1 X_2] & \dots & E[X_1 X_m] \\ E[X_2 X_1] & E[X_2^2] & \dots & E[X_2 X_m] \\ \vdots & \vdots & \ddots & \vdots \\ E[X_m X_1] & E[X_m X_2] & \dots & E[X_m^2] \end{bmatrix}$

Variance of \mathbf{X}

$$E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] = \begin{bmatrix} Var(X_1) & Cov(X_1 X_2) & \dots & Cov(X_1 X_m) \\ Cov(X_1 X_2) & Var(X_2) & \dots & Cov(X_2 X_m) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1 X_m) & Cov(X_2 X_m) & \dots & Var(X_m) \end{bmatrix}$$

K

Covariance matrix

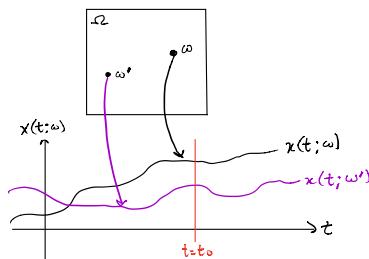
Cross-Correlation (b/t two random vectors)

$$E[\mathbf{X} \mathbf{Y}^T] = \left\{ E[X_i Y_j] \right\}_{1 \leq i \leq m, 1 \leq j \leq n}$$

$$\text{Cov}(\mathbf{X}, \mathbf{Y}) = \left\{ \text{Cov}(X_i, Y_j) \right\}_{1 \leq i \leq m, 1 \leq j \leq n}$$

Random Process

Each sample $\omega \in \Omega$ in the sample space is mapped to a time function $X(t, \omega)$



Sampling a random process at a particular time t_0 , you get a random variable $X(t_0, \omega_1), X(t_0, \omega)$ random variables

$X(t, \omega)$ is a deterministic function of t , called a realization or a sample path of this random process

$X(t) \rightarrow \text{s.p.}$ $x(t) \rightarrow \text{sample path}$

Stationarity: A random process is stationary if its statistical characteristics do not change over time.

Specifically,

$$\forall \Delta \quad F_{X(t_1), \dots, X(t_n)}(x_1, \dots, x_n) = F_{X(t_1+\Delta), \dots, X(t_n+\Delta)}(x_1, \dots, x_n)$$

Wide Sense Stationary: A random process is WSS if

$$(i) E[X(t)] = \mu$$

$$(ii) R_X(t_1, t_2) = E[X(t_1)X(t_2)] = R(t) \quad \text{where } \tau = t_2 - t_1$$

Independence Review

Independent Events: $A \perp B$ if $P[A \cap B] = P[A]P[B]$
 or equivalently: $P[A|B] = P[A]$

Independent Random Variables:

$X \perp\!\!\!\perp Y$ if $f_{XY}(x,y) = f_X(x)f_Y(y)$

or equivalently,

$$f_{XY}(x,y) = f_X(x)$$

Conditional Independent Events

$A \perp\!\!\!\perp B$ conditioned on C if

$$P[A \cap B | C] = P[A|C]P[B|C]$$

or equivalently,

$$P[A | B \cap C] = P[A | C]$$

Conditional Independent Random Variables:

$X \perp\!\!\!\perp Y$ if $f_{XY|Z}(x,y|z) = f_X(x|z)f_Y(y|z)$

or equivalently, $f_{XYZ}(x,y,z) = f_X(x|z)f_Y(y|z)$

Increment: The increment of a r.p. $\{X(t)\}$ over an interval $[a, b]$ is the r.v. $\underbrace{X(b) - X(a)}_{\text{another r.v.}}$

Independent Increments: $\{X(t)\}$ has independent increments if $\forall n$, and for all $t_0 < t_1 < \dots < t_n$ the increments $X(t_1 - t_0), \dots, X(t_n - t_0)$ are independent

Stationary Increments

$\{X(t)\}$ has stationary increments if the distribution of $X(t + \tau) - X(t)$ depends only on τ , not t .

Key Idea: When characterizing processes w/ independent increments over non-overlapping intervals

Counting Function

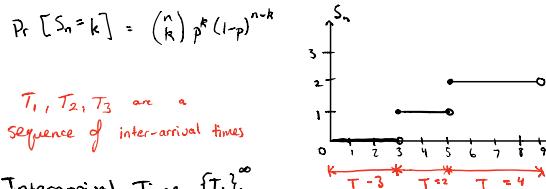
$f(t) \quad (t \geq 0)$ is a counting function if $f(0) = 0$,
 $f(t)$ takes non-negative integers, is non-decreasing, and is right continuous.

Binomial Counting Process

$$X_i \sim \text{Bernoulli}(p)$$

$$S_n = \sum_{i=1}^n X_i$$

$$S_n \sim \text{Binomial}(n, p)$$



Interarrival Time $\{T_i\}_{i=1}^\infty$

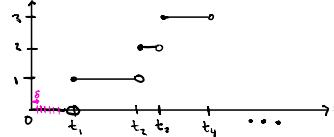
In above example,

$$T_i \sim \text{Geometric}(p)$$

$$\Pr[T_1 = k] = (1-p)^{k-1} p$$

What about continuous counting processes?

Say events arrive continuously w/ rate λ avg arrival/unit time



Partition time into δ -length intervals.

