Probability Space

$(\Omega, \mathcal{F}, Pr)$ — Pr: probability measure on Ω; Ω: sample space; $\mathcal{F}$: collection of events I want to assign probabilities to; level of resolution we observe our random experiment at

$\mathcal{F}$ Axioms

(i) $\Omega \in \mathcal{F}$

(ii) If $A \in \mathcal{F}$, $A^c \in \mathcal{F}$ ← countable intersections also there! $A \cap B = (A^c \cup B^c)^c$

(iii) If $A_i \in \mathcal{F}$, $\bigcup_i A_i \in \mathcal{F}$

Pr $\mathcal{F} \to [0,1]$

(i) $0 \le Pr(A) \le 1$

(ii) $Pr(\Omega) = 1$ — If one event occurs the other events do not

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

Properties that follow

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

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

$Pr(\Omega) = 1 = Pr(A) + Pr(A^c)$

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

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

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

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

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

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

Union Bound

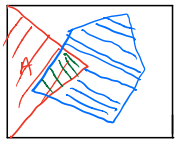
$Pr(\bigcup_{i=1}^n A_i) \le \sum_{i=1}^n Pr(A_i)$

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

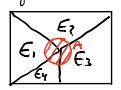
If $\{E_1, E_2, \dots, E_k\}$ partition Ω then

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

This makes more sense since you know some event which occurred

generative models

Example of Partition for $k=4$



No intersections $\bigcup_{i=1}^k E_i = \Omega$

Bayes Rule

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

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

prior knowledge · generative model · $Pr(A)$ 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

$Pr(A_1 \cap A_2 \cap \dots \cap A_k) = Pr(A_1) Pr(A_2) \cdots Pr(A_k)$ $\forall k \in [2,n]$, $k \in \mathbb{Z}$

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

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

Random Variables

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

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

M^{th} Central Moment: $\mathbb{E}[(X - \mathbb{E}[X])^n]$

Cumulative Distribution Function

The CDF is defined as

$$F_X(x) = Pr(X \le x) \quad \forall x \in \mathbb{R}$$
$$= Pr(\{\omega | X(\omega) \le x\})$$

Properties of CDF

① $\lim_{x \to \infty} F_X(x) = 0$, $\lim_{x \to -\infty} F_X(x) = 1$

② $\forall x < y$ $F_X(x) \le F_X(y)$

③ F is right continuous: $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$

④ $Pr[x \le X \le y] = F_X(y) - F_X(x)$

PMF for Discrete Random Variables

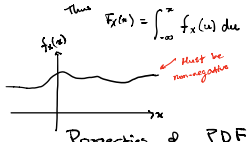
$P_X(x) = Pr[X = x]$

$F_X(x) = \sum_{u \le x} P_X(u)$

PDF for Continuous Random Variables

$f_X(x) = \frac{dF_X(x)}{dx}$ measures how fast we accumulate probability

Thus $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$



Properties of PDF

① $f_X(x) \ge 0$ $\forall x$

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

③ $Pr[x \le X \le y] = \int_x^y f_X(u)\, du$

Variance

How much a rv varies from its expectation

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

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

Correlation

Correlation between X and Y

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

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

Covariance

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

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

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

Properties

Independence of rv's X, Y: $f_{X,Y}(x,y) = f_X(x) f_Y(y) \Rightarrow$ uncorrelatedness

X, Y, Z r.v.'s

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

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

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 $Y = g(X)$ ($f_X(x)$, g)

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

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

Preservation of Order

If $Pr[X \ge Y] = 1$

then $\mathbb{E}[X] \ge \mathbb{E}[Y]$

Integration by Parts

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

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i\, Pr(X = i)$$
$$= \sum_{i=0}^{\infty} \underbrace{Pr(X > i)}_{1 - F_X(i)}$$ ← Tail Probability

To see this, observe

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

$\sum_{i=0}^{\infty} Pr(X > i)$ $i=0$ ✓ ✓; $i=1$ ✓; ⋮; 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

$$P_{X|Y}(x|y) \triangleq Pr[\overset{A}{X = x} \mid \overset{B}{Y = y}]$$
$$= \frac{Pr(\{X = x\} \cap \{Y = y\})}{Pr(\{Y = y\})} = \frac{P_{X,Y}(x,y)}{P_Y(y)}$$

Similarly for continuous r.v.'s

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

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

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

Cauchy-Schwartz Inequality

$$|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$$

Correlation Coefficient

(normalization comes from Cauchy-Schwarz inequality)

