Applied Probability

Manish Agarwal

Oct 2022

This is based on the course MITx6.431x by Prof. Tikilis.

Contents

1 Probability rules

If Ω is the set of all possible outcomes, we define event A a subset of the sample space. We then assign probability to the events. The Axioms of probability include

- Non-negativity: $\boldsymbol{P}[A] \geq 0$.
- Normalization: $\boldsymbol{P}[\Omega] = 1$.
- Countable Additivity: if $A \cap B = \emptyset$, then $\boldsymbol{P}[A \cup B] = \boldsymbol{P}[A] + \boldsymbol{P}[B]$.

There are three important conditioning rules

- Multiplication rule: $\boldsymbol{P}[A \cap B] = \boldsymbol{P}[B|A]\boldsymbol{P}[A] = \boldsymbol{P}[B] = \boldsymbol{P}[A|B]\boldsymbol{P}[B]$.
- total probability rule: $\boldsymbol{P}[A] = \sum_i [A|B_i]\boldsymbol{P}[B_i]$.
- Bayes' rule: Definition of conditional probability given by $\boldsymbol{P}[A|B] = \frac{\boldsymbol{P}[A \cap B]}{\boldsymbol{P}[B]}$.

Two event A and B are independent if $\boldsymbol{P}[A \cap B] = \boldsymbol{P}[A]\boldsymbol{P}[B]$. For more than two events to be independent each combination has to be independent.

The expectation of a random variance X is $\boldsymbol{E}[X] = \sum\limits_x x p_X(x) = \int_x x f_X(x) dx$. An indicator of an event is 1_A and $\boldsymbol{E}[1_A] = \boldsymbol{P}[A]$. Also, $\boldsymbol{E}[aX+b] = a\boldsymbol{E}[X]+b$. Variance is defined as $Var[X] = \boldsymbol{E}[(X-\boldsymbol{E}[X])^2] = \boldsymbol{E}[X^2]-\boldsymbol{E}[X]^2$. Also, $Var[aX + b] = a^2 Var[X]$. Law of total expectations states

$$\boldsymbol{E}[A] = \boldsymbol{E}\left[\boldsymbol{E}[A|B]\right] = \sum_{i=1}^{n} \boldsymbol{E}[X|A_i]\boldsymbol{P}[A_i].$$

Conditional PMFs are similarly defined, $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{P_Y(y)}$ and the conditional expectation written as $\boldsymbol{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y)$. X, Y are independent if $\boldsymbol{E}[XY] = \boldsymbol{E}[X]\boldsymbol{E}[Y]$. A CDF is defined as $F_X(x) = \boldsymbol{P}[X \leq x]$. and the PDF is given by $f_X(x) = \frac{dF_X}{dx}(x)$. For a general function $Y = g(X)$ we have $f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}}{dy}(y)|$. we have the following defintion of covaraince

$$Cov[X, Y] = \boldsymbol{E}[XY] - \boldsymbol{E}[X]\boldsymbol{E}[Y].$$

The law of total variance is given by

$$Var[X] = \boldsymbol{E}[Var[X|Y]] + Var[\boldsymbol{E}[X|Y]].$$

2 Bayesian Inference and MAP

Maximum a posteriori probability (MAP) estimate is defined as $\max_\theta \boldsymbol{f}_{\Theta|X}(\theta|x)$. The Least mean sqaures (LMS) estimate is defined as $\min_\theta \boldsymbol{E}[(\Theta-\hat{\theta})^2|X=x] = \boldsymbol{E}[\Theta|X=x]$. We have a biased coin with bias Θ with prior $f_\Theta(.)$ which is uniform in $[0,1]$. For n trials we get K heads. Using Bayes' rule we can write $f_{\Theta|K}(\theta|k) \propto \theta^k(1-\theta)^{n-k}$, using Binomial distribution. This is a $Beta(k+1, n-k+1)$ distribution.

However, if our prior itself was a Beta distribution $f_\Theta(\theta) = \frac{1}{c}\theta^\alpha(1-\theta)^\beta$ with $\alpha, \beta \geq 0$. Thus we have the posterior $f_{\Theta|K}(\theta|k) \propto \theta^{\alpha+k}(1-\theta)^{\beta+n-k}$. Thus, the posterior remains a beta distribution. The MAP estimate is $\hat{\theta}_{MAP} = \max_\theta \log f_{\Theta|K}(\theta|k) = \frac{\alpha+k}{\alpha+\beta+n}$. The LMS estimate is $\hat{\theta}_{LMS} = \boldsymbol{E}[\Theta|K=k] = \int_0^1 f_{\Theta|K}(\theta|k)d\theta = \frac{\alpha+k+1}{\alpha+\beta+n+2}$.

To measure the performance we can use probability of error $\boldsymbol{P}[\hat{\Theta} \neq \Theta|X=x]$ or $\boldsymbol{P}[\hat{\Theta} \neq \Theta]$ or the MSE $\boldsymbol{E}[(\hat{\Theta}-\Theta)^2|X=x]$ or $\boldsymbol{E}[(\hat{\Theta}-\Theta)^2]$.

2.1 Linear models with normal noise

We want to ultimately consider the model $X_i = \sum_{j=1}^m a_{ij}\Theta_j + W_j$, but we will build up to it slowly.

- For a normal variable $X \sim \mathcal{N}(\mu, \sigma^2)$ we have the pdf $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Thus, for a negative quadratic exponential density $ce^{-\alpha x^2-\beta x-\gamma}$ to be identified as a normal pdf we need $\alpha > 0$ and has a mean $-\frac{\beta}{2\alpha}$ and $\sigma^2 = \frac{1}{2\alpha}$.
- We start with the simplest model $X = \Theta + W$, where $\Theta, W \sim \mathcal{N}(0,1)$ and $\Theta \perp W$. Here Θ is the parameter we wish to estimate, and W is some noise in the system. X is the observations in our model. Under the Bayesian program inference about Θ is the calculation of the posterior $f_{\Theta|X}(\theta|x)$. Now, this posterior can be written as

  $$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}, \; f_X(x) = \int f_\Theta(\theta)f_{X|\Theta}(x|\theta)d\theta.$$

