Chapter 0: A general introduction to the course

The example considers the study of the impact advertising expenditure may have on the interest of a product, which is measured by the percentage of π of people who remembered having watched the advertisement on TV. A common model for this example is logistic regression as the following:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta\mathbf{x} \qquad (0.1)$$

where $\mathbf{x}$ is the expenditure of the advertisement. The larger the value of β, the more effective the advertisement campaign is in boosting awareness.

Once the model is built, our task is to use methods in statistics to estimate α, β. In the course 'Applied Statistics', we learnt linear regression method, i.e., given observed data $\{(\pi_i, x_i)\}_{1\leqslant i\leqslant n}$, where i denotes some place. we use least square method as the following:

$$\min_{\alpha,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{\pi_i}{1-\pi_i} - \alpha - \beta x_i\right]^2\right\}.$$

In this course, we shall study the inference problem of α, β via Bayesian approach, in which we take the observed data $\{(\pi_i, x_i)\}_{1\leqslant i\leqslant n}, \alpha, \beta$ as random variables. We call the parameters α, β (to be estimated) as unknowns, and the observed data $\{(\pi_i, x_i)\}_{1\leqslant i\leqslant n}$ as knowns. The distribution of (α, β) given the known is called the **posterior distribution**. Obtaining posterior distribution is an important step but not the final one, one must draw meaningful information from it.

The model (0.1) is extremely simple but usually fails to describe the real situation, for instance, the parameters (α, β) may change with the time t, in which the

model needs to be modified as

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \alpha_t + \beta_t \mathbf{x}_t.$$

Now, the number of unknown quantities has risen dramatically from 2 to $2n$ if our observations last for n, consequently the inference problem will be much more complex to handle.

For complex problems in statistics, people often develop approximation methods to get a not exact but acceptable solution. In this course, we only study **stochastic approximation**, in which stochastic simulation plays a central role. We will learn techniques for sampling random variables and Monte Carlo (MC) method.

However, a simple MC is not enough to handle high dimensional statistical problems. So we will study a much more advanced stochastic simulation, Markov Chain Monte carlo (MCMC), to sample high dimensional random variables. We shall learn two typical sampling methods: Gibbs sampling, Metropolis-Hastings algorithms.

Chapter 1. stochastic Simulation

1.1. Generation of discrete random quantities

We start from a random quantity U generating from uniform distribution on the interval $[0, 1]$, denoted by $U \sim U([0, 1])$. U can be easily sampled by randomly choosing a number from $[0, 1]$.

In the generic case of a random quantity x assigning values $\{x_1, \cdots, x_k\}$ with respective probabilities $p_1, \cdots p_k$ subject to $p_1 + \cdots + p_k = 1$, the interval $[0, 1]$ is split

into k intervals $I_1, \cdots, I_k$ with $I_i = (F_{i-1}, F_i]$ where $F_0 = 0, F_i = p_1 + \cdots + p_i$ for $1 \leq i \leq k$. We take a random quantity u from $U([0,1])$, if $u \in I_i$, the random quantity X takes the value x_i. We can see that $P(X = x_i) = P(u \in I_i) = F_i - F_{i-1} = p_i$.

- Bernoulli distribution

$$P(X=1) = p \quad P(X=0) = 1-p.$$

$$k = 2, x_1 = 1, x_2 = 0, F_0 = 0, F_1 = p, F_2 = 1.$$

- Binomial distribution $\text{Bin}(n,p)$

$$p_i = P(X=i) = \begin{pmatrix} n \\ i \end{pmatrix} p^i (1-p)^{n-i}, i = 0, 1, \cdots, n.$$

$$k = n+1; x_1 = 0, x_2 = 1, \cdots, x_k = n$$

$$F_0 = 0, F_i = p_1 + \cdots + p_i \text{ for } i = 1, 2, \cdots, n+1$$

1.2. Generation of continuous random quantities

1.2.1. Probability integral transforms

Consider drawing a random quantity X from a continuous probability distribution with the distribution function F. We know F is a continues nondecreasing function if F has an inverse F^{-1}, then $X = F^{-1}(U)$, where U is a random quantity drawn from $U([0,1])$, is a random quantity as desired. Indeed,

$$P(X \leqslant z) = P\left(F^{-1}(U) \leqslant z\right) = P(U \leqslant F(z)) = F(z), \forall z \in \mathbb{R}$$

Example: Exponential distribution Exp(1).