Assumptions:

For $\delta \rightarrow 0$

- ① The probability of having more than one arrivals in δ -interval is negligible ($\delta \rightarrow 0$)
- ② Whether there is an arrival within a δ -interval is independent of arrivals in other δ -intervals

$$\{P = \lambda \delta\}$$

These assumptions make this similar to the binomial discrete case w/ $p = \lambda \delta$

$\delta \rightarrow 0$

$$p = \lambda \delta \rightarrow 0$$

$$n = \frac{t}{\delta} \rightarrow \infty$$

$$S_n \sim \text{Binomial}\left(\frac{t}{\delta}, \lambda \delta\right) \rightarrow \text{Poisson}(nt)$$

Properties

① If JG, Uncorrelated \Rightarrow independence NOT generally true

$$\text{Cov}(X, Y) = 0 \quad \text{joint distribution factors into product of marginal distributions}$$

② If you have two rvs which are marginally Gaussian AND they're independent, can conclude they are jointly Gaussian

③ If XY jointly Gaussian then XM is Gaussian USEFUL

$$\text{mean: } E[XY] = \bar{X} \bar{Y}_{\text{mean}} = E[X]E[Y] = \frac{1}{2} (E[X^2] + E[Y^2])$$

$$\text{variance: } \text{Var}(XY) = \text{Var}(X) + \text{Var}(Y) - 2 \text{Cov}(X, Y)$$

$$= \text{Var}(X)(1 - \rho_{XY}) + \text{Var}(Y)(1 - \rho_{XY})$$

in general, a rv next variable being conditioned on is a CONSTANT

Gaussian Random Processes

Definition:

$X(t)$ is Gaussian if $\forall n$, and $\forall t_1, t_2, \dots, t_n$

$X(t_1), \dots, X(t_n)$ are JOINTLY Gaussian

Properties Completely specified by first two moments

- ① completely specified by its mean function $\mu(t)$ and autocorrelation function $R_X(t)$ /autocovariance function $C_X(t)$.

Example

$$f_{X(t_1), X(t_2)}(\cdot, \cdot) \sim N\left(\begin{bmatrix} \mu(t_1) \\ \mu(t_2) \end{bmatrix}, \begin{bmatrix} C_{XX}(t_1, t_1) & C_{XX}(t_1, t_2) \\ C_{XX}(t_2, t_1) & C_{XX}(t_2, t_2) \end{bmatrix}\right) \quad \text{"BROAD RIGHT VALUES!"}$$

$$\textcircled{2} \text{ WSS} \Rightarrow S$$

Definition: A Brownian motion (also called a Wiener Process) with parameter σ^2 is a random process $\{X(t)\}_{t \geq 0}$ such that

- ① $X(0) = 0$
- ② $\{X(t)\}_{t \geq 0}$ has independent increments
- ③ $X(t) - X(t_1) \sim N(0, \sigma^2(t-t_1))$
- ④ Every sample path is continuous

Properties of Brownian Motion

$$\textcircled{1} \quad \mu(t) \triangleq E[X(t)] = E[X(t)-X(0)] = 0$$

$$\begin{aligned} \textcircled{2} \quad R_X(t_1, t_2) &= C_{XX}(t_1, t_2) \quad t_2 > t_1 \\ &= E[(X(t_1) - X(t_0))(X(t_2) - X(t_0))] \\ &= E[(X(t_1))^2] + E[X(t_1)]E[X(t_2)] - E[X(t_1) - X(t_0)]E[X(t_2) - X(t_0)] \\ &= E[(X(t_1) - X(t_0))^2] + E[X(t_1) - X(t_0)]E[X(t_2) - X(t_0)] \\ &= \sigma^2 t, \end{aligned}$$

In general,

$$C_{XX}(t_1, t_2) = R_X(t_1, t_2) = \min(t_1, t_2) \sigma^2$$

$$\textcircled{3} \quad X(t) \sim N(0, \sigma^2 t)$$

Also have $X(t_1), X(t_2), \dots, X(t_n)$ JG & uncorrelated

$$\begin{bmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_n) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} X(t_1) - X(t_0) \\ X(t_2) - X(t_0) \\ \vdots \\ X(t_n) - X(t_0) \end{bmatrix}$$

Markov Process

A random process $\{X(t)\}_{t \geq 0}$ is a Markov process if $\forall n$, and t, t_1, t_2, \dots, t_n

$$\begin{array}{c} X(t_0) \\ X(t_1) \\ X(t_2) \\ \vdots \\ X(t_n) \end{array} \rightarrow \begin{array}{c} t_0 \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{array}$$

$$\Pr[X(t_{n+1}) = x_{n+1} | X(t_0) = x_0, X(t_1) = x_1, \dots, X(t_n) = x_n] = \Pr[X(t_{n+1}) = x_{n+1} | X(t_n) = x_n]$$

i.e. conditioned on where you are now, the future is independent of the past.

We call the value we see the state.

Once you know the current state, future state DEPENDS on past states

State: $X(t)$ is called the state at time t .

All the values $X(t)$ can take is called the state-space \mathcal{X}

If \mathcal{X} is discrete, then we call this a Markov chain.