  If we know the realized value of $\Theta = \theta$, we simply have $X = \theta + W$ and hence, $f_{X|\Theta}(x|\theta) \sim \mathcal{N}(\theta, 1)$, because Θ is independent of W. We can now calculate the posterior

  $$f_{\Theta|X}(\theta|x) \propto e^{-\frac{\theta^2}{2}}e^{-\frac{(x-\theta)^2}{2}}$$

  which we recognize as a Normal pdf as well. Thus, to estimate Θ we simply use $\hat{\theta}_{MAP} = \hat{\theta}_{LMS} = \boldsymbol{E}[\Theta|X = x] = \min_\theta \left(\frac{1}{2}\theta^2 + \frac{1}{2}(x-\theta)^2\right) = \frac{x}{2}$. This gives us $\hat{\Theta}_{MAP} = \boldsymbol{E}[\Theta|X] = \frac{X}{2}$. This generalizes for any Normal distribution.
- We now consider the case of multiple observations, i.e. we have $X_i = \Theta + W_i, \; i = 1, \ldots, n$, and we want to estimate Θ such that $\Theta \sim \mathcal{N}(x_0, \sigma_0^2)$ and $W_i \sim \mathcal{N}(0, \sigma_i^2)$ and Θ, W_i are independent. Here, we first notice that $f_{X_i|\Theta}(x_i|\theta) \propto e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}$ and $(X_i|\Theta=\theta) \sim \mathcal{N}(\theta, \sigma_i^2)$. We now notice that given $\Theta = \theta$, W_i are independent implies X_i are independent. Thus, we can write the joint distribution as

  $$f_{X|\Theta}(x|\theta) = f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f_{x_i|\Theta}(x_i|\theta)$$

  And hence, the posterior is

  $$f_{\Theta|X}(\theta|x) \propto e^{-\frac{(\theta-x_0)^2}{2\sigma_0^2}} \prod_{i=1}^n e^{-\frac{(x_i-\theta)^2}{2\sigma_i^2}}$$

  We can take the derivative of the quadratic term in the exponential to get the MAP estimate. Giving $\hat{\Theta}_{MAP} = \frac{\sum_{i=0}^n x_i/\sigma_i^2}{\sum_{i=0}^n 1/\sigma_i^2}$. Thus we recognize that the posterior is normal and LMS and MAP estimate coincide. This estimate are linear, i.e. of the form $\sum a_i x_i$ kind.

- To estimate the performance we use the the mean squared error $\boldsymbol{E}[(\Theta - \hat{\Theta})^2|X = x] = E[(\Theta - \hat{\theta})^2|X = x] = Var[\Theta|X = x]$. From the quadratic expression we have $\alpha = \sum_{i=0}^{n} \frac{1}{2\sigma_i^2}$. Thus mean squared error is $1/2\alpha = 1/\sum_{i=0}^{n} \frac{1}{\sigma_i^2}$. To estimate the unconditional variance $\boldsymbol{E}[(\Theta - \hat{\Theta})^2] = \int \boldsymbol{E}[(\Theta - \hat{\Theta})^2|X = x] f_X(x)dx = 1/\sum_{i=0}^{n} \frac{1}{\sigma_i^2}$. In the special case of all σs being equal we have the mean squared error is $\frac{\sigma^2}{n+1}$.

2.2 Least mean squares (LMS) estimation

Next, we use **Least mean squares (LMS) estimation**, $\hat{\theta} = \min_{\theta} \boldsymbol{E}[(\Theta - \hat{\theta})^2|X = x] = \boldsymbol{E}[\Theta|X = x]$. Here we try to minimize mean squared error (MSE) $\boldsymbol{E}[(\Theta - \hat{\theta})^2]$, conditional or unconditional. $\hat{\Theta}_{LMS} = \boldsymbol{E}[\Theta|X]$ minimizes $\boldsymbol{E}[(\Theta - g(X))^2]$ over all estimators $\hat{\Theta} = g(X)$. This is same as MAP if the posterior is unimodal and symmetric around the mean, e.g. when the posterior is normal, the case in linear-normal models. If we define the error $\tilde{\Theta} = \hat{\Theta} - \Theta$, then we see the it is unbiased, i.e. $\boldsymbol{E}[\tilde{\Theta}] = 0$, as well as $\boldsymbol{E}[\tilde{\Theta}|X = x] = 0$. We can also calculate $Cov[\tilde{\Theta}, \hat{\Theta}] = 0$. Finally, $Var[\Theta] = Var[\hat{\Theta}] + Var[\tilde{\Theta}]$.

We sometimes simplify the problem of finding the functional estimator $\hat{\Theta} = g(X)$ by considering **linear least mean square estimators** of the form $\hat{\Theta}_{LLMS} = aX + b$, where we minimize $\boldsymbol{E}[(\Theta - \hat{\Theta}_{LLMS})^2]$ with respect to a, b. In the special case, when $\boldsymbol{E}[\Theta|X]$ is linear in X, then $\hat{\Theta}_{LMS} = \hat{\Theta}_{LLMS}$. If a is known we know that the optimal $b = \boldsymbol{E}[\Theta] - a\boldsymbol{E}[X]$, using the LMS estimation result. Thus, we need to minimize $\boldsymbol{E}[(\Theta - aX - \boldsymbol{E}[\Theta - aX])^2] = Var[\Theta - aX] = Var[\Theta] + a^2 Var[X] - 2xCov[\Theta, X]$. This gives $a = \frac{Cov[\Theta,X]}{Var[X]}$. Thus,

$$\hat{\Theta}_{LLMS} = \boldsymbol{E}[\Theta] + \frac{Cov[\Theta, X]}{Var[X]}(X - \boldsymbol{E}[X]) = \boldsymbol{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - \boldsymbol{E}[X])$$

Thus, for LLMS estimator only means, variances and covariances matter. We note that correlation plays a critical role in linear estimators in reducing error.

$$\boldsymbol{E}[(\hat{\Theta}_{LLMS} - \Theta)^2] = (1 - \rho^2)Var[\Theta].$$

This can be extended to multiple observations where we solve a set of linear equations. The LMS estimate $\boldsymbol{E}[\Theta|X]$ is same as $\boldsymbol{E}[\Theta|X^3]$, but they are different for LLMS: $\hat{\Theta} = aX + b$ versus $\hat{\Theta} = aX^3 + b$. We can also consider polynomial of X or any function of X and the model still remains linear.