Exp (1) has a probability density function: $f(z) = \begin{cases} e^{-z}, & z \geqslant 0, \\ 0, & z < 0. \end{cases}$

Its distribution function is $F(z) = \begin{cases} 1 - e^{-z}, & z \geqslant 0, \\ 0, & z < 0. \end{cases}$

We only need to concentrate on $F(z)$ on $[0, \infty)$, and have

$$F^{-1}(z) = -\log(1 - z).$$

So $F^{-1}(U) = -\log(1-U)$ has a probability distribution Exp (1). Because $1 - U \sim U([0, 1])$, we have $-\log U \sim \text{Exp}(1)$.

For a distribution function which does not have an inverse, we define a generalized inverse as the following:

$$F^{-}(z) = \inf\{x \in \mathbb{R} : F(x) \geqslant z\}.$$

Assignment: Sample a random quantity $Z \sim \text{Exp}(\lambda)$ for some $\lambda > 0$.

1.2.2. Bivariate techniques

If $X = (X_1, X_2)$ has a joint density $f_X(x_1, x_2)$ and $g_{\cdot}(x_1, x_2) \to (y_1, y_2)$ is a differentiable transform with the inverse $(x_1, x_2) = g^{-1}(y_1, y_2)$. Then the random vector $(Y_1, Y_2) = g(X_1, X_2)$ has a density

$$f_Y(y_1, y_2) = f_X\left(g^{-1}(y_1, y_2)\right) J(y_1, y_2)$$

where $J(y_1, y_2)$ is the Jacobi matrix $J(y_1, y_2) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$.

Example: Let $U_1 \sim U([0, 1]), U_2 \sim U([0, 1])$ be independent, define

$$x_1 = \sqrt{-2\log u_1}\cos(2\pi u_2), x_2 = \sqrt{-2\log u_1}\sin(2\pi u_2)$$

i.e. $(x_1, x_2) = g(u_1, u_2) = \left(\sqrt{-2\log u_1}\cos(2\pi u_2), \sqrt{-2\log u_1}\sin(2\pi u_2)\right)$.

$$g^{-1}(x_1, x_2) = \left(\exp\left(-\frac{x_1^2 + x_2^2}{2}\right), \frac{1}{2\pi}\tan^{-1}\left(\frac{x_2}{x_1}\right)\right),$$

$$f_U(u_1, u_2) = 1 \text{ for } (u_1, u_2) \in [0,1] \times [0,1].$$

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial u_1}{\partial x_1} & \frac{\partial u_2}{\partial x_2} \\ \frac{\partial u_1}{\partial x_2} & \frac{\partial u_2}{\partial x_2} \end{vmatrix} = \begin{vmatrix} \frac{\partial g_1^{-1}(x_1,x_2)}{\partial x_1} & \frac{\partial g_2^{-1}(x_1,x_2)}{\partial x_1} \\ \frac{\partial g_1^{-1}(x_1,x_2)}{\partial x_2} & \frac{\partial g_2^{-1}(x_1,x_2)}{\partial x_2} \end{vmatrix} = \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}$$

Hence $f_X(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$.