Jointly Gaussian

$$X \sim \mathcal{N}(\mu, \sigma^2) \text{ if } f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E[X] = \mu$$

$$\text{Var}(X) = \sigma^2$$

X and Y are jointly Gaussian if

Z = aX + bY is Gaussian $\forall a, b \in \mathbb{R}$

For jointly Gaussian X and Y w/ $|J_{XY}| < 1$

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{|J_{XY}|}} e^{-\frac{1}{2} \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{1}{2} \frac{(y-\mu_Y)^2}{\sigma_Y^2}}$$

where

$$K = \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Cov}(Y, Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho_{XY} \sigma_X \sigma_Y \\ \rho_{XY} \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix}$$

$$K^{-1} = \frac{1}{|K|} \begin{bmatrix} \sigma_X^2 & -\rho_{XY} \sigma_X \sigma_Y \\ -\rho_{XY} \sigma_X \sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Characterization of Discrete-Time Markov Chain:

A discrete-time Markov chain $\{X_n\}_{n \geq 0}$ is characterized by

- the PMF $p(i)$ of its initial state X_0 :

$$p(i) \triangleq \Pr_{i \in \mathcal{X}}[X_0 = i]$$

- and one-step transition probabilities:

$$\Pr_{i,j \in \mathcal{X}}[X_{n+1} = j | X_n = i]$$

Remark: the joint PMF is given by

$$\Pr[X_0 = i_0, \dots, X_n = i_n] = p(i_0) \Pr[X_1 = i_1 | X_0 = i_0] \dots \Pr[X_n = i_n | X_{n-1} = i_{n-1}]$$

i.e. for Markov process only need the marginal distribution of the initial state and the transition probabilities from one state to the next.

Need $p(i_0)$ and $\{\Pr[X_{n+1} = j | X_n = i]\}_{i,j \in \mathcal{X}}$

Homogeneous Markov Chain \rightarrow Time INVARIANT

$$\Pr[X_{n+1} = j | X_n = i] \stackrel{\text{doesn't depend on } n}{=} p_{i,j}$$

Transition Matrix:

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & \dots & p_{1,M} \\ p_{2,1} & p_{2,2} & \dots & p_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M,1} & p_{M,2} & \dots & p_{M,M} \end{bmatrix} \quad \sum_j p_{i,j} = 1 \quad i \in \mathcal{I}$$

n -Step Transition Matrix

$$\mathbf{P}^{(n)} = \left\{ p_{i,j} \triangleq \Pr[X_n = j | X_0 = i] \right\}$$

Chapman-Korogorov Equation

$$p_{i,j}^{(n+1)} = \sum_{k \in \mathcal{X}} p_{i,k}^{(n)} p_{k,j}^{(n)}$$

$$\begin{aligned} \mathbf{P}^{(m+n)} &= \mathbf{P}^{(m)} \mathbf{P}^{(n)} \\ \mathbf{P}^{(n)} &= \mathbf{P} \mathbf{P}^{(n-1)} \\ &= \mathbf{P} \mathbf{P} \mathbf{P}^{(n-2)} \\ &\vdots \\ &= \mathbf{P}^n \end{aligned}$$

State Probability

This is the distribution at time n .

$$\mathbf{p}(n) = [\Pr[X_1 = 1] \ \dots \ \Pr[X_n = m]]^\top$$

$$\mathbf{p}(0) = [\Pr[X_0 = 0] \ \dots \ \Pr[X_0 = m]]^\top$$

$$\mathbf{p}(1) = \mathbf{p}(0) \mathbf{P}$$

$$\vdots$$

$$\mathbf{p}(n) = \mathbf{p}(0) \mathbf{P}^n$$

$$\{X_n\}_{n \geq 0} : \hat{p}(0) = \Pr[X_0 = 1, \dots, X_n = m]$$

$$\begin{cases} \text{pmf of } X_0 \\ \mathbf{P} \end{cases}$$

row vector by definition

$$\hat{p}(n) \triangleq \Pr[X_n = 1, \dots, X_n = m]$$

$$\begin{aligned} \Pr[X_n = j] &= \sum_{i=1}^m \Pr[X_n = i] \Pr[X_n = j | X_0 = i] \\ &= \hat{p}(0) \mathbf{p}^{(n)} \end{aligned}$$

Markov Process and Independent Increments:

- If $\{X(t)\}_{t \geq 0}$ has independent increments and $X(0) = c$ (a constant), then $\{X(t)\}_{t \geq 0}$ is a Markov process.

- The converse is not true: a Markov process may have dependent increments.

Examples:

- Random walk is a discrete-time Markov chain with $\mathcal{X} = \mathbb{Z}$.
- Brownian motion is a continuous-time Markov process with $\mathcal{X} = \mathbb{R}$.
- Poisson process is a continuous-time Markov chain with $\mathcal{X} = \mathbb{Z}^+$.

Gaussian Random Vectors

\mathbf{X} is a Gaussian random vector if its coordinates are JOINTLY GAUSSIAN

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix} \quad \text{a Gaussian random vector} \Leftrightarrow \mathbf{a}_i^\top \mathbf{X} + b_i \text{ is Gaussian} \forall i \in \mathbb{R}, \mathbf{a}_i \in \mathbb{R}^m$$

Use $\mathbf{X} \sim N(\mu, K)$ to denote Gaussian random vector

If $\mathbf{X} \sim N(\mu, K)$, then

- ① Any subvector of \mathbf{X} is a Gaussian random vector

$$- X_i \sim N(\mu_i, K_{ii})$$

$$- \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_j \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_j \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1j} \\ K_{21} & K_{22} & \dots & K_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ K_{j1} & K_{j2} & \dots & K_{jj} \end{bmatrix}\right)$$

② $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

is a Gaussian random vector

$$\mathbb{E}[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\top]$$

$$\mathbf{Y} \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}K\mathbf{A}^\top)$$

- ③ If K is diagonal, every rv in \mathbf{X} is independent of every other rv in \mathbf{X}

- ④ If \mathbf{X}_{rv} and \mathbf{Y}_{rv} are jointly Gaussian then they are independent iff $\text{Cov}(\mathbf{X}, \mathbf{Y}) = 0$