Example 2.1. *Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ, uniformly distributed between zero and one hour.*

*Assume that Juliet was late by an amount x on their first date, how should Romeo use this information to update the distribution of Θ? **Solution**: We note that $f_\Theta(\theta) = \mathbf{1}_{0\le\theta\le1}$ and $f_{X|\Theta}(x|\theta) = \frac{1}{\theta}\mathbf{1}_{0\le x\le\theta}$. We can, thus, find the posterior $f_{\Theta|X}(\theta|x) = \frac{f_\Theta \theta f_{X|\Theta}(x|\theta)}{f_X(x)} = \frac{1}{\theta|\log x|}\mathbf{1}_{x\le\theta\le1}$. We see the prior and posterior here*

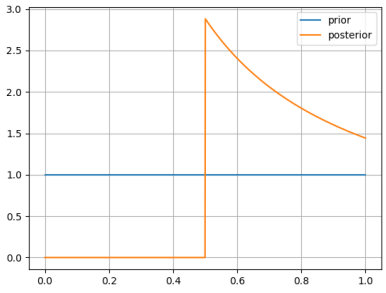


Figure 1: Prior and posterior for $X = 0.5$

How should Romeo update the distribution of Θ if he observes that Juliet is late by $x_1, \ldots, x_n$ on the first

n dates? Assume that Juliet is late by a random amount $X_1, \ldots, X_n$ on the first n dates where, given θ, $X_1, \ldots, X_n$ are uniformly distributed between zero and θ and are conditionally independent. ***Solution****: We find the joint likelihood as* $f_{X_1,\ldots,X_n|\Theta}(x_1, \ldots, x_n|\Theta) = \frac{1}{\theta^n}\mathbf{1}_{x_1\leq\theta}1_{x_n\leq\theta} = \frac{1}{\theta^n}\mathbf{1}_{\max(X_n)\leq\theta}$. *Thus, the posterior is* $f_{\Theta|X_1,\ldots,X_n}(\theta|x_1, \ldots, x_n) = \frac{f_\Theta(\theta)f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n|\theta)}{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)} \propto \frac{1}{\theta^n}\mathbf{1}_{\max(X_n)\leq\theta}$.

Find the MAP estimate of Θ based on the observation $X = x$. ***Solution****: Looking at the distribution we note that the posterior takes the highest value at $\hat{\theta}_{MAP} = x$.*

Find the LMS estimate of Θ base don the observation $X = x$. ***Solution****: This is defined as* $\hat{\theta}_{LMS} = \boldsymbol{E}[\Theta|X = x] = \int_x^1 \theta\frac{1}{\theta|\log x|}d\theta = \frac{1-x}{|\log x|}$.

Calculate the conditional mean squared error for the MAP and LMS estimates and compare the results. ***Solution****: For any estimator $\hat{\theta}$, conditional MSE is $\boldsymbol{E}[(\hat{\Theta} - \Theta)^2|X = x]$ which is equal to $\int_x^1(\hat{\theta} - \theta)^2\frac{1}{\theta|\log x|}d\theta = \hat{\theta}^2 - 2\hat{\theta}\frac{1-x}{|\log x|} + \frac{1-x^2}{2|\log x|}$. Thus, the MSE for MAP estimate is $x^2 - 2x\frac{1-x}{|\log x|} + \frac{1-x^2}{2|\log x|}$ and the MSE for LMS estimate is $\frac{1-x^2}{2|\log x|} - \left(\frac{1-x}{\log x}\right)^2$. The MSE of LMS is the minimum of all as it is designed for it.*

Derive the LLMS estimator of Θ based on X. ***Solution****: The LLMS estimate is given by* $\hat{\theta}_{LLMS} = \boldsymbol{E}[\Theta] + \frac{Cov[\Theta,X]}{Var[X]}(X - \boldsymbol{E}[X])$. *We can calculate,* $\boldsymbol{E}[\Theta] = \frac{1}{2}, \boldsymbol{E}[X] = \boldsymbol{E}[\boldsymbol{E}[X|\Theta]] = \boldsymbol{E}[\frac{\Theta}{2}] = \frac{1}{4}, Var[X] = \boldsymbol{E}[Var[X|\Theta]] + Var[\boldsymbol{E}[X|\Theta]] = \boldsymbol{E}[\frac{\Theta^2}{12}] + Var[\frac{\Theta}{2}] = \frac{1}{12}(Var[\Theta] + (\boldsymbol{E}[\Theta])^2) + \frac{1}{2}Var[\Theta] = \frac{7}{144}, and Cov[\Theta, X] = \boldsymbol{E}[\Theta X] - \boldsymbol{E}[\Theta]\boldsymbol{E}[X] = \boldsymbol{E}[\boldsymbol{E}[\Theta X|\Theta]] - \frac{1}{8} = \boldsymbol{E}[\frac{\Theta^2}{2}] - \frac{1}{8} = \frac{1}{24}$. *This gives* $\hat{\theta}_{LLMS} = \frac{6}{7}x + \frac{2}{7}$.

Calculate the conditional mean squared error of the LLMS estimate and compare it to other estimators. ***Solution****: The MLE for this is obtained by plugging $\hat{\theta}_{LLMS}$ into the MSE formula above. We see the comparision of the three MSE here* □

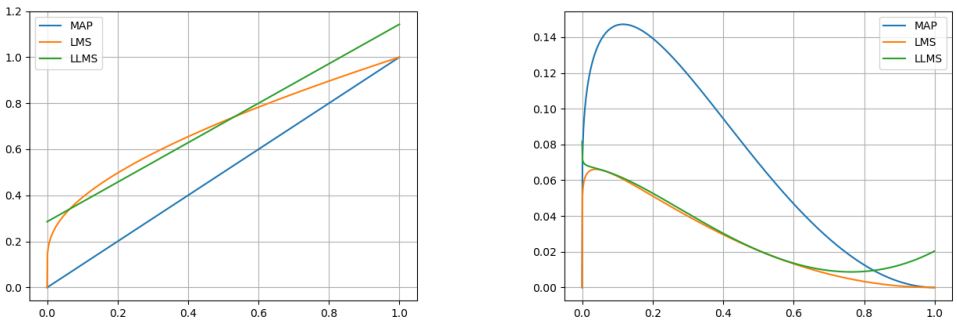


Figure 2: Estimators and MSE for the three estimators for $X = x$

3 Statistics

For a non-negative random variable X with known mean we have the **Markov inequality**

$$\boldsymbol{P}[X \geq c] \leq \frac{\mu}{c}.$$

Chebysheb inequality is valid for a random variable with known mean and variance

$$\boldsymbol{P}[|X - \mu| \geq c] \leq \frac{\sigma^2}{c^2}.$$

Hoeffding's inequality further strengthens these bounds for bounded random variables $a \leq X_i \leq b$. For $S_n = \sum X_i$ we have

$$\boldsymbol{P}[|S_n - \boldsymbol{E}[S_n]| \geq t] \leq 2e^{-\frac{2t^2}{n(b-a)^2}}$$

Central limit theorem gives a much tighter bound for $n \to \infty$.