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \quad (|\rho_{X,Y}| \le 1)$$

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

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_m) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_m)$$

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

Linear Minimum Mean Squared Error

Now our estimator $g(Y)$ takes the form $g(Y) = aY + b$

The MSE becomes

$$\begin{aligned} MSE &= \mathbb{E}[(X - g(Y))^2] \\ &= \mathbb{E}[(X - aY - b)^2] \end{aligned}$$

Want

$$\{a^*, b^*\} = \underset{a,b}{\mathrm{argmin}}\ \mathbb{E}[(X - aY - b)^2]$$

Use orthogonality principle

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

Consider the following cases

① $a = 0, b = 1$ ($X, Y \Leftrightarrow \mathbb{E}[XY]$)

$$X - a^*Y - b^* \perp 1 \Rightarrow \mathbb{E}[X - a^*Y - b^*] = 0$$

($W \to \mathbb{E}[W] = 0$)

$$\mathbb{E}[X] - \mathbb{E}[a^*Y] - \mathbb{E}[b^*] = 0$$

(UNBIASED estimator)

$$\to b^* = \mathbb{E}[X] - a^*\mathbb{E}[Y]$$

② $a = 1, b = 0$

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

$$\begin{aligned} &\mathbb{E}[(X - a^*Y - b^*)Y] \\ &= \mathbb{E}[(X - a^*Y - \mathbb{E}[X] - a^*\mathbb{E}[Y])Y] \\ &= \mathbb{E}[(X - \mathbb{E}[X] - a^*(Y - \mathbb{E}[Y]))Y] \end{aligned}$$

(first ①)

$$\Rightarrow \mathrm{Cov}((X - \mathbb{E}[X] - a^*(Y - \mathbb{E}[Y])), Y) = 0$$

(Note: $\mathrm{Cov}(W,Y) = \mathbb{E}[WY]$ here)

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

$$a^* = \mathrm{Cov}(X,Y)/\mathrm{Var}(Y) \quad (\mathrm{Cov}(X,Y))$$

Aside

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

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

So

$$\begin{aligned} \hat{X}^*_{LMMSE} &= a^*Y + b^* \\ &= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}Y + \mathbb{E}[X] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\mathbb{E}[Y] \\ &= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y]) + \mathbb{E}[X] \end{aligned}$$

and the LMSE is

$$\begin{aligned} MSE &= \mathbb{E}[W^2] = \mathrm{Cov}(W,W) \\ &= \mathrm{Cov}((X - \mathbb{E}[X] - a^*(X - \mathbb{E}[Y]), (X - \mathbb{E}[X] - a^*(X - \mathbb{E}[Y])) \\ &= \mathrm{Cov}(X,X) + a^{*2}\mathrm{Cov}(Y,Y) - 2a^*\mathrm{Cov}(X,Y) \\ &= \mathrm{Var}(X) + \left(\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\right)^2\mathrm{Var}(Y) - 2\left(\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\right)\mathrm{Cov}(X,Y) \\ &= \mathrm{Var}(X) - \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(Y)} \end{aligned}$$

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

$$= X - \left(\mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y])\right)$$

$$= X - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}Y - \left(\mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\mathbb{E}[Y]\right)$$

Know W & Y are Jointly Gaussian

$$\mathbb{E}[W] = 0$$

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

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

Need W is error given by best LMMSE

$$\mathbb{E}[Wg(Y)] = 0 \quad \forall\, g(Y) \Rightarrow$$

Trivially true since W independent of Y.

$$\mathbb{E}[Wg(Y)] = \mathbb{E}[W]\mathbb{E}[g(Y)] = 0$$

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/ respect to the norm

$\|\hat{x}\| = \sqrt{\langle \hat{x}, \hat{x} \rangle}$ 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 (r.v.'s NOT vectors)

$$\langle X, Y \rangle \triangleq \mathbb{E}[XY]$$

Inner Product ⇒ norm/length ⇒ distance metric

length/norm of X = $\sqrt{\langle X, X \rangle} = \sqrt{\mathbb{E}[X^2]}$

Angle b/t X, Y; θ_{XY}: $\cos(\theta_{XY}) = \frac{\langle X,Y \rangle}{\|X\|\|Y\|} = \frac{\mathbb{E}[XY]}{\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}}$ (Cauchy-Schwarz ⇒ this ≤ 1)

Projection of X onto Y (a random variable)

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

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

$$\Pi_Y(X) = 0$$

Pythagorean (all X_i pairwise orthogonal)

$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}[X_i^2]$$

Jointly Gaussian Random Variables

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

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

Note: K is our covariance matrix.