MMSE Estimate of \mathbf{X} using \mathbf{Y}

$$\begin{aligned} \text{MMSE} \quad \mathbb{E}[||\mathbf{X} - \hat{\mathbf{X}}||^2] &= \mathbb{E}\left[\left(\mathbf{X} - \hat{\mathbf{X}}\right)^\top \left(\mathbf{X} - \hat{\mathbf{X}}\right)\right] \\ &= \sum_{i=1}^m \mathbb{E}[X_i - \hat{X}_i]^2 \end{aligned}$$

Need to "design"

$$\hat{\mathbf{X}} = g(\mathbf{Y}) = \begin{bmatrix} g_1(\mathbf{Y}) \\ g_2(\mathbf{Y}) \\ \vdots \\ g_m(\mathbf{Y}) \end{bmatrix}$$

Problem decoupled into m independent MMSE estimation problems, one for each X_i .

$$\hat{X}_i = g_i(\mathbf{Y})$$

Our best MMSE estimator for X_i is

$$\mathbb{E}[X_i | \mathbf{Y}]$$

Need conditional distribution

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

LMMSE Estimator of \mathbf{X}

Similar to rv case,

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{A}\mathbf{Y} + \mathbf{b}_{\text{rv}} \\ &= \mathbb{E}[\mathbf{X}] + \text{Cov}(\mathbf{X}, \mathbf{Y}) \text{Cov}^{-1}(\mathbf{Y})(\mathbf{Y} - \mathbb{E}[\mathbf{Y}]) \end{aligned}$$

If \mathbf{X}, \mathbf{Y} are jointly Gaussian, (similar to rv case)

$$\mathbb{E}[\mathbf{X} | \mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \text{Cov}(\mathbf{X}, \mathbf{Y}) \text{Cov}^{-1}(\mathbf{Y})(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])$$

Poisson Process:

Definition: $\{N(t)\}_{t \geq 0}$ is Poisson process w/ rate λ if it is a counting process w/ independent increments and $N(t) - N(s) \sim \text{Pois}(\lambda(t-s)) \quad \forall t > s$.

Interarrival Time (note: T_i iid τ_i)

Look at T_1

$$\Pr[T_1 > t] = (1 - e^{-\lambda})^t$$

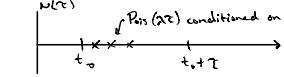
$$\lim_{t \rightarrow \infty} (1 - e^{-\lambda})^t = e^{-\lambda t}$$

$$\text{i.e. } T_1 \sim \exp(\lambda) \quad \mathbb{E}[T_1] = \frac{1}{\lambda}$$

leads to ANOTHER definition

Definition: $\{N(t)\}_{t \geq 0}$ is a counting process w/ rate λ if it is a counting process w/ interarrival times iid exponentially distributed

Definition: $\{N(t)\}_{t \geq 0}$ is a counting process w/ rate λ if it is a counting process such that $\forall T > 0$, $N(T) \sim \text{Pois}(\lambda T)$, and given $N(\tau) = n$ (n arrival times in interval T), are iid $U[0, T]$



Facts about Poisson Process

$$\{N(t)\}_{t \geq 0} \text{ w/ rate } \lambda \rightarrow \text{arrival rate/unit time} \quad \frac{\# \text{arrivals}}{\text{time}} = \frac{n}{T} \xrightarrow{T \rightarrow \infty} \lambda$$

① Process has independent increments

$$N(t_2) - N(t_1) \perp \! \! \! \perp N(t_3) - N(t_2)$$

length doesn't matter on interval, they just need NOT overlap

② #arrivals in interval $[t_1, t_2] = N(t_2) - N(t_1)$ where

$$N(t_2) - N(t_1) \sim \text{Pois}(\lambda(t_2 - t_1))$$

③ The interarrival times $\{T_i\}_{i=1}^\infty$ are iid where

$$T_i \sim \exp(\lambda)$$

④ Given $N(\tau) = k$, these k arrivals are iid $U[0, \tau]$ i.e. $t_i \sim U[0, k\tau]$

⑤ $\mu_n(t) = \mathbb{E}[N(t)] = \lambda t$ i.e. arrival rate * time
 $\text{Var}(N(t)) = \lambda t$ (Poisson property)

$$R_n(s, t) \triangleq \mathbb{E}[N(s)N(t)] \quad \text{assume } s \leq t \quad \text{have var.}$$

$$= \mathbb{E}[N(s)(N(s) + N(t) - N(s))]$$

$$= \mathbb{E}[N(s)^2] + \mathbb{E}[N(s)(N(t) - N(s))] \quad \text{independent increments}$$

$$= \lambda s + (\lambda s)^2 + \mathbb{E}[N(s)] \mathbb{E}[N(t) - N(s)]$$

$$= \lambda s + (\lambda s)^2 + \lambda s(\lambda(t-s)) \quad \text{if } t-s \leq s \quad \text{if } t-s > s$$

$$C_n(s, t) = \lambda \min(s, t) \quad \text{← applies } \perp \! \! \! \perp \text{ s.t. (note above we assumed s < t)}$$

Stationary Markov Chain:

Let π be a stationary distribution of a Markov chain $\{X_n\}_{n=0}^\infty$. If X_0 is distributed according to π , then $\{X_n\}_{n=0}^\infty$ is a stationary process.

Interpretations of Stationary Distribution

Long-term occupancy rate:

For a Markov chain with $p(i)$ given by a stationary distribution π , we have $\Pr(X_n = i) = \pi_i$ for all n . Thus, π_i can be interpreted as the long-term fraction of time the chain spends in state i . For example of Coin A and Coin B, the stationary distribution is $\pi = \begin{bmatrix} 0.4 & 0.6 \end{bmatrix}$, which can be interpreted as seeing a "head" 40% of the time and a "tail" 60% of the time.