Let $X_1, X_2, \ldots$ are iid observations with finite mean μ and variance σ^2. The sample mean is a random variable defined as $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Notice that $\boldsymbol{E}[X_i] = \mu$. We further note that $\boldsymbol{E}[M_n] = \boldsymbol{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}[X_i] = \mu$. Similarly, $\boldsymbol{V}[Mn] = \frac{1}{n^2}\sum_{i=1}^{n}\boldsymbol{V}[X_i] = \frac{\sigma^2}{n}$. Thus, using Chebysheb inequality we have $\boldsymbol{P}[|M_b - \mu| \geq \epsilon] \leq \frac{\boldsymbol{V}[M_n]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0$, for a fixed $\epsilon > 0$. This is called the **weak law of large numbers**

$$\text{for } \epsilon > 0, \boldsymbol{P}[|M_n - \mu| \geq \epsilon] \xrightarrow[n\to\infty]{} 0.$$

We can look at the sample sum $S_n = \sum_{i=1}^{n} X_i$ whose variance is $n\sigma^2$ or the mean $M_n = \frac{1}{n}\sum_{i=1}^{n}$ whose variance is $\frac{\sigma^2}{n}$, both these distributions are uninteresting in the limit $n \to \infty$. However if we scale the variable as $\frac{S_n}{\sqrt{n}}$ we see the variance is always σ^2. What distribution does this approach as $n \to \infty$ is answered by central limit theorem. Let $X_1, \ldots, X_n$ be iid with finite mean μ and variance σ^2. We introduce the standardized variable $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$, which has $\boldsymbol{E}[Z_n] = 0$ and $\boldsymbol{V}[Z_n] = 1$. Then the **central limit theorem** state that (where Z is standard Normal random variable)

$$\lim_{n\to\infty} \boldsymbol{P}[Z_n < z] = \boldsymbol{P}[Z \leq z]$$

This is a very deep result, as the converging distribution is independent of the distribution of X_i, depending only on the mean and variance of X and X being iid. Note that the converges of CDF implies converges of PDF. In fact if X are not identical distributed but with same mean and variance, the result still holds. Weak dependence (e.g. local dependence) there are versions of CLT which can be applied. For example, if for some $m \geq 0$, $X_1, X_2, \ldots$ is a stationary m-dependent sequence with $\boldsymbol{E}[X_i] = \mu$ and $\boldsymbol{V}[X_i] = \sigma^2$ then $\sqrt{n}\left(\bar{X}_n - \mu\right) \to \mathcal{N}\left(0, \sigma^2 + 2\sum_{k=1}^{m} Cov[X_i, X_{i+k}]\right)$.

In practice we can apply CLT for moderate $n \approx 30$ as well, and symmetry and uni-modality help as well. De Moivre-Laplace CLT to the binomial is a special case where we use Normal distribution to approximate Binomial variable (sum of Bernoulli random variables) using CLT. It uses the 1/2 approximation.

In classical statistics we consider θ to be constant, and the observations come from a distribution parametrized by θ and our job then is to come up with an estimator $\hat{\Theta}$ (which is a random variable) and infer the value $\hat{\theta}$ and give some confidence interval around it. The kinds of problem we encounter are: **Hypothesis testing** where we choose from two possible scenarios, **Estimation problem** where we design an estimator to keep the error small, which uses subjective loss functions. In contrast, Bayes rule is relatively unambiguous.

For an estimate $\hat{\theta}_n$, e.g. mean, we want it to be **unbiased** $\boldsymbol{E}[\hat{\Theta}] = \theta$, and **consistent** $\hat{\Theta}_n \xrightarrow[n\to\infty]{p} \theta$. We also want the **mean squared error** $\boldsymbol{E}[(\hat{\Theta}_n - \theta)^2]$ to be small. MSE can be decomposed as $\boldsymbol{E}[(\hat{\Theta} - \theta)^2] = \boldsymbol{V}[\hat{\Theta} - \theta] + \boldsymbol{E}[\hat{\Theta} - \theta]^2$, which is the sum of the variance and square of the bias. $\sqrt{\boldsymbol{V}[\hat{\Theta}]}$ is called the standard error of the estimator, used in the calculation of confidence interval $[\hat{\Theta}^-, \hat{\Theta}^+]$, which is a random interval. The general method to find estimates is the **Maximum likelihood estimator** given by $\hat{\theta}_{ML} = argmax_\theta f_X(x;\theta)$. ML estimators are generally consistent, asymptotically normal, and asymptotically efficient (least variance).

Example 3.1. *Consider the class average in an exam in a few different settings. In all cases, assume that we have a large class consisting of equally well-prepared students. Think about the assumptions behind the central limit theorem, and choose the most appropriate response under the given description of the different settings. The options are*

1. *The class average is approximately normal.*
2. *The class average is not approximately normal because the student scores are strongly dependent.*
3. *The class average is not approximately normal because the student scores are not identically distributed.*

- *Consider the class average in an exam of a fixed difficulty.*

  ***Solution**: Since the students are equally well-prepared and difficulty level is fixed, the only randomness is a student's score come from luck or accidental mistakes of the student. It is then plausible to assume that each student's score will be an independent random variable drawn from the same distribution, and the CLT applies. option 1.*

- *Consider the class average in an exam that is equally likely to be very easy or very hard.*

  ***Solution**: Here, the score of each student depends strongly on the difficulty level of the exam, which is random but common for all students. This creates a strong dependence between the student scores, and the CLT does not hold.*

- *Consider the class average if the class is split into two equal-size sections. One section gets an easy exam and the other section gets a hard exam.*

  ***Solution**: The scores of different students are not identically distributed. However, if Y_i be the score of the ith student from the first section and let Z_i bet he score of the ith student in the second section. Then the class average is the average of the random variable $\frac{Y_i+Z_i}{2}$. Under our assumptions, these later random variables are iid, and hence CLT applies.*

- *Consider the class average if every student is randomly and independently given either an easy exam or a hard exam.*

  ***Solution**: Unlike in the second problem above, here the student scores are iid and hence CLT applies.*

4 Stochastic processes

A stochastic process is an infinite sequence of random variables $X_1, X_2, \ldots$. At the fundamental level we are interested in the joint distribution $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ of this infinite series, for any n. The assumption of independence makes it easy to tackle. In the case of a single experiment, the stochastic process represents the realization of a single path from the infinite dimensional sample space Ω, which is a set of infinite sequences.

4.1 The Bernoulli process

Bernoulli process is a sequence of independent Bernoulli trials X_i. At each trial i, $\boldsymbol{P}[X_i = 1] = p$ with $0 < p < 1$. Different trials are **independent** and are **time-homogeneous**, that is the probability law is independent of i. The number of success S in n steps is $S = \sum_{i=1}^{n} X_i$. This is a Binomial distribution,

$$\boldsymbol{P}[S = k] = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n.$$

Thus, $\boldsymbol{E}[S] = np$ and $\boldsymbol{V}ar[S] = np(1-p)$. The **time until the first arrival** is defined as $T_1 = \min\{i : X_i = 1\}$, so

$$\boldsymbol{P}[T_1 = k] = (1-p)^{k-1} p, \ k = 1, 2, \ldots,$$

the familiar Geometric distribution.