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

$$= \mathbb{E}\left[\begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}\begin{pmatrix} X - \mu_X \\ Y - \mu_Y \end{pmatrix}^T\right]$$

$$= \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

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

Definition: Joint pdf of such a distribution is

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

Properties of Jointly Gaussian Random Variables

① J.G. ⇒ M.G. (⇍)

② Uncorrelated JG rv's are independent (special!)

independence ⇒ uncorrelated (⇏ 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

$$\hat{X}_{MMSE} = \mathbb{E}[X|Y] = \hat{X}_{LMMSE} = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y])$$

$$MMSE = \mathrm{Var}(X)(1 - \rho_{XY}^2) = \mathrm{Var}(X) - \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(Y)}$$

⑤ For J.G X, Y

$$\mathbb{E}[X|Y=y] = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,y)}{\mathrm{Var}(y)}(y - \mathbb{E}[y])$$

$$\mathbb{E}[X|Y] = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - \mathbb{E}[Y])$$

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

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

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

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi\mathrm{Var}(X|Y)}} e^{-(\cdots)} \to \mu: \mathbb{E}[X|Y], \quad \sigma^2: \mathrm{Var}(X|Y)$$

Proof! $X = \hat{X}_{MMSE} + W$

where W is independent of Y Given $Y = y$ (an observation),

- $\hat{X}_{MMSE}$ is a constant
- W maintains its distribution

Thus $X \sim N(\mu_W + \hat{X}_{MMSE}, \sigma_W^2)$

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

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

Minimum Mean Squared Error

$$\hat{X}_{MMSE} = \underset{\hat{X}}{\mathrm{argmin}}\ \mathbb{E}[(X - \hat{X})^2]$$

MMSE of X using a constant

$$\hat{X}_{MMSE} = a^*$$

where

$$a^* = \underset{a \text{ in } \mathbb{R}}{\mathrm{argmin}}\ \mathbb{E}[(X - a)^2]$$

} it's $\mathbb{E}[X]$!

$$\begin{aligned} MSE &= \mathbb{E}[(X - a)^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X] + \mathbb{E}[X] - a)^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(\mathbb{E}[X] - a)] + \mathbb{E}[(\mathbb{E}[X] - a)^2] \end{aligned}$$

($\to 0$)

Want to minimize

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

$$\to \mathrm{Var}(X) + (\mathbb{E}[X] - a)^2$$

Take $a = \mathbb{E}[X]$

MMSE of X given $Y = y$ (similar proof)

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

$$MSE(\hat{X}_{MMSE}) = \mathbb{E}[\mathrm{Var}(X|Y)]$$

The error

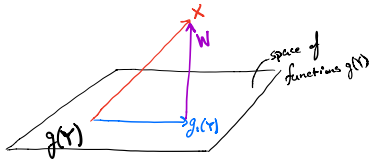
$$W = X - \mathbb{E}[X|Y]$$

has

$$\mathbb{E}[W] = \mathbb{E}[X - \mathbb{E}[X|Y]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$$

we call this type of estimator UNBIASED

VISUALIZATION



$X, Y, \hat{X}$

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

want $\min_{\hat{X}} \mathbb{E}[W^2]$

- Orthogonality principle:

$$W \perp g(Y) \quad \forall\, g$$

} necessary AND sufficient condition

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

claim $X - \mathbb{E}[X|Y] \perp g(Y)\ \forall\, g$ (inner product $\langle X - \mathbb{E}[X|Y], g(Y) \rangle$)

Proof $\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = 0$

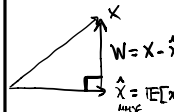
LHS

$$\mathbb{E}_{XY}[Xg(Y)] - \mathbb{E}_Y[(\mathbb{E}_{X|Y}[X|Y])g(Y)]$$

$$\mathbb{E}_Y[\mathbb{E}_{X|Y}[Xg(Y)|Y]] - \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]g(Y)]$$

$$\mathbb{E}_Y[g(Y)\mathbb{E}_{X|Y}[X|Y]] - \mathbb{E}_Y[g(Y)\mathbb{E}_{X|Y}[X|Y]] = 0 \checkmark$$