Stationary Distribution of Markov Chain

Stationary Distribution:

Let $\pi = \{\pi_i, i \in \mathcal{X}\}$ be a probability distribution (a row vector). It is called a **stationary distribution** of a Markov chain with transition matrix P if

$$\pi = \pi P$$

Computing Stationary Distribution:

$$\begin{cases} \pi \\ \sum_i \pi_i = 1 \end{cases}$$



Recurrent Class

Communication Class and Irreducibility

Accessible and Communicating:

- If, for some $n \geq 0$, $p_{i,j}^{(n)} > 0$, we say that j is **accessible** from i and write $i \rightarrow j$.
- If $i \rightarrow j$ and $j \rightarrow i$, we say that i and j **communicate** and write $i \leftrightarrow j$.
- "Communication" is an **equivalence relation**, i.e.,
 - reflexive**: $i \leftrightarrow i$
 - symmetric**: $i \leftrightarrow j$ iff $j \leftrightarrow i$
 - transitive**: if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.

Decomposition of State Space into Communication Classes:

The state space \mathcal{X} can be decomposed into disjoint exhaustive **communication classes**. First put state 1 and all states communicating with 1 in a class C_1 . Then pick a state i in $\mathcal{X} \setminus C_1$. Put i and all states communicating with i into another class C_2 . Continue this process until all states have been assigned.

Periodicity

Period of State:

Let $\mathcal{N}_i \triangleq \{n \geq 1 : p_{i,i}^{(n)} > 0\}$. The **period** $d(i)$ of state i is defined as

$$d(i) \triangleq \begin{cases} \gcd\{\mathcal{N}_i\}, & \text{if } \mathcal{N}_i \neq \emptyset \\ 1 & \text{otherwise} \end{cases}$$

If $d(i) = 1$, state i is said to be **aperiodic**. If $d(i) > 1$, state i is periodic with period $d(i)$.

Remark: If $p_{i,i}^{(n)} > 0$, then n is an integer multiple of $d(i)$, and $d(i)$ is the largest integer with this property. Returns to state i are only possible via paths whose lengths are multiples of $d(i)$.

Period of a Class

All states in a communication class have the same period, also called the class period.

Sufficient Conditions for Aperiodicity

Either of the following is a sufficient condition for an irreducible Markov chain to be aperiodic

- $\exists i \in \mathcal{X}$, s.t. $p_{i,i} > 0$ (**self loop**)
- $\exists n > 0$, s.t. $P^n > 0$ (**common path length for all state pairs**).

Transience and Recurrence

Probability of Return:

- $f_i^{(n)} \triangleq \Pr\{X_1 \neq i, \dots, X_{n-1} \neq i, X_n = i \mid X_0 = i\}$ is the probability of returning, for the first time, to state i after n steps.
- $f_i \triangleq \sum_{n=1}^{\infty} f_i^{(n)}$ is the probability of ever returning to state i .

Recurrence and Transience:

- State i is **recurrent** if $f_i = 1$.
- State i is **transient** if $f_i < 1$.

The Number of Returns:

$$N_i \triangleq \sum_{n=1}^{\infty} \mathbb{1}_{[X_n=i \mid X_0=i]}$$

is the number of times the chain returns to state i . We have

$$\begin{cases} \Pr\{N_i = \infty\} = 1, & \text{if } i \text{ recurrent} \\ \mathbb{E}[N_i] = \frac{f_i}{1-f_i} < \infty, & \text{if } i \text{ transient} \end{cases}$$

i.e., if i is recurrent, the chain returns to i infinitely often. Otherwise, the chain visits i only a finite number of times, and the expected number of visits to i is $\frac{f_i}{1-f_i}$.

Criteria for Recurrence:

- i is recurrent iff $f_i = 1$.
- i is recurrent iff $\sum_{n=1}^{\infty} p_{i,i}^{(n)} > 0$.
- If $i \leftrightarrow j$, then i is recurrent iff j is recurrent.
- The states of a finite-state, irreducible Markov chain are all recurrent.

Positive Recurrence and Null Recurrence

Return Time:

Let $\{X_n\}_{n=0}^{\infty}$ be a Markov chain. Define, for $i \in \mathcal{X}$,

$$T_i \triangleq \min\{n \geq 1 : X_n = i \mid X_0 = i\}.$$

Remark: T_i tells us how long it takes for a chain, started at state i , to return to i for the first time.

T_i is a random variable with PMF given by $[f_i^{(1)}, f_i^{(2)}, \dots]$.

Positive Recurrence and Null Recurrence:

A recurrent state i is **positive recurrent** if $\mathbb{E}[T_i] < \infty$ and **null recurrent** otherwise.

Example: For the simple random walk on \mathbb{Z} with $p = \frac{1}{2}$, it can be shown that $f_i^{(2n)} \sim \frac{C}{n^{3/2}}$. While $f_i = 1$, the expected return time

$$\mathbb{E}[T_i] = \sum_{n=1}^{\infty} 2n f_i^{(2n)} \sim \sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty$$

Thus, every state is null recurrent.

Occupancy Rate:

Define the occupancy rate r_i of state i (i.e., the long-run fraction of time the chain spent in state i) as

$$r_i = \lim_{n \rightarrow \infty} \frac{\sum_{m=1}^n \mathbb{1}_{[X_m=i \mid X_0=i]}}{n}$$

• If i is transient, the expected number of visits to i is finite, thus $r_i = 0$.

• If i is null recurrent, the expected number of visits to i is infinite but grows sublinearly with n , thus $r_i = 0$.

• If i is positive recurrent, $r_i = \frac{1}{\mathbb{E}[T_i]} > 0$.