The fresh start property, or memorylessness, is a property of this process, due to independence. If we choose N randomly or determined causally (e.g., $N = T_1$, after three consecutive success), then the process $X_{N+1}, X_{N+2}, \ldots$ is a Bernoulli process independent of the previously observed values $X_1, \ldots, X_N$.

Example 4.1. *At each slot, a server is either busy with probability p, or idle with probability $1-p$. What is the distribution of busy periods?*

***Solution**: We first wait for the first busy slot which is geometric with parameter p. Then using memorylessness we thereafter wait for the first idle slot, which should be geometrically distributed with parameter $1-p$. Now the total width of consecutive busy slots is same as the first time of arrival for the first idle slot in the second process. Thus, the distribution of busy periods in the original process is simply Geometric$(1-p)$.* □

To find the **time of k-th arrival** Y_k, can be seen as the sum of inter-arrival times $T_k = Y_k - Y_{k-1}$, for $k \geq 2$. Thus, $Y_k = T_1 + T_2 + \cdots + T_k$. We notice that T_i are independent of each other and have $Geometric(p)$ distribution. Thus, $\boldsymbol{E}[Y_k] = \frac{k}{p}$ and $\boldsymbol{V}ar[Y_k] = \frac{k(1-p)}{p^2}$. And for calculating $\boldsymbol{P}[Y_k = t]$ we notice that if we have a success at slot t we need $k-1$ successes in previous $t-1$ slots and they are independent of each other. Thus, the probability of kth arrival at the step t is given by

$$\boldsymbol{P}[Y_k = t] = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \ t = k, k+1, \ldots.$$

This is called the Pascal distribution.

Say we **merge two independent Bernoulli processes** X_t, Y_t with parameters p, and q respectively, using binary sum to get a new process Z_t. We can easily see the independence of Z_t across time (because X_t is independent across time and Y_t is independent across time, and X_t is independent of Y_t). The probability of success for the process Z_t is $1 - (1-p)(1-q) = p + q - pq$. Thus Z_t is a Bernoulli process with parameter $p + q - pq$. Let us now consider **splitting a Bernoulli process into two**. If we have a Bernoulli process X_t with parameter p. We then split the process, using an independent coin flip with probability q to get two streams Y_t, Z_t, such that $X_t = Y_t + Z_t$. We can easily see the independence property of Y_t, and Z_t. The probability of success for process Y_t is pq and for process Z_t is $p(1-q)$, both of which are Bernoulli processes. Notice, that the two streams Y_t, Z_t are not independent of each other.

The Poisson approximation to the Binomial: We are interested in the case of large n and small p such that $\lambda = np$. We consider the number of arrival S in n slots, $p_S(k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$, for $k = 0, 1, \ldots, n$. We

take $p = \frac{\lambda}{n}$, for fixed k we have $p_S(k) = \lim_{n\to\infty} \frac{n}{n}\frac{n-1}{n}\ldots\frac{n-k+1}{n}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} = \frac{\lambda^k}{k!}e^{-\lambda}$. Thus, the Binomial distribution converges to a Poisson distribution in this special limit.

4.2 The Poisson process

The Poisson process is the continuous time extension of Bernoulli process. In here we don't divide the times into slots but let it be continuous, thus arrivals can happen at any time. The number of arrivals in disjoint time intervals are assumed to be **independent**. We assume **time homogeneity**, i.e. for a given τ, $\boldsymbol{P}[k,\tau]$ is the probability of k arrivals in the interval of duration τ, irrespective of the location of the interval. This implies the relation $\sum_{k=0}^{\infty}\boldsymbol{P}[k,\tau] = 1$. To tackle the situation of multiple arrivals at the same time we introduce one more assumption of **small interval probabilities**. For very small δ we have $\boldsymbol{P}[k,\delta] = \begin{cases} 1-\lambda\delta+\mathcal{O}(\delta^2) & k=0 \\ \lambda\delta+\mathcal{O}(\delta^2) & k=1 \\ 0+\mathcal{O}(\delta^2) & k>1 \end{cases}$. That is, for very small interval we have minuscule probability of more than 1 arrivals, and in that interval it is a Bernoulli process with probability of success $\lambda\delta$, to first order. This introduces λ as the probability per unit time, or the arrival rate. Thus, Poisson process models rare arrival events which are uncoordinated, e.g., radioactive decay, financial market shocks.

If N_τ denotes the arrivals in $[0,\tau]$, then $\boldsymbol{P}[k,\tau] = \boldsymbol{P}[N_\tau = k]$. We divide the interval into n small intervals of length δ with $n = \tau/\delta$ slots. From, the definition of Poisson process we know that the sum of probability that a given slot has ≥ 2 arrivals is $\frac{\tau}{\delta}\mathcal{O}(\delta^2) \xrightarrow[\delta\to 0]{} 0$. Thus, if we neglect the possibility of more than 1 arrival in a slot, we can say that the probability of k arrivals in Poisson process is the same as the probability that k slots have arrival (which is a Binomial distribution and hence as Bernoulli process). Now, the Bernoulli pmf is $p_S(k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0,\ldots,n$. We, thus, have $N_\tau \approx$ *Binomial* with $p = \lambda\delta+\mathcal{O}(\delta^2)$. Thus $np = \lambda\tau+\mathcal{O}(\delta) \approx \lambda\tau$. We showed that, as $n\to\infty$, $p\to 0$ for fixed $k = 0,1,\ldots$ we have $p_S(k) \to \frac{\lambda^k}{k!}e^{-\lambda}$. Thus we have a Poisson distribution

$$\boldsymbol{P}[k,\tau] = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!},\ k = 0,1,\ldots$$

The Poisson process is increasingly accurately described by a Bernoulli process if we have a very fine time discretization. And the approximation becomes exact in the limit.

We can now calculate $\boldsymbol{E}[N_\tau] = \lambda\tau$ and $\boldsymbol{V}ar[N_\tau] = \lambda\tau$. These can be easily derived using the Bernoulli process approximation and taking a limit. Thus, we see that $\lambda = \frac{\boldsymbol{E}[N_\tau]}{\tau}$ is the arrival rate.