MSE Using Pythagorean Theory



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

$$MSE = \mathbb{E}[W^2] = \mathbb{E}[X^2] - \mathbb{E}[\hat{X}^2]$$

$$= \mathrm{Var}(X) + (\mathbb{E}[X])^2 - (\mathrm{Var}(\hat{X}_{MMSE}) + (\mathbb{E}[\hat{X}_{MMSE}])^2)$$

$$\mathbb{E}[X] = \mathbb{E}[\hat{X}] \quad = \mathrm{Var}(X) - \mathrm{Var}(\hat{X}_{MMSE})$$

Random Vectors

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}, \quad \mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_m] \end{bmatrix}$$

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

Notice Symmetry

$$\mathbb{E}[\underset{m\times 1}{X}\underset{1\times m}{X^T}] = \begin{bmatrix} \mathbb{E}[X_1^2] & \mathbb{E}[X_1X_2] & \cdots & \mathbb{E}[X_1X_m] \\ \mathbb{E}[X_2X_1] & \mathbb{E}[X_2^2] & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_mX_1] & \mathbb{E}[X_mX_2] & \cdots & \mathbb{E}[X_m^2] \end{bmatrix}_{m\times m}$$

Variance of $X_{m\times 1}$

$$\mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T\right] = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_m) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_m,X_1) & \mathrm{Cov}(X_m,X_2) & \cdots & \mathrm{Var}(X_m) \end{bmatrix}_{m\times m}$$

K → Covariance matrix

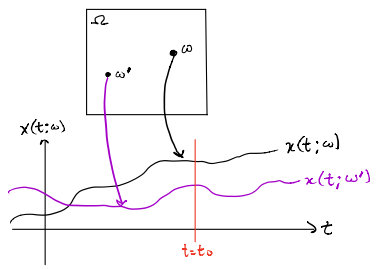
Cross-Correlation (b/t two random vectors)

$$\mathbb{E}[\underset{m\times 1}{X}\underset{1\times n}{Y^T}] = \{\mathbb{E}[X_iY_j]\}_{\substack{1\le i\le m \\ 1\le j\le n}} \quad \text{(}i\text{ rows, }j\text{ columns)}$$

$$\mathrm{Cov}(X,Y) = \{\mathrm{Cov}(X_i,Y_j)\}_{\substack{1\le i\le m \\ 1\le j\le n}}$$

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

$X(t_0,\omega')$, $X(t_0,\omega)$ random variables

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

$X(t)$ → r.p. $\quad x(t)$ → sample path

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

Specifically,

$$\forall \Delta \quad F_{X(t_1),\dots,X(t_n)}(x_1,\dots,x_n) = F_{X(t_1+\Delta),\dots,X(t_n+\Delta)}(x_1,\dots,x_n)$$
$$\forall n \quad \{t_i\}_{i=1}^n, \Delta.$$

WIDE Sense Stationary: A random process is WSS if

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

(ii) $R_X(t_1,t_2) \triangleq \mathbb{E}[X(t_1)X(t_2)]$
$= R(\tau)$

where $\tau = t_2 - t_1$

Independence Review

Independent Events: $A \perp\!\!\!\perp B$ if $\Pr[A\cap B] = \Pr[A]\Pr[B]$

or equivalently: $\Pr[A|B] = \Pr[A]$

Independent Random Variables:

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

or equivalently,

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

Conditional Independent Events

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

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

or equivalently,

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

Conditional Independent Random Variables:

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

or equivalently,

$$f_{X|Y,Z}(x|y,z) = f_{X|Z}(x|z)$$

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 $t_0<t_1<\cdots<t_n$ the increments $X(t_1)-X(t_0), \dots, X(t_n)-X(t_{n-1})$ 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

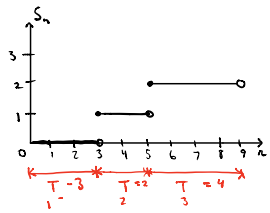
Counting Function

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

Binomial Counting Process

$$X_i \sim \text{Bernoulli}(p)$$
$$S_n = \sum_{i=1}^n X_i$$
$$S_n \sim \text{Binomial}(n,p)$$

$$\Pr[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$



T_1, T_2, T_3 are a sequence of inter-arrival times

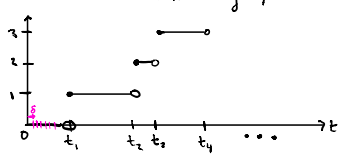
Interarrival Time $\{T_i\}_{i=1}^\infty$

In above example,

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

What about continuous counting processes?