Recurrent Class

Recurrent Class:

A communication class is a **positive (null) recurrent class** if one of its members is positive (null) recurrent.

Positive recurrence, null recurrence, and transience are class properties, i.e., if one state in a communication class has the property, all states in this class have the property.

Closed Class:

A communication class \mathcal{C} is closed if for all $i \in \mathcal{C}$, $j \notin \mathcal{C}$, we have $p_{i,j} = 0$.

Remark: You can go into a closed set, but you can not go out once you are in.

Every recurrent class is closed.

Proof: Prove by contradiction. Assume there exists $i \in \mathcal{C}$, $j \notin \mathcal{C}$ with $p_{i,j} > 0$. Since \mathcal{C} is recurrent, the chain must be able to come back to \mathcal{C} once leaving \mathcal{C} by transitioning from i to j . In other words, there must exist $k \in \mathcal{C}$ with $p_{j,k} > 0$. Since \mathcal{C} is a communication class, $p_{j,k} > 0$ implies $j \rightarrow k$. Together with $p_{k,j} > 0$, we conclude that j communicates with i . This contradicts with $j \notin \mathcal{C}$.

Canonical Decomposition

Canonical Decomposition:

The state space \mathcal{X} of a Markov chain can be decomposed as

$$\mathcal{X} = \mathcal{T} \cup \mathcal{C}_1 \cup \mathcal{C}_2 \cup \dots,$$

where \mathcal{T} consists of transient states (\mathcal{T} is not necessarily one communicating class), $\{\mathcal{C}_i\}$ are closed, disjoint communication classes of recurrent states. If we relabel the states so that the states in each class have consecutive labels with states in \mathcal{C}_1 having the smallest indexes, then the transition matrix \mathbf{P} can be rewritten as

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{P}_3 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & & 0 \\ \mathbf{R}_1 & \mathbf{R}_2 & \mathbf{R}_3 & & \cdots & \mathbf{Q} \end{pmatrix}$$

where \mathbf{P}_i is a square stochastic matrix that governs the transitions within \mathcal{C}_i . Transitions from states in \mathcal{T} to states in \mathcal{C}_i are governed by \mathbf{R}_i . Transitions among states in \mathcal{T} are governed by \mathbf{Q} .

Existence of Stationary Distribution

The number of stationary distributions for a discrete-time Markov chain takes three possible values: 0, 1, or ∞ :

- If all states are transient or null recurrent, then the chain has no stationary distribution.
- If the chain has a single recurrent class and the recurrent class is positive recurrent, then the chain has a unique stationary distribution $\{\pi_i\}$ given by

$$\pi_i = \begin{cases} \frac{1}{\mathbb{E}[T_i]} & \text{if } i \text{ is positive recurrent} \\ 0 & \text{if } i \text{ is transient} \end{cases}$$

- If the chain has multiple positive recurrent classes, then it has an infinite number of stationary distributions. Specifically, let the states be ordered according to the canonical form with the positive recurrent classes be the first K recurrent classes. Let $\pi^{(k)}$ be the stationary distribution for the k th positive recurrent class. Then for any

$$0 \leq \alpha_1, \dots, \alpha_K \leq 1 \text{ with } \sum_{k=1}^K \alpha_k = 1,$$

$$(\alpha_1 \pi^{(1)}, \alpha_2 \pi^{(2)}, \dots, \alpha_K \pi^{(K)}, 0, \dots, 0)$$

is a stationary distribution.

Absorption Probability

- Modify original Markov chain by adding absorbing states $\sigma, \chi+1$ corresponding to winning "game".
- Define the conditional absorption probability

$$g_i \triangleq \Pr[\text{Absorbed at state } \sigma \mid \text{current state } i]$$

- Use Markovian property to get system of equations.

$$\cdot \mathbb{E}[g_i] \rightarrow "A \text{ wins"} \quad 1 - \mathbb{E}[g_i] \rightarrow "B \text{ wins"}$$

Absorption Time

Define the conditional expected absorption time

$$e_i \triangleq \mathbb{E}[\text{Remaining time until } \sigma \mid \text{currently at state } i]$$

Based on Markovian property, get system of equations

- Expected absorption time is $\mathbb{E}[e_i]$ for non-absorbed

Continuous-Time Markov Chain

Continuous-Time Markov Chain:

A continuous-time Markov chain $\{X(t)\}_{t \geq 0}$ satisfies

- the Markov property: $\forall t_0 < t_1 < \dots < t_{n+1}$,

$$\Pr[X(t_{n+1}) = j \mid X(t_n) = i, X(t_{n-1}) = i_{n-1}, \dots, X(t_0) = i_0] = \Pr[X(t_{n+1}) = j \mid X(t_n) = i]$$

- the state space \mathcal{X} is discrete.