The **time until the first arrival**, T_1, which is a continuous variable, has a cdf $\boldsymbol{P}[T_1 \leq t] = 1-\boldsymbol{P}[T_1 > t] = 1-\boldsymbol{P}[0,t] = 1-e^{-\lambda t}$. Thus, we have the pdf as

$$f_{T_1}(t) = \lambda e^{-\lambda t},\ t \geq 0,$$

which is an Exponential distribution with parameter λ. This has memorylessness property, conditioned on $T_1 > t$, the pdf of $T_1 - t$ is again exponential. We can look for the pdf $f_{Y_k}(y)$ of the **time of kth arrival**, Y_k next. We note that $f_{Y_k}(y)\delta \approx \boldsymbol{P}[y \leq Y_k \leq y+\delta] = \boldsymbol{P}[k-1,y]\lambda\delta + \boldsymbol{P}[k-2,y]\mathcal{O}(\delta^2)+\cdots \approx \boldsymbol{P}[k-1,y]\lambda\delta$, where we have a small interval δ at the end when at least one arrival happens. This gives,

$$f_{Y_k}(y;k) = \lambda\boldsymbol{P}[k-1,y] = \frac{\lambda^k y^{k-1}e^{-\lambda y}}{(k-1)!},\ y \geq 0.$$

This is called the Erlang distribution of kth order. For $k = 1$ this reduces to the exponential distribution.

Since Poisson process is a limiting case of Bernoulli process, it also has the fresh start property. Therefore we can see the time of kth arrival Y_k as the sum of inter-arrival times $T_k = Y_k - Y_{k-1}$ for $k \geq 2$. Thus, $Y_k = T_1+T_2+\cdots+T_k$. T_k are independent of each other and have *Exponential*(λ) distribution. Thus, $\boldsymbol{E}[Y_k] = \frac{k}{\lambda}$ and $\boldsymbol{V}ar[Y_k] = \frac{k}{\lambda^2}$ giving us the mean and variance of Erlang distribution.

Example 4.2. *Let X and Y be independent Erlang random variables with common parameter λ and of order m and n, respectively. Is the random variable $X+Y$ Erlang?*

***Solution**: The random variable X can be viewed as the sum of m iid exponential random variables. Similarly, Y can be viewed as the sum of n iid exponential random variables. Furthermore, since X and Y are independent, we take two collections of random variables to be independent. Thus $X+Y$ can be interpreted as the sum of $m+n$ iid exponentials, and is Erlang of order $m+n$.* □

Example 4.3. *Fish are caught as a Poisson process with parameter λ. We fish for two hours and stop if we catch at least one fish in that time period, otherwise we continue until the first fish is caught. Find the expected number of fish caught and the expected time of fishing.*

***Solution**: The pdf for the poison process is $p(k,t) = \frac{(\lambda t)^k e^{-kt}}{k!}$. The total fishing time F can be divided into parts, the first takes exactly 2 hrs, while the second takes $F-2$ hours. Now, $E[F] = 2 + E[F-2]$. Further, we have the cases of $F=2$ and $F>2$, giving us $E[F-2] = P[F=2] \times 0 + P[F>2] \times E[F-2|F>2]$. Notice, $P[F>2] = f(0,2)$. $E[F-2|F>2]$ is simply a new Poisson process starting at time t, so its value is $\frac{1}{\lambda}$. Thus, we have $E[F] = 2 + \frac{p(0,2)}{\lambda}$.*

To find the expected number of fish we again divide it as $T = T_{t\leq 2} + T_{t>2}$. The expected number of fish caught in two hours is given by $E[T_{t\leq 2}] = E[N_2] = 2\lambda$. Under the second scenario you have a stopped process with probability $1-p(0,2)$, while the new Poisson process executes with probability $p(0,2)$. Thus, $E[T_{t>2}] = 0 \times (1-p(0,2)) + 1 \times p(0,2)$. This gives $E[T] = 2\lambda + p(0,2)$. □

We consider the sum of two independent Poisson random variables X and Y, with $\lambda = 1$ for simplicity. For a given Poisson process we take consecutive intervals of length μ and ν with M and N arrivals respectively, which are random variables. Clearly $M \sim Poisson(\mu)$ and $N \sim (\nu)$ and they are independent. Their sum is the original Poisson process with interval $\mu+\nu$ with $M+N$ arrivals. Thus, $M+N \sim Poisson(\mu+\nu)$. This is similar to the fact that sum of independent normal distributions is a normal distribution, a very rare property.

Merging independent Poisson processes produces a Poisson process. If we have X_t and Y_t as two Poisson processes with parameter λ_1 and λ_2. For a given time interval the total arrivals in the merges process is the sum of the arrivals in the two independent processes. Any occurrences across non-intersecting times are independent. For the limiting case of length δ, keeping only linear terms we find that the probability of 0 arrivals is $1-(\lambda_1+\lambda_2)\delta$ and of 1 arrival is $(\lambda_1+\lambda_2)\delta$. Rest of the arrivals have at least an order $\mathcal{O}(\delta^2)$. This establishes the small interval probability property. Thus the merged process is a Poisson process with parameter $\lambda_1+\lambda_2$. To find the probability of the event in the merged stream coming from the first stream is $\frac{\lambda_1}{\lambda_1+\lambda_2}$. Similarly, the probability of the kth arrival coming from the first stream is also given by $\frac{\lambda_1}{\lambda_1+_2}$ (based on independent small interval property of Poisson processes).

Example 4.4. *Say we have three light bulbs with life exponentially distributed with parameter λ, independent of each other. Find the expected time until the first burn out and until all of them burn out.*

***Solution**: We look at three Poisson process with parameter where the bulb's failure represent the first arrival. For the case of first burn out $min(X_1, X_2, X_3)$ we note that the combined process is a Poisson process with parameter 3λ and hence the expected time is $\frac{1}{3\lambda}$. For the case of the all the three bulbs burning out $max(X_1, X_2, X_3)$ we note that the first arrival time for the combined process is $\frac{1}{3\lambda}$. Thereafter we forget about the process where the bulb failed and look at the combined processes of the remaining two bulbs with first time of arrival (i.e. the failure of second bulb), which is a Poisson process with parameter 2λ, as $\frac{1}{2\lambda}$. Finally, the remaining Poisson process will the the first arrival, i.e the third bulb failing in an expected time of $\frac{1}{\lambda}$. This gives the total arrival time of $\frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda} = \frac{11}{6\lambda}$.* □

We want to **Split a Poisson process** arrivals into two different streams using a parameter q for the first stream A and $1-q$ to the second stream B. The splitting and probabilities are independent. The probability of arrival in an interval δ for stream A is $\lambda\delta q$. This shows that the two streams are Poisson with parameters λq and $\lambda(1-q)$. However, since Poisson process is continuous, a point information about stream A does not give any

information about any time segment in stream B, or vise-versa. Thus the two process are independent of each other!