Example: Ratio of uniforms methods (Not required in theory, but required in python program)

In data science, we often come across a distribution whose density is known up to a constant, e.g. a random variable X with a density

$$f_X(x) = \frac{1}{C}e^{-x^4}$$

the constant $C = \int_{-\infty}^{+\infty} e^{-x^4}dx$ is not known. Taking

$$f^*(x) = e^{-x^4},$$

define $C_f = \left\{(x_1, x_2) : 0 \leqslant x_1 \leqslant \sqrt{f^*(x_2/x_1)}\right\}$. Let (U_1, U_2) be the random vector uniformly distributed on C_f. Then $Y = \frac{U_2}{U_1}$ has a density $\frac{f^*(y)}{\int_{-\infty}^{+\infty} f^*(y)dy}$.
Indeed, let $x = u_1, y = \frac{u_2}{u_1}$, i.e.

$$(x, y) = g(u_1, u_2) = \left(u_1, \frac{u_2}{u_1}\right).$$

$$(u_1, u_2) = g^{-1}(x, y) = (x, xy).$$

$$J(x, y) = \begin{vmatrix} \frac{\partial g_1^{-1}(x,y)}{\partial x} & \frac{\partial g_2^{-1}(x,y)}{\partial x} \\ \frac{\partial g_2^{-1}(x,y)}{\partial y} & \frac{\partial g_2^{-1}(x,y)}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & y \\ 0 & x \end{vmatrix} = x.$$

Let $f_{(X,Y)}(x,y) = x \cdot k$, where $k = \frac{1}{\text{Area}\left(C_f\right)}$, since $f\left(U_1,U_2\right)\left(u_1,u_2\right) = k$ for $(u_1,u_2) \in C_f$. Now we consider the distribution of Y, which is given by

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{+\infty} f_{(X,Y)}(x,y)dx \\ &= \int_{-\infty}^{+\infty} x \cdot k 1_{C_f} dx \\ &= \int_0^{\sqrt{f^*(y)}} kxdx \\ &= kf^*(y) = \frac{f^*(y)}{\int_{-\infty}^{+\infty} f^*(y)dy}. \end{aligned}$$

Assignment: Make a python or R program to generate 1000 random quantities from the probability distribution with a density $f_X(x) = \frac{1}{C}e^{-\frac{|x|^5}{2}}$ with $C = \int_{-\infty}^{+\infty} e^{-\frac{|x|^5}{2}}dx$.

1.2.3. Methods based on mixture

Example 1: Mixture of Gaussian distribution

Let $X_1 \sim N\left(\mu_1,\sigma_1^2\right), X_2 \sim N\left(\mu_2,\sigma_2^2\right)$ with $\sigma_1 > 0$ and $\sigma_2 > 0$, and let X_1 and X_2 be independent. Let $\Theta \sim \text{Ber}(p)$ with $p \in (0,1)$ and let Θ be independent of X_1 and X_2. Then, $(\Theta)X_1 + (1-\Theta)X_2$ is a simplest Gaussian mixture model.

Assignment: $\Theta X_1 + (1-\Theta)X_2$ has a probability density

$$\frac{p}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-p}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right).$$

(Hint: $P\left(\Theta X_1 + (1-\Theta)X_2 \leqslant x\right) = P\left(X_1 \leqslant x \mid \Theta = 1\right)P(\Theta = 1) + P\left(X_2 \leqslant x | \Theta = 0\right)P(\Theta = 0)$ and the independence of Θ of X_1 and X_2.)

We now aim to sample a random quantity from the distribution of the above

Gaussian Mixture Model.

Pseudo-code

- Sample a random quantity $\Theta \sim \text{Ber}(p)$

  If $\Theta = 1$, sample a random quantity $X \sim N(\mu_1, \sigma_1^2)$,

  If $\Theta = 0$, sample a random quantity $X \sim N(\mu_2, {\sigma_2}^2)$.

- Return $Y = X$.

Assignment: Realize the above psendocode by programming to create 1000 random quantities, and draw a histogram. (Hint: To draw a random quantity X from $N(\mu, \sigma^2)$, first draw a random quantity $Z \sim N(0,1)$ and take $X = \mu + \sigma Z$.)

Example 2: t-distribution.