Homogeneity:

A continuous-time Markov chain $\{X(t)\}_{t \geq 0}$ is **homogeneous** (time invariant) if

$$\Pr[X(t_{n+1}) = j \mid X(t_n) = i] = p_{i,j}(t_{n+1} - t_n), \quad \forall t_{n+1}, t_n$$

or equivalently,

$$\Pr[X(s+t) = j \mid X(s) = i] = p_{i,j}(t), \quad \forall s, t$$

Transition Matrix and Stationary Distribution

Transition Matrix:

The transition matrix for a time period of length t :

$$\mathbf{P}(t) = \left\{ p_{i,j}(t) \triangleq \Pr[X(t) = j \mid X(0) = i] \right\}_{i,j \in \mathcal{X}}$$

Chapman-Kolmogorov Equations:

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s), \quad \text{equivalently, } p_{i,j}(t+s) = \sum_{k \in \mathcal{X}} p_{i,k}(t)p_{k,j}(s)$$

State Probability at time t :

Let $\mathbf{p}(t) \triangleq \left\{ \Pr[X(t) = i] \right\}_{i \in \mathcal{X}}$ be a row vector of the state probabilities at time t .

$$\mathbf{p}(t) = \mathbf{p}(0)\mathbf{P}(t)$$

Stationary Distribution:

A probability distribution π over \mathcal{X} is a **stationary distribution** if

$$\pi = \pi\mathbf{P}(t), \quad \forall t$$

Question: How to characterize $\mathbf{P}(t)$ for all t ? Is there a single matrix that fully characterizes $\{\mathbf{X}(t)\}_{t \geq 0}$, similar to the one-step transition matrix for DTMC?

Characterizations of CTMC

Characterization of CTMC:

A CTMC can be characterized by

- the **holding time** $T_i \sim \exp(\lambda_i)$ in each state i (the exponential distribution is due to the Markov property that dictates memoryless holding time);
- an **embedded DTMC** with transition probabilities $p_{i,j}$ ($p_{i,i} = 0$ for all i) that governs the state transition when the holding time in the current state is up.

Remarks: The dynamics of a CTMC can be viewed as follows. Each time a state, say i , is entered, a holding time $T_i \sim \exp(\lambda_i)$ is selected. When the holding time is up, the next state j is selected according to the embedded DTMC $\{p_{i,j}\}$.

An Equivalent Characterization:

Based on properties of exponential distribution, an equivalent characterization of a CTMC with $\{\lambda_i\}_{i \in \mathcal{X}}$ and $\{p_{i,j}\}$ is as follows.

- Each time a state, say i , is entered, a transition time $T_{i,j} \sim \exp(\lambda_i p_{i,j})$ is selected for each state $j \neq i$.
- When the minimum transition time $\min_j\{T_{i,j}\}$ is up, transit to the state k with the minimum transition time, i.e., $T_{i,k} = \min_j\{T_{i,j}\}$.

Characterizations of CTMC

Characterization of CTMC:

A CTMC can be characterized by

- the **holding time** $T_i \sim \exp(\lambda_i)$ in each state i (the exponential distribution is due to the Markov property that dictates memoryless holding time);
- an **embedded DTMC** with transition probabilities $p_{i,j}$ ($p_{i,i} = 0$ for all i) that governs the state transition when the holding time in the current state is up.

Remarks: The dynamics of a CTMC can be viewed as follows. Each time a state, say i , is entered, a holding time $T_i \sim \exp(\lambda_i)$ is selected. When the holding time is up, the next state j is selected according to the embedded DTMC $\{p_{i,j}\}$.

An Equivalent Characterization:

Based on properties of exponential distribution, an equivalent characterization of a CTMC with $\{\lambda_i\}_{i \in \mathcal{X}}$ and $\{p_{i,j}\}$ is as follows.

- Each time a state, say i , is entered, a transition time $T_{i,j} \sim \exp(\lambda_i p_{i,j})$ is selected for each state $j \neq i$.
- When the minimum transition time $\min_j\{T_{i,j}\}$ is up, transit to the state k with the minimum transition time, i.e., $T_{i,k} = \min_j\{T_{i,j}\}$.

Transition Rate Matrix

Transition Rate Matrix Q:

Based on the equivalent characterization, a CTMC is fully characterized by the **transition rate matrix** $\mathbf{Q} = \{q_{i,j}\}_{i,j \in \mathcal{X}}$ where

- $\square q_{i,j} = \lambda_i p_{i,j}$ for $i \neq j$ is the rate of transitioning from state i to state j .
- $\square q_{i,i} = -\sum_{j \neq i} q_{i,j}$ is the rate of leaving state i (the negative sign indicates the direction of leaving); note that $-q_{i,i} = \lambda_i$ is the holding time parameter.
- \square each row of \mathbf{Q} sums up to 0.

From Q to P(t):

$$\mathbf{P}'(t) = \mathbf{PQ}$$

with initial condition $\mathbf{P}(0) = \mathbf{I}$. Solving this differential equation leads to

$$\mathbf{P}(t) = e^{t\mathbf{Q}} = \sum_{k=0}^{\infty} \frac{(t\mathbf{Q})^k}{k!}$$

Proof: to show that \mathbf{P} satisfies the above differential equation, we first show that $\mathbf{Q} = \mathbf{P}'(0)$ and then apply Chapman-Kolmogorov equations to $\mathbf{P}(t + \delta)$ when evaluating $\mathbf{P}'(t)$.

Stationary Distribution

Obtaining Stationary Distribution:

The stationary distribution π can be obtained by solving the following linear equations:

$$\begin{cases} \pi \mathbf{Q} = 0 \\ \sum_{i \in \mathcal{X}} \pi_i = 1 \end{cases}$$

Proof: Take derivative on both sides of $\pi = \pi \mathbf{P}(t)$ and plug in the differential equation for $\mathbf{P}(t)$, we have

$$\pi \mathbf{P}'(t) = 0 \Rightarrow \underbrace{\pi \mathbf{P}(t)}_{=\pi} \mathbf{Q} = 0 \Rightarrow \pi \mathbf{Q} = 0$$

Global Balance Equations:

$\pi \mathbf{Q} = 0$ can be written as

$$\pi_i q_{i,i} = \sum_{j \neq i} \pi_j q_{j,i}, \quad \forall i$$

which are referred to as the **global balance equations**. They state that the rate of probability flow out of state i (the left-hand side) is equal to the rate of flow into state i (the right-hand side). More generally, the net flow through any closed loop must be zero for the chain to be in equilibrium. Based on this, you can create your own set of balance equations to solve for the stationary distribution.