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



Partition time into δ-length intervals.

Assumptions.

For $\delta\to 0$

① The probability of having more than one arrivals in δ-interval is negligible (i.e. $\to 0$)

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/ $p=\lambda\delta$

$$\delta\to 0$$
$$p = \lambda\delta \to 0$$
$$n = \frac{t}{\delta} \to \infty$$
$$S_n \sim \text{Binomial}\left(\tfrac{t}{\delta}, \lambda\delta\right) \longrightarrow \text{Poisson}(\lambda t)$$

Properties

① If JG, Uncorrelated ⇒ independence (NOT generally true)

$\mathrm{Cov}(X,Y)=0$ ↑ 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 X,Y jointly Gaussian then $X|Y$ is Gaussian USEFUL

with mean $\mathbb{E}[X|Y] = \hat{X}_{\text{Linear}} = \mathbb{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y-\mathbb{E}[Y])$

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

$= \mathrm{Var}(X)(1-\rho_{XY}^2)$ ← For every realization of Y, have the same result

in general, a rv w/ it variable being conditioned on

a CONSTANT

Gaussian Random Processes

Definition:

$X(t)$ is Gaussian if $\forall n$, and $\forall \{t_i\}_{i=1}^n$ r.v.s $X(t_1),\dots,X(t_n)$ are JOINTLY Gaussian

Properties "Completely specified by first two moments"

① completely specified by its mean function $\mu_X(t)$ and autocorrelation function $R_X(t,s)$ / autocovariance function $C_X(t,s)$.

Example "KNOW RIGHT VALUES"

$$f_{X(t_1),X(t_2)}(x_1,x_2) \sim \mathcal{N}\left(\begin{bmatrix}\mu_X(t_1)\\ \mu_X(t_2)\end{bmatrix}, K = \begin{bmatrix} C_X(t_1,t_1) & C_X(t_1,t_2) \\ C_X(t_2,t_1) & C_X(t_2,t_2)\end{bmatrix}\right)$$

② WSS ⇒ S

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

① $X(0)=0$

② $\{X(t)\}_{t\ge 0}$ has independent increments

③ $X(t_2)-X(t_1) \sim \mathcal{N}(0, \sigma^2(t_2-t_1))$

④ Every sample path is continuous

Properties of Brownian Motion

① $\mu_X(t) \triangleq \mathbb{E}[X(t)] = \mathbb{E}[X(t)-X(0)] = 0$

② $R_X(t_1,t_2) = C_X(t_1,t_2) \quad t_2\ge t_1$
$= \mathbb{E}[X(t_1)X(t_2)]$
$= \mathbb{E}[X(t_1)(X(t_1)+X(t_2)-X(t_1))]$
$= \mathbb{E}[X^2(t_1)] + \mathbb{E}[X(t_1)]\mathbb{E}[X(t_2)-X(t_1)]$
$= \mathbb{E}[(X(t_1)-X(0))^2] + \mathbb{E}[X(t_1)-X(0)]\mathbb{E}[X(t_2)-X(t_1)]$ ← just illustrates increments
$= \sigma^2 t_1$

In general,

$$C_X(t_1,t_2) = R_X(t_1,t_1) = \min(t_1,t_2)\sigma^2$$

③ $X(t)\sim\mathcal{N}(0,\sigma^2 t)$

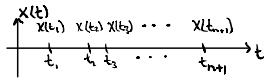
Also have $X(t_1), X(t_2), \dots, X(t_n)$ JG $\forall n$.

i.e.

$$\begin{bmatrix} X(t_1) \\ \vdots \\ X(t_n)\end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots \\ 1 & 1 & 1 & 0 \cdots \\ \vdots & & & \end{bmatrix} \begin{bmatrix} X(t_1)-X(0) \\ X(t_2)-X(t_1) \\ \vdots \\ X(t_n)-X(t_{n-1})\end{bmatrix}$$

Markov Process

A random process $\{X(t)\}_{t\ge 0}$ is a Markov process if $\forall n$, and $t_1<t_2<\cdots<t_{n+1}$



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

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 ⫫ 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

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

$\mathbb{E}[X]=\mu$

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

X and Y are jointly Gaussian if $Z = aX+bY$ is Gaussian $\forall a,b\neq 0\in\mathbb{R}$

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

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{|K|}} e^{-\frac{1}{2}\left(\begin{pmatrix}x-\mu_x\\ y-\mu_Y\end{pmatrix}^T K^{-1}\begin{pmatrix}x-\mu_x\\ y-\mu_Y\end{pmatrix}\right)}$$

where

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

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

Characterization of Discrete-Time Markov Chain:

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

- the PMF $\mathbf{p}(0)$ of its initial state X_0:

$$\mathbf{p}(0) \triangleq \left[\underbrace{\Pr[X_0 = i]}_{p_i(0)}\right]_{i\in\mathcal{X}}$$

- and one-step transition probabilities:

$$\left\{\Pr[X_{n+1} = j | X_n = i]\right\}_{n\geq 0, i,j\in\mathcal{X}}$$

Remark: the joint PMF is given by

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

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 $p_X(x_0)$ and $\{\Pr[X_{m+1} = j | X_m = i]\}_{i,j\in\mathcal{X}}$

Homogeneous Markov Chain ⇒ Time INVARIANT (↓ doesn't depend on $\forall m$ / n)

$$\Pr[X_{m+1} = j | X_m = i] \triangleq P_{i,j}$$

Transition Matrix:

$$\mathbb{P} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,M} \\ P_{2,1} & P_{2,2} & \cdots & P_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M,1} & P_{M,2} & \cdots & P_{M,M} \end{bmatrix}$$

$\mathcal{X} = \{1, 2, \ldots, M\}$

$0 \leq P_{i,j} \leq 1$

$\sum_j P_{i,j} = 1 \quad \forall i$

n-Step Transition Matrix

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

Chapman-Kolmogorov Equation

$$P_{i,j}^{(m+n)} = \sum_{k\in\mathcal{X}} \underbrace{P_{i,k}^{(m)}}_{i\times k} \underbrace{P_{k,j}^{(n)}}_{k\times j} \quad \forall i,j$$

$$\mathbb{P}^{(m+n)} = \mathbb{P}^{(m)}\mathbb{P}^{(n)}$$

$$\mathbb{P}^{(n)} = \mathbb{P}\mathbb{P}^{(n-1)} = \mathbb{P}\mathbb{P}\mathbb{P}^{(n-2)} = \vdots = \mathbb{P}^n$$

State Probability

This is the distribution at time n.

$\mathbf{p}(n) = [\Pr[X_n = 1] \cdots \Pr[X_n = m]]$ ← vector

$\mathbf{p}(0) = [\Pr[X_0 = 0] \cdots \Pr[X_0 = m]]$

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

$\vdots$

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

$\{X_n\}_{n\geq 0}$: $\vec{p}(0) = \begin{cases} \Pr[X_0 = 1, \ldots, X_0 = m] \\ \text{pmf of } X_0 \\ \mathbb{P} \end{cases}$

row vector by definition $\vec{p}(n) \triangleq \Pr[X_n = 1, \ldots, X_n = m]$

$$\Pr[X_n = j] = \sum_{i=1}^{m} \Pr[X_0 = i]\Pr[X_n = j | X_0 = i] = \vec{p}(0)\mathbb{P}^{(n)}$$

Gaussian Random Vectors

X is a Gaussian random vector if its coordinates are JOINTLY GAUSSIAN

$X \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix}$ a Gaussian random vector $\Leftrightarrow a_1X_1 + \cdots + a_mX_m$ is Gaussian $\forall a_i \in \mathbb{R}$, $1 \leq i \leq m$.

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

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

① Any subvector of X is a Gaussian random vector

- $X_i \sim \mathcal{N}(\mu_i, K_{i,i})$
- $\begin{bmatrix} X_i \\ X_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_i \\ \mu_j \end{bmatrix}, \begin{bmatrix} K_{i,i} & K_{i,j} \\ K_{j,i} & K_{j,j} \end{bmatrix}\right)$

② $\forall A, b$

$$\underset{m\times 1}{Y} = \underset{n\times m}{A}\underset{m\times 1}{X} + \underset{n\times 1}{b}$$

is a Gaussian random vector $\quad \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T]$

$$Y \sim \mathcal{N}(A\mu + b, AKA^T)$$

③ If K is diagonal, every rv in X is independent of every other rv in X

④ If $\underset{m\times 1}{X}$ and $\underset{n\times 1}{Y}$ are jointly Gaussian then they are independent iff $\mathrm{Cov}(X, Y) = 0$

MMSE Estimate of X using Y

MSE $\mathbb{E}[\|X - \hat{X}\|^2] = \mathbb{E}\left[\underset{1\times m}{(X - \hat{X})^T}\underset{m\times 1}{(X - \hat{X})}\right]$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i - \hat{X}_i]$$