The t-distribution arises as the sampling distribution of the t-test. Let $X_1, \cdots, X_n$ be i.i.d with the distribution $N(\mu, \sigma^2)$, the sample mean and the sample variance are respectively as the below:

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n), \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

The t-statistic is defined as

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}},$$

T has a t-distribution with a degree $n-1$, whose density function reads as

$$f(x) = \frac{\tau\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\tau\left(\frac{n-1}{2}\right)}\left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

where $\tau(\cdot)$ is defined as $\tau(z) = \int_0^\infty x^{z-1}e^{-x}dx$ for $z > 0$.

How to sample random quantity Z with t-distribution with $(n-1)$ degree?

A remarkable theorem about T is

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}}$$

where $Z \sim N(0,1)$ and $Y \sim \chi^2(n-1)$, chi-square distribution with a degree $n-1$, moreover Z and Y are independent.

Pseudo-Code:

- Sample a random quantity Y from the distribution $\chi^2(n-1)$,
- Sample a random quantity $Z \sim N\left(0, \frac{Y}{n-1}\right)$
- Return $X = Z$.

Assignment: Realize the above pseudo-code to generate 1000 random quantities by programming, and draw a histogram.

1.3. Resampling methods

Resampling method consists of two steps, the first one provides a random quantity from an approximate distribution, while the second is a correction mechanisms. In what follows, we denote by f the target distribution and by g the auxiliary distribution.

We only study the classical accept-rejection method in this course.

We aim to draw a random quantity X from a distribution whose probability density is f, which may be difficult to be drawn. We introduce an auxiliary probability

distribution g and take the following assumption:

(A) Assume that there exists some $M > 0$ such that

$$\frac{f(x)}{g(x)} \leq M \quad \text{for all } x.$$

Pseudo code for accept-rejection method:

- Sample a random quantity Y from the distribution g
- Sample a random quantity U from $U([0,1])$.

  If $U \leqslant \frac{f(Y)}{Mg(Y)}$, accept, ie. $X = Y$

  If $U > \frac{f(Y)}{Mg(Y)}$, reject and repeat the procedure.

Theoretical proof for this algorithm: Let us consider the one dimension case. For any $x \in \mathbb{R}$, consider

$$\begin{aligned}
P(X \leqslant x) &= P\left(Y \leqslant x \mid U \leq \frac{f(Y)}{Mg(Y)}\right) \\
&= \frac{P\left(Y \leqslant x, U \leq \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Mg(Y)}\right)} \\
&= \frac{\int_{-\infty}^{x}\left(\int_0^{\frac{f(Y)}{Mg(y)}} du\right) g(y)dy}{\int_{-\infty}^{+\infty}\left(\int_0^{\frac{f(y)}{Mg(y)}} du\right) g(y)dy} \left(\because \frac{f(y)}{Mg(y)} \leqslant 1 \text{ for all } y\right) \\
&= \frac{\int_{-\infty}^{x} f(y)dy}{\int_{-\infty}^{+\infty} f(y)dy}
\end{aligned}$$

So, X has a distribution with the density f.

Assignment: Repeat the proof of the algorithm.

Remark 1. *Accept-rejection method can be used to sample a random quantity whose distribution is known up to a constant. An example is to sample a random quantity X from the distribution with the following density*

$$\frac{1}{C} \cdot \frac{1}{1+|x-2|^3} \qquad (1.3.1)$$

where $C = \int_{-\infty}^{+\infty} \frac{1}{1+|x-2|^3} dx$, which can not be explicitly computed out.

Assignment: Write the pseudo-code for sampling a random quantity from the distribution (1.3.1) via accept-rejection method (Hint: take $M = 5$). Make a program to tun the accept-rejection procedure 1000 times, count how many random quantities are created, and draw a histogram.

Chapter 2. Bayesian Inference

2.1. Bayes' Theorem

In statistics, we often need to use the observed data $x_1, \ldots, x_n$ to estimate the parameter θ in a model by some statistical methods such as maximum likelihood function.