WSS and Correlation Functions

Wide Sense Stationarity (WSS):

A random process is **wide sense stationary (WSS)** if

- (i) $\mathbb{E}[X(t)] = m$
- (ii) $R_X(t_1, t_2) \triangleq \mathbb{E}[X(t_1)X(t_2)] = R_X(t_1 - t_2)$
- (iii) $R_X(\tau), \text{ where } \tau = t_1 - t_2$

Properties of the Autocorrelation function $R_X(\tau)$:

Let $R_X(\tau)$ be the autocorrelation function of a zero-mean WSS random process. Then

1. $R_X(0) = \mathbb{E}[X^2(t)]$ is the average power of $X(t)$.
2. $R_X(\tau)$ is even: $R_X(-\tau) = R_X(\tau)$.
3. $R_X(0) \geq |R_X(\tau)|$.

Joint WSS:

Two random processes $\{X(t)\}_{t=-\infty}^{\infty}$ and $\{Y(t)\}_{t=-\infty}^{\infty}$ are **jointly wide sense stationary (WSS)** if

- (i) both $\{X(t)\}_{t=-\infty}^{\infty}$ and $\{Y(t)\}_{t=-\infty}^{\infty}$ are WSS;
- (ii) $R_{X,Y}(t_1, t_2) \triangleq \mathbb{E}[X(t_1)Y(t_2)] = R_{X,Y}(t_1 - t_2) = R_{X,Y}(\tau), \text{ where } \tau = t_1 - t_2$

Power Spectrum Density

Power Spectrum Density:

- \square The **power spectrum density** $S_X(f)$ of a discrete-time WSS random process $\{X_n\}_{n=-\infty}^{\infty}$ is the discrete-time Fourier transform of the autocorrelation function $R_X(k)$:

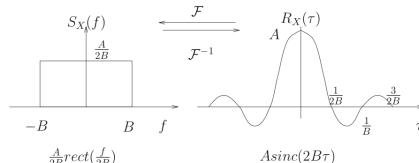
$$S_X(f) \triangleq \sum_{k=-\infty}^{\infty} R_X(k) e^{-j2\pi fk} \quad (-\frac{1}{2} < f \leq \frac{1}{2})$$

$$R_X(k) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f) e^{j2\pi fk} df$$

- \square The **power spectrum density** $S_X(f)$ of a continuous-time WSS random process $\{X(t)\}_{t=-\infty}^{\infty}$ is the Fourier transform of the autocorrelation function $R_X(\tau)$:

$$S_X(f) \triangleq \int_{-\infty}^{\infty} R_X(\tau) e^{-j2\pi f\tau} d\tau$$

$$R_X(\tau) = \int_{-\infty}^{\infty} S_X(f) e^{j2\pi f\tau} df$$



Power Spectrum Density

Properties of Power Spectrum Density:

For a real-valued random process $\{X(t)\}_{t=-\infty}^{\infty}$,

1. $S_X(f) \geq 0$ for all f ;
2. $S_X(f)$ is real and even;
3. Average power:

$$R_X(0) = \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f) df \quad (\text{discrete time})$$

$$R_X(0) = \int_{-\infty}^{\infty} S_X(f) df \quad (\text{continuous time})$$

Cross Power Spectrum Density:

For jointly WSS random processes $\{X(t)\}_{t=-\infty}^{\infty}$ and $\{Y(t)\}_{t=-\infty}^{\infty}$, the cross power spectrum density $S_{X,Y}(f)$ is the Fourier transform of the crosscorrelation function $R_{X,Y}(\tau)$.

Linear Filtering of Random Processes

Discrete Time:

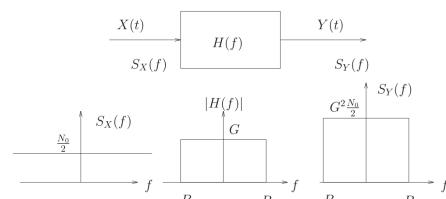
$$Y_n = \sum_{l=-\infty}^{\infty} h_l X_{n-l}$$

$$S_Y(f) = |H(f)|^2 S_X(f) \quad (-\frac{1}{2} < f \leq \frac{1}{2}) \quad (1)$$

Continuous Time:

$$Y(t) = \int_{-\infty}^{\infty} h(s) X(t-s) ds$$

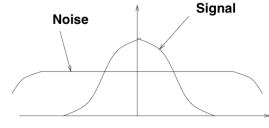
$$S_Y(f) = |H(f)|^2 S_X(f) \quad (2)$$



White Noise and Gaussian Processes

White Noise:

- Continuous-time white noise $\{X(t)\}$: a WSS process with zero mean and PSD $S_X(f) = N_0/2$ within the frequency range of interest $f \in [-W, W]$.
- Discrete-time white noise $\{X_n\}$: a sequence of zero-mean and uncorrelated random variables each with variance $\text{Var}(X_n) = \sigma^2$, i.e., $R_X(k) = \sigma^2 \delta_k$ and $S_X(f) = \sigma^2$ ($-\frac{1}{2} < f \leq \frac{1}{2}$).



Gaussian Processes

- Definition: A random process $X(t)$ is Gaussian if for all n and t_1, \dots, t_n , random variables $X(t_1), \dots, X(t_n)$ are jointly Gaussian.
- Properties:
 - The output of a linear filter driven by a Gaussian process is Gaussian.
 - Wide sense stationarity implies strict stationarity.