Limiting probabilities

- $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.6 & 0.4 \\ 0.6 & 0.4 & 0 \end{bmatrix}$ $P^5 = \begin{bmatrix} .06 & .3 & .64 \\ .18 & .38 & .44 \\ .38 & .44 & .18 \end{bmatrix}$ $P^{30} = \begin{bmatrix} .23 & .385 & .385 \\ .23 & .385 & .385 \\ .23 & .385 & .385 \end{bmatrix}$

- $P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$ $\lim_{n\to\infty} P^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}$

- What is the interpretation of $\lim_{n\to\infty} p_{ij}^{(n)} = [\lim_{n\to\infty} P^n]_{ij}$?

- $\pi_j = \lim_{n\to\infty} p_{ij}^{(n)}$ denotes the probability of being in state j at time n when starting in state i.

- For an M state DTMC, $\pi = (\pi_1, \ldots, \pi_M)$ denotes the limiting distribution.

- How do we obtain the limiting distribution π? Does it always exist?

Stationary distribution

The **stationary distribution** of a Markov chain is defined as a solution to the equation $\pi = \pi P$.

- πP is essentially the p.m.f of X_1 having picked X_0 according to π.
- $\pi = \pi P$ says that, if the initial distribution is π, then the distribution of X_1 is also π.
- Continuing this argument, the p.m.f of X_n for any n is π and there is no dependence on the starting state.
- MCMC algorithms use this idea (at stationarity successive states of the Markov chain have p.m.f π) to sample from target distribution π.

Limiting vs Stationary distribution

- Obtain stationary distribution for the Markov Chain with transition probability $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.6 & 0.4 \\ 0.6 & 0.4 & 0 \end{bmatrix}$
- The limiting distribution need not exist for some Markov chains, but the stationary distribution exists. For example for $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.
- The limiting distribution if it exists, is same as the stationary distribution.

Statistical Inference

- Statistical Inference methods deal with drawing inference about an unknown model/random variable/random process from observations/data.
- There is an unknown quantity θ^* that we would like to estimate using data $\mathcal{D}$. eg: ML, signal/communication systems.
- Broadly, you can give 3 types of estimates for θ^*.
  1. Point Estimation: Here you want to give a point estimate which is a single numerical value that is your best guess for θ^*.
  2. Interval Estimation: here you give an interval on say $\mathbb{R}$ where θ^* is bound to lie with some certainty.
  3. Hypothesis testing: In binary hypothesis testing, you have two hypothesis ($H_0 : \theta = \alpha_1$ and $H_1 : \theta = \alpha_2$) and you use data $\mathcal{D}$ to decide which is true.

Statistical Inference

- There are two approaches to Statistical Inference:
  1) Bayesian 2) Frequentist (or classical)

- In Bayesian Inference, the unknown quantity is modelled as a random variable with a distribution that keeps changing as more and more data becomes available.

- Bayesian inference assumes a prior distribution $p_\Theta(\theta)$ on the unknown parameter θ^* and uses the likelihood $p_{X|\Theta}(x|\theta)$ for observing data x to obtain the posterior $p_{\Theta|X}(\theta|x)$

- In Bayesian inference, prior and posterior distribution reflect our state of knowledge.

Frequentist or Classical Inference

- Classical Inference models the unknown quantity as a constant and come up with estimators that are deterministic functions of the observed data.
- Given data, these estimators are deterministic functions of the data, but in reality are also random variables.
- For example sample mean as an estimator for the mean.

Classical Inference: Point Estimation

- Let θ^* denote the unknown parameter of a random variable X (typically mean, variance, scale, shape etc) and suppose we observe i.i.d samples of X which are recorded in the dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$.
- In frequentist approach, we estimate θ^*, by defining a point estimator $\hat{\Theta}$ as a function of the random samples $X_1, \ldots X_n$ as $\hat{\Theta} = h(X_1, \ldots X_n)$.
- While $\hat{\Theta}$ is a random variable, given $\mathcal{D}$ the estimator takes the value $\hat{\Theta} = h(x_1, \ldots x_n)$.
- Example : Sample mean $\hat{\mu}_n = \frac{\sum_{i=1}^{n} X_i}{n}$.