Example 2.1: Consider a series of measurements about a physical quantity μ with measurement error $\varepsilon_i \sim N(0, \sigma^2)$, where σ is not known. In this case, the measurements are

$$x_i = \mu + \varepsilon_i, \quad i = 1, 2, \cdots n.$$

The likelihood function of $\theta = (\mu, \sigma)$ is

$$\ell(\theta) = \prod_{i=1}^{n} P_\theta(x_i) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right].$$

The maximum likelihood estimator is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \ell(\theta)$$

In this estimation, we only consider it from the observed data $x_1, \cdots, x_n$.

However, in practice, we may have some prior knowledge about θ, which is natural to be taken into account in the estimate. We need to use Bayes' theorem.

2.1.1. Prior, posterior and predictive distributions

For the parameter θ to be estimated, we assign it a prior distribution $P(\theta)$ according to our prior knowledge about it.

Given a θ, the observed data x has a conditional distribution $f(x \mid \theta)$. The distribution of θ after observing x is denoted by $p(\theta \mid x)$. By Bayes' Theorem,

$$p(\theta \mid x) = \frac{f(x \mid \theta)p(\theta)}{f(x)}$$

where $f(x) = \int f(x \mid \theta)p(\theta)d\theta$. Indeed,

$$\begin{aligned} p(\theta \mid x) = \frac{f(x,\theta)}{f(x)} &= \frac{f(x \mid \theta)p(\theta)}{\int_{-\infty}^{+\infty} f(x,\theta)d\theta} \\ &= \frac{f(x \mid \theta)p(\theta)}{\int_{-\infty}^{+\infty} f(x \mid \theta)p(\theta)d\theta}. \end{aligned}$$

Because $f(x)$ does not depend on θ, we may write

$$p(\theta \mid x) \propto f(x \mid \theta)p(\theta)$$

In statistical models, $f(x \mid \theta)$ is the likelihood function $\ell(\theta)$ given the observed data x. ($e.g.$ $x = (x_1, \cdots, x_n)$ in Example 2.1). We usually suppress the dependence on x and write

$$\ell(\theta) = f(x \mid \theta), \pi(\theta) = p(\theta \mid x).$$

Then

$$\pi(\theta) \propto \ell(\theta)p(\theta)$$

Example 2.1 (continued) Suppose that σ^2 is known, so $\theta = \mu$ is the parameter to be estimated. Let us consider this model in Bayesian inference setting. Suppose that a prior distribution $p(\mu) = \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{(\mu-\mu_0)^2}{2\tau_0^2}}$, where μ_0 and τ_0 are both known. The likelihood function is

$$\begin{aligned}
\ell(\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&\propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right)
\end{aligned}$$

where $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$, and the second $\propto$ is due to

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 &= n\mu^2 - 2\sum_{i=1}^{n} x_i \cdot \mu + \sum_{i=1}^{n} x_i^2 \\
&= n\left(\mu^2 - 2\bar{x}\mu\right) + \sum_{i=1}^{n} x_i^2 \\
&= n(\mu-\bar{x})^2 + \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\
&= n(\mu-\bar{x})^2 + n\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)}_{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{\mu})^2}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\pi(\mu) &= \ell(\mu)p(\mu) \\
&\propto \exp\left(-\frac{1}{2}\cdot\frac{(\mu-\bar{x})^2}{\sigma^2/n}\right)\exp\left(-\frac{1}{2}\cdot\frac{(\mu-\mu_0)^2}{\tau_0^2}\right) \\
&= \exp\left(-\frac{1}{2}\cdot\frac{(\mu-\mu_1)^2}{\tau_1^2} - \frac{1}{2}\cdot\frac{(\bar{x}-\mu_0)^2}{n^{-1}\sigma^2+\tau_0^2}\right)
\end{aligned}$$

where $\tau_1^{-2} = n\sigma^{-2} + \tau_0^{-2}, \mu_1 = \tau_1^2 \left(n\sigma^{-2}\bar{x} + \tau_0^{-2}\mu_0\right)$. So

$$\pi(\mu) \propto \exp\left(-\frac{1}{2} \cdot \frac{(\mu - \mu_1)^2}{\tau_1^2}\right),$$

this means that the posterior distribution of μ is $N(\mu_1, \tau_1^2)$ with

$$\mu_1 = \tau_1^2 \left(n\sigma^{-2}\bar{x} + \tau_0^{-2}\mu_0\right), \tau_1^2 = \frac{1}{n\sigma^{-2} + \tau_0^{-2}}.$$