Random incidence paradox: For a Poisson process with parameter λ that runs forever, the average inter arrival time is $\frac{1}{\lambda}$. We can choose a time T randomly and investigate the average time $T - T_1$ before when an incident occurred and the time to next incident $T + T_2$. The sum of these two intervals $T_1 + T_2$, should intuitively, estimate the average inter-arrival time $\frac{1}{\lambda}$. However, when the experiment is performed we get a value of $\frac{2}{\lambda}$. To explain this, we note that the two variables are themselves exponential, and independent, and hence there sum have a mean of $\frac{2}{\lambda}$. How can this be? This is because average inter-arrival time is a different statistical sample than we selecting a time and conducting our experiment. Essentially, we have a higher chance to pick longer intervals so our sampling is different from uniform sampling.

This phenomenon is not specific to Poisson process. Consider an arrival process which are equally probable to be 5 or 10 minutes. If we take an average we get an average inter arrival time of 7.5 minutes. If we randomly choose a point in time and measure the inter arrival interval in which our chosen point falls. Since the 10 minute interval is twice more likely to be picked versus a 5 minute interval we get an average of $\frac{2}{3} \times 10 + \frac{1}{3}5 \approx 8.3$ minutes. This is much more clear now. These calculations can be generalized to **renewal processes** which have iid interarrival times, for some general distribution. If a typical interarrival interval T has probability p_k of having length k, then the probability that the observer sees an interval S of length k is proportional to kp_k. Since the probabilities need to sum to 1, we have $\boldsymbol{P}[S = k] = \frac{kp_k}{\sum_k kp_k} = \frac{kp_k}{\boldsymbol{E}[T]}$. It follows that $\boldsymbol{E}[S] = \sum_k k\frac{kp_k}{\boldsymbol{E}[T]} = \frac{\boldsymbol{E}[T^2]}{\boldsymbol{E}[T]}$. This is true for continuous cases as well and matches the results above.

Different sampling methods can give different results. We will explain this using an example. If half the buses are empty and other half have 50 passengers we have on an average 25 passengers per bus. But if we ask the passengers how crowded their bus is, i.e. the total number of people in their bus you will get an answer of 50. Different samplings lead to different results, because they are measuring different things.

Normal approximation to Poisson process: If we fix p and let $n \to \infty$, we are in the setting where the central limit theorem applies. When we let $n \to \infty, p \to 0$, while keeping the product np fixed, the Poisson approximation applies. For example for a $Binomial(n, p)$ if we fix p and let $n \to \infty$ we get the Normal approximation; while if we keep np fixed, and let $n \to \infty, p \to 0$ we get the Poisson distribution. In the case of $Poisson(n)$ we get the Normal distribution as $n \to \infty$.

4.3 The Markov Process

In a Markov process the current state is sufficient to determine the future state distribution, i.e. we can write $state(t+1) = f(state(t), noise)$. For given finite states X, let X_n be the state after n transitions. Let us denote the initial state is X_0. Using the Markov property the **transition probabilities** are given by

$$p_{ij} = P[X_{n+1} = j | X_n = i].$$

We intend to find the N-step transition probabilities $r_{ij}(n) = P[X_{n+s} = j | X_s = i], \forall s \geq 0$. Notice that $\sum_{j=1}^{m} r_{ij}(n) = 1, \forall i, n$. We notice the recurring relation $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}$. Another recursion for the same can be written as $r_{ij}(n) = \sum_{k=1}^{m} p_{ik}r_{kj}(n-1)$. For the initial random state we have $P[X_n = j] = \sum_{i=1}^{m} P[X_0 = i]r_{ij}(n)$. One would be interested to know what is the long term distribution of states, if it is steady at all, and what is the influence of initial states on the steady behavior?

We are interested in knowing if $r_{ij}(n) \xrightarrow[n\to\infty]{P} \pi_j$? That is, if $r_{ij}(n)$ converge to something? This happens only if periodicity is absent from the chain. The initial state does matter for final distribution if there are portions in the chain not accessible from other parts of the chain. Absent that, the initial states play no role in the final distribution. A state i is called **recurrent** if staring from i, and from wherever you can go, there is a way of returning to i, otherwise it is called **transient**. In the long run the probability of being in transient state goes to zero. We call a group of states a **recurrent class** if they can all communicate with each other. If a chain has more than one classes, the long run probabilities will depend on the initial states. A single recurrent

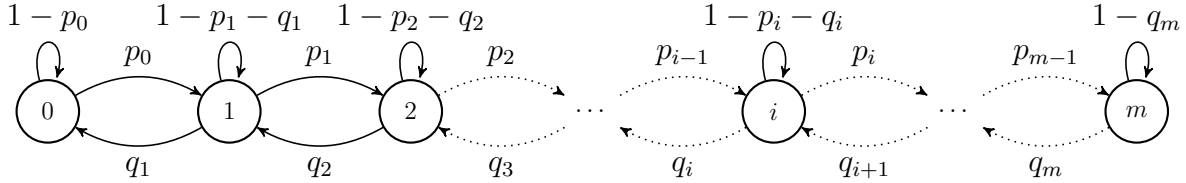
class does not guarantee vanishing dependence of initial state. We need that class to be aperiodic.

The states in a recurrent class as **periodic** if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group. A periodic recurrent class can't have a self transition.

Thus, to answer the question if $r_{ij}(n) = P[X_n = j | X_0 = i]$ converge to some π_j we need to answer under what conditions is there a convergence, and is it independence of initial state i. The main theorem of Markov chain states that if all the recurrent states are all in the single class, and single recurrent class is not periodic then we do have convergence and it is independent of the initial conditions. Thereafter, to calculate this stead sate probability we start from the recursion $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{k_j}$ and take the limit $n \to \infty$ to get the **balance equations** $\pi_j = \sum_{k=1}^{m} \pi_k p_{kj}$ for $j = 1, \ldots, m$. This is a set of m equations with m unknowns. Imposing the **normalization** $\sum_{j=1}^{m} \pi_j = 1$ gives a unique solution.

The frequency interpretation of probability applies to the steady state probability as well. In stead state, π_j represents the fraction of time the Markov system is in state j. Similarly, the fraction of times the transition happens over the edge $i \to j$ is equal to $\pi_i p_{ij}$. Finally, the frequency of transitions into state j is $\sum_k \pi_k p_{kj} = \pi_j$.