Need to "design" $\hat{X} = g(Y) = \begin{bmatrix} g_1(Y) \\ g_2(Y) \\ \vdots \\ g_m(Y) \end{bmatrix}$

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

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

Our best MMSE estimator for X_i is

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

Need conditional distribution

$$f_{X_i|Y}(x|y) = \frac{f_{X_i,Y}(x_i, y_1, y_2, \ldots, y_m)}{f_Y(y_1, \ldots, y_m)}$$

LMMSE Estimator of X

Similar to rv case,

$$\underset{m\times 1}{\hat{X}} = \underset{m\times n}{A}\underset{n\times 1}{Y} + \underset{m\times 1}{b} = \mathbb{E}[X] + \underset{m\times n}{\mathrm{Cov}(X,Y)}\underset{n\times n}{\mathrm{Cov}^{-1}(Y)}\underset{n\times 1}{(Y - \mathbb{E}[Y])}$$

If X, Y are jointly Gaussian, (similar to rv case)

$$\mathbb{E}[X|Y] = \mathbb{E}[X] + \underset{m\times n}{\mathrm{Cov}(X,Y)}\underset{n\times n}{\mathrm{Cov}^{-1}(Y)}\underset{n\times 1}{(Y - \mathbb{E}[Y])}$$

Poisson Process:

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

Interarrival Time (note: T_i iid $\forall i$)

Look at T_1

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

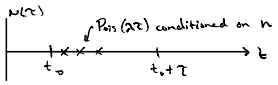
$$\lim_{\delta\to 0} (1 - \lambda\delta)^{t/\delta} = e^{-\lambda t}$$

i.e $T_i \sim \exp(\lambda)$ $\qquad \mathbb{E}[T_i] = \frac{1}{\lambda}$

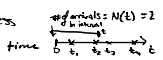
Leads to ANOTHER definition

Definition: $\{N(t)\}_{t\geq 0}$ is Poisson 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 $\forall$ any interval τ is a counting process such that $\forall \tau > 0$ $N(\tau) \sim \mathrm{Pois}(\lambda\tau)$, and given $N(\tau) = n$ (n arrival times in interval τ), are iid $\mathcal{U} \sim [0, \tau]$

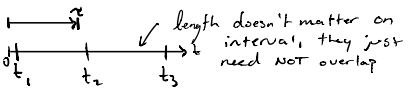


Facts about Poisson Process $\{N(t)\}_{t\geq 0}$ w/ rate λ. $\lambda \to$ arrival rate/unit time



① Process has independent increments

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



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

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

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 $\mathcal{U}[0, \tau]$ i.e $t_i \sim \mathcal{U}[0, k]$

⑤ $\mu_N(t) = \mathbb{E}[N(t)] = \lambda t$ i.e arrival rate* time

$\mathrm{Var}(N(t)) = \lambda t$ (Poisson property)

$R_N(s,t) \triangleq \mathbb{E}[N(s)N(t)]$ assume $s \leq t$ (if $s > t$ have var.)

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

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

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

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

$C_N(s,t) = \lambda \min(s,t)$ ← applies $\forall s,t$ (note above we assumed $s \leq t$)

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 $\mathcal{X} = \mathbb{Z}^+$.

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 $\mathbf{P}$ if

$$\pi = \pi\mathbf{P}$$

Computing Stationary Distribution:

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

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.

Intepretations of Stationary Distribution

Long-term occupancy rate:

For a Markov chain with $\mathbf{p}(0)$ 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 the example of Coin A and Coin B, the stationary distribution is $\pi = [\frac{2}{3}\ \frac{1}{3}]$, which can be interpreted as seeing a "head" 40% of the time and a "tail" 60% of the time.

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 $i \to j$.
- If $i \to j$ and $j \to i$, 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 $\mathcal{X} \backslash C_1$. 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

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

Transience and Recurrence

Probability of Return:

- ☐ $f_i^{(n)} \triangleq \Pr\{X_1 \neq i, \cdots, X_{n-1} \neq i, X_n = i \mid X_0 = i\}$ 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:

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

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 $\sum_{n=1}^{\infty} p_{i,i}^{(n)} = \infty$.*
- ☐ *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, returns to i for the first time. It is a random variable with PMF given by $[f_i^{(1)}, f_i^{(2)}, \cdots]$.