So, we can see that the information of the data has effect on the parameter μ.

If we increase τ_0, we get less information from the prior distribution $p(\mu)$.

2.1.2. Summarizing the information from the posterior distribution.

Once the posterior distribution is available, one may seek to summarize its information through a few elements such as the location and dispersion measures. The main location measures are mean, mode and median, while the main dispersion measures are variance, standard deviation, precision, interquantile range and curvature at the mode.

Let θ be a scalar, we give the definitions of the above concepts:

- Posterior mean $= \int \theta\pi(\theta)d\theta$
- Posterior variance $= \int \left(\theta - \int \theta\pi(\theta)d\theta\right)^2 \pi(\theta)d\theta$
- Posterior median is the median of the distribution π, ie. a point θ_0 such that

$$\int_{-\infty}^{\theta_0} \pi(\theta)d\theta = \frac{1}{2}.$$

- Credibility interval C : given an $\alpha > 0$ (e.g. $\alpha = 0.05$), an interval C is called $100(1-\alpha)\%$ credibility interval for θ if $\int_C \pi(\theta)d\theta = 1-\alpha$. We usually look for a credibility interval with the shortest length. This interval is called highest posterior density (HPD) interval.

Example 2.1 (continued) We have shown that the posterior distribution of $\mu \sim N(\mu_1, \tau_1^2)$. So the posterior mean, median and mode are all μ_1, the posterior variance is τ_1^2 and the HPD interval of $(1-\alpha)$ is $\left(\mu_1 - \tau_1 z_{\alpha/2}, \mu_1 + \tau_1 z_{\alpha/2}\right)$ where $z_{\alpha/2}$ satisfies

$$\int_{z_{\alpha/2}}^{\infty} \varphi(x)dx = \frac{\alpha}{2}$$

2.2. A brief discussion on conjugate distributions

From the above Example 2.1, we choose the prior distribution $p(\mu) \propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right)$, which makes the derivation of the posterior distribution very simple as a change of parameters:

$$\tau_1^{-2} = n\sigma^{-2} + \tau_0^{-2}, \quad \mu_1 = \tau_1^2\left(n\sigma^{-2}\bar{x} + \tau_0^{-2}\mu_0\right)$$

Example 2.1 (continued). After observing n_1 samples $x_1^{(1)}, x_2^{(1)}, \cdots, x_{n_1}^{(1)}$ from the population with the distribution $N(\mu, \sigma^2)$, we obtain a posterior distribution about μ as $N(\mu_1, \tau_1^2)$ with $\tau_1^{-2} = n_1\sigma^{-2} + \tau_0^{-2}, \mu_1 = \tau_1^2\left(n_1\sigma^{-2}\bar{x}^{(1)} + \tau_0^{-2}\mu_0\right)$, with $\bar{x}^{(1)} = \frac{1}{n_1}\left(x_1^{(1)} + \cdots + x_{n_1}^{(1)}\right)$. Now a new sequence of samples $x_1^{(2)}, \cdots, x_{n_2}^{(2)}$ are observed, we take $N(\mu_1, \tau_1^2)$ as the prior distribution, and get the new posterior distribution as $N(\mu_2, \tau_2^2)$ with

$$\tau_2^{-2} = n_2\sigma^{-2} + \tau_1^{-2}, \mu_2 = \tau_2^2\left(n_2\sigma^{-2}\bar{x}^{(2)} + \tau_1^{-2}\mu_1\right),$$

where $\bar{x}^{(2)} = \frac{1}{n_2}\left(x_1^{(2)} + \cdots + x_{n_2}^{(2)}\right)$. In other words, the new posterior distribution is obtained based on the observation $x^{(1)}$ and $x^{(2)}$.