Birth-death process is a special Markov chain process, for which convergence exists. Now using the fre-



quency interpretation and concentrating on the interaction between node i and $i+1$ in the long run, we can argue that the number of observations for events $i \to i+1$ should be same as $i+1 \to i$. Thus we can say $\pi_i p_i = \pi_{i+1} q_{i+1}$. Using the normalization of $\sum_{j=1}^{m} \pi_j = 1$ we can find π_0 and then other π_j. Say we have the common parameter $p_i = p$ and $q_i = q$ and we define $\rho = \frac{p}{q}$. For this case we have $\pi_{i+1} = \pi_i \rho$. Thus, we can write $\pi_i = \pi_0 \rho^i, i = 0, 1, \ldots, m$. For $p = q$ (symmetric random walk) we have $\pi_i = \pi_0$, i.e. $\pi_i = \frac{1}{m+1}$. For $n \to \infty$ and $p < q$, the stable system we have $\pi_0 = 1 - \rho$ and $\pi_i = (1-\rho)\rho^i$, which is a shifted geometric distribution. Thus $\boldsymbol{E}[X_n] = \frac{\rho}{1-\rho}$ in steady state.

Extending this to chains with more than one recurrent classes, if we start from within a class we will end up in that class and can use the same procedure as above to calculate π_j for that class, in isolation from the rest of the chain. However, if we start in one of the transient states, it will depend on the probabilities. And hence, the initial conditions matter in this case. How large should n be for the steady state probabilities to take hold; it depends on the **mixing time**. To get an order of magnitude of the mixing time one can look at the average steps needed for a transition in the chain.

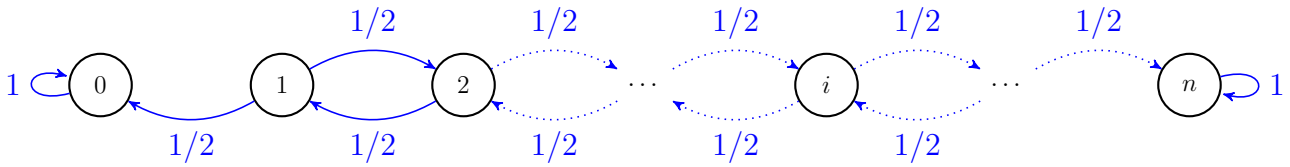
Example 4.5. *We have to establish m phone lines given phone calls are placed with exponential distribution with parameter λ, with each phone call duration follows an exponential distribution with parameter μ. We want to determine m such that no customer has to wait with 99% probability, given $\lambda = 30$ call/min and $\mu = 1/3$ per min.*

***Solution**: We discretize the time into δ length bins small enough so that the probability of more than one event in a bin is $\mathcal{O}(\delta^2)$. Thus, looking this as a merged Poisson process, we have $P[\text{a new call arrives}] \approx \lambda\delta$ and if we have i busy calls, then $P[\text{a departure}] \approx i\mu\delta$. We define the state space as number of lines being used at any bin ranging from 0 to m, as in the birth-death process. We also have $p_i = \lambda\delta$ and $q_i = i\mu\delta$. The balance equation is given by $\pi_{i-1}\lambda\delta = \pi_i i\mu\delta \implies \pi_i = \pi_{i-1}\frac{\lambda}{i\mu}$. This gives, $\pi_i = \pi_0 \frac{\lambda^i}{\mu^i i!}$. Using normalization we can solve it to $\pi_0 = \frac{1}{\sum_{i=0}^{m} \frac{\lambda^i}{\mu^i i!}}$. We want π_m to be less than 1%. This will calculate to 106.* □

The calculation of **absorption probabilities** a_i to an absorption state k with $p_{kk} = 1$ depends on the initial location i. These give a system of linear equations of the form $a_i = \sum_{j=1}^{m} p_{ij}a_j, \forall i$ and can be solved easily. This is simply an application of law of total probability. Similarly, **expected time to absorption** μ_i to each a particular state k from initial location i can be calculated using the law of total expectations. This gives a system of linear equations of the form $\mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j$, for a chain with one absorbing state, which can be solved easily. To calculate the time to absorption to any of the absorbing states, we can club them together into a single absorbing state and apply the calculations.

For the **mean first passage time** from i to s is defined as $t_i = \boldsymbol{E}[min\{n \geq 0 : X_n = s\}|X_0 = i]$, we make the target state an absorbing state by removing any outgoing paths from that state and redo the calculations as above. To calculate **mean recurrence time** of state s defined as $t_s^* = \boldsymbol{E}[min\{n \geq 1 : X_n = s\}|X_0 = i]$, we again use the law of total expectation on the original chain and get the expression $t_s^* = 1 + \sum_{j}^{m} p_{sj}t_j$ where t_j is the the mean first passage time.

Example 4.6. ***Gambler's ruin****: A gambler starts with i dollar. Each time she bets \$1 in a fair game, until she either has 0 or n dollars. What is the probability a_i that she ends up with having n dollars? What is the expected wealth at the end? How long does the gambler expect to stay in game? What if the game is unfair?* ***Solution****: We note the boundary conditions $a_i = 0, a_n = 1$. Further for $0 < i < n$ we have $a_i = \frac{1}{2}a_{i+1} + \frac{1}{2}a_{i-1}$*



which has the solution $a_i = \frac{i}{n}$. To find the expected wealth $0 \times \frac{n-i}{n} + n \times \frac{i}{n} = i$. Thus there is no free lunch. The expectation to absorption is $\mu_0 = \mu_n = 0$ and for $0 < i < n$ we have $\mu_i = 1 + \frac{1}{2}\mu_{i+1} + \frac{1}{2}\mu_{i-1}$ has a solution $\mu_i = i(n-i)$. For an unfair game assume $p < \frac{1}{2}$ of winning, and define $r = \frac{1-p}{p}$. For this case we get $a_i = \frac{1-r^i}{1-r^n}$ and $\mu_i = \frac{r+1}{r-1}\left(i - n\frac{1-r^i}{1-r^n}\right)$. □

Example 4.7. *Mary loves gambling. She starts out with \$200 and keeps playing rounds of the same game. For each round, she can bet either \$100 or \$200, assuming she has sufficient funds, and wins with probability p. Assume that whether she wins is independent across rounds and is unaffected by the size of her bet. If she wins, she receives back double the amount she bet, and if she loses, she gets nothing. Mary stops when she has either \$0 or \$400. What is her optimal betting strategy? That is, the strategy that gives her the greatest probability of reaching \$400.*

Solution*: If Mary has 300 at any point, she can either bet 100 or 200. In case of a loss it is advantageous to have had bet 100, so 100 is the optimal bet if the current state is 300. Also, it is clear that when Mary has 100 she can only bet 100.*

If Mary has 200. If she bets 200 she has a probability of success of p. If she bets 100, and from the previous argument betting 100 at states 100 and 300, we need to find the probability of absorption to the state 400. This gives $a_{100} = a_{200}p, a_{200} = a_{100}(1-p) + a_{300}p, a_{300} = a_{200}(1-p) + p$. This solves to $a_{200} = \frac{p^2}{1-2p+2p^2}$. Betting 200 is better if $p > a_{200} \implies p < \frac{1}{2}$. □