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

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

- ☐ 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 transiting 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 $j \to i$. Together with $p_{i,j} > 0$, 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

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

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 kth positive recurrent class. Then for any $0 \leq \alpha_1, \ldots, \alpha_K \leq 1$ with $\sum_{k=1}^{K} \alpha_k = 1$,

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

is a stationary distribution.

Absorption Probability

- Modify original Markov chain by adding absorbing states 0, X+1 corresponding to winning "game"
- Define the conditional absorption probability

$$g_i \triangleq \Pr\left[\text{Absorbed at state } 0 \mid \text{current state } i\right]$$

- Use Markovian property to get system of equations.
- $\sum \pi_i g_i \to$ "A wins" $\quad 1 - \sum \pi_i g_i \to$ "B wins"

Absorption Time

Define the conditional expected absorption time

$$e_i \triangleq \mathbb{E}\left[\text{Remaining time until Absorbed} \mid \text{Currently at state } i\right]$$

Based on Markovian property, get system of equations

- Expected absorption time is $\sum \pi_i e_i$ for non-adsorbed i

Continuous-Time Markov Chain

Continuous-Time Markov Chain:

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

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

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

- ☐ 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:

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

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 $\{X(t)\}_{t \geq 0}$, 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

- $q_{i,j} = \lambda_i p_{i,j}$ for $i \neq j$ is the rate of transiting from state i to state j.
- $q_{i,i} = -\sum_{j\neq i} q_{i,j}$ 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.
- 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

$$\begin{aligned} &(i) \quad \mathbb{E}[X(t_i)] = m \\ &(ii) \quad R_X(t_1,t_2) \triangleq \mathbb{E}[X(t_1)X(t_2)] \\ &\qquad = R_X(t_1 - t_2) \\ &\qquad = R_X(\tau), \quad \text{where } \tau = t_1 - t_2 \end{aligned}$$

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

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

Power Spectrum Density

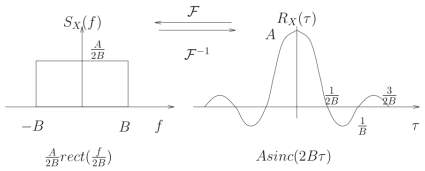
Power Spectrum Density:

- 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 $R_X(k)$:

$$\begin{aligned} S_X(f) &\triangleq \sum_{k=-\infty}^{\infty} R_X(k)e^{-j2\pi fk} \qquad (-\frac{1}{2} < f \leq \frac{1}{2}) \\ R_X(k) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f)e^{j2\pi fk}df \end{aligned}$$

- 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 $R_X(\tau)$:

$$\begin{aligned} S_X(f) &\triangleq \int_{-\infty}^{\infty} R_X(\tau)e^{-j2\pi f\tau}d\tau \\ R_X(\tau) &= \int_{-\infty}^{\infty} S_X(f)e^{j2\pi f\tau}df \end{aligned}$$



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:

$$\begin{aligned} R_X(0) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} S_X(f)df \quad \text{(discrete time)} \\ R_X(0) &= \int_{-\infty}^{\infty} S_X(f)df \quad \text{(continuous time)} \end{aligned}$$

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 $R_{X,Y}(\tau)$.

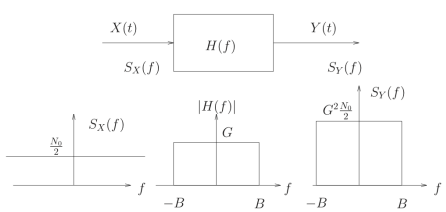
Linear Filtering of Random Processes

Discrete Time:

$$\begin{aligned} Y_n &= \sum_{l=-\infty}^{\infty} h_l X_{n-l} \\ S_Y(f) &= |H(f)|^2 S_X(f) \qquad (-\frac{1}{2} < f \leq \frac{1}{2}) \end{aligned} \tag{1}$$

Continuous Time:

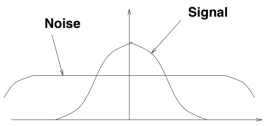
$$\begin{aligned} Y(t) &= \int_{-\infty}^{\infty} h(s)X(t-s)ds \\ S_Y(f) &= |H(f)|^2 S_X(f) \end{aligned} \tag{2}$$



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 $t_1, \cdots, t_n$, random variables $X(t_1), \cdots, X(t_n)$ are jointly Gaussian.
- Properties:
  - The output of a linear filter driven by a Gaussian process is Gaussian.
  - Wide sense stationarity implies strict stationarity.