Because we take the prior distribution $p(\theta) \sim N(\mu_0, \tau_0^2)$, we can obtain the posterior distribution $\pi(\theta) \sim N(\mu_1, \tau_1^2)$ and continue the update of posterior distributions for the data flow $x^{(1)}, x^{(2)}, \cdots, x^{(n)}, \cdots$ in the following way:

$$\tau_{i+1}^{-2} = n_{i+1}\sigma^{-2} + \tau_i^{-2}, \mu_{i+1} = \tau_{i+1}^2 \left(n_{i+1}\sigma^{-2}\bar{x}^{(i+1)} + \tau_i^{-2}\mu_i\right)$$

This $p(\theta)$ is a conjugate distribution w.r.t. $\ell(\theta)$. The conjugate distributions in Bayesian inference is a deep theory, and we will not enter it but give several examples.

Example 2.2 Consider the Poisson model: Let $x_1, x_2 \ldots\ x_n$ be a sequence of observation from Poisson distribution Poi (λ). Given a λ,

$$\ell(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Given a prior distribution $p(\lambda) \sim G(\alpha, \beta)$, gamma distr. with the parameter α, β, ie. $p(\lambda) = \lambda^{\alpha-1}e^{-\beta\lambda}$.

$$\begin{aligned}
\pi(\lambda) &\propto \ell(\lambda)p(\lambda) \\
&\propto \prod_{i=1}^{n} \left(\lambda^{x_i}e^{-\lambda}\right) \lambda^{\alpha-1}e^{-\beta\lambda} \\
&= \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-n\lambda - \beta\lambda}
\end{aligned}$$

i.e. $\pi(\lambda) \propto \lambda^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-n\lambda - \beta}$. So $\pi(\lambda) \sim G\left(\sum_{i=1}^{n} x_i + \alpha, n\lambda + \beta\right)$. We say $G(\alpha, \beta)$ is a conjugate distribution to Poisson model.

Example 2.3: Consider Bernoulli model: Let $x_1, \cdots, x_n$ be a sequence of observations from the Bernoulli distr. Ber(θ). Then,

$$\ell(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

(recall $x_i \in \{0,1\}$). Take the prior distribution $P(\theta) \sim \text{Beta}(\alpha, \beta)$ with $\alpha > 0, \beta > 0$, i.e.

$$p(\theta) = \frac{\tau(\alpha+\beta)}{\tau(\alpha)\tau(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1].$$

So

$$\begin{aligned}\pi(\theta) &\propto \ell(\theta)p(\theta)\\ &\propto \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{\sum_{i=1}^n x_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n x_i+\beta-1}.\end{aligned}$$

i.e. $\pi(\theta) \sim \text{Beta}\left(\sum_{i=1}^n x_i + \alpha, n - \sum_{i=1}^n x_i + \beta\right)$.

2.3. Dynamic models (Time series)

Dynamic linear models are defined by a pair of equations, called the observation equation and the system equation, respectively given by

$$y_t = F_t'\beta_t + \varepsilon_t, \qquad \varepsilon_t \sim N\left(0, \sigma_t^2\right), \quad (2.3.1)$$
$$\beta_t = G_t\beta_{t-1} + \omega_t, \quad \omega_t \sim N\left(0, W_t\right), \quad (2.3.2)$$

where $y_t \in \mathbb{R}$ is an observation at the time $t, F_t \in \mathbb{R}^d, \beta_t \in \mathbb{R}^d$, $G_t \in \mathbb{R}^{d\times d}$, and $W_t \in \mathbb{R}^{d\times d}$ is the covariance matrix. Moreover. the errors ε_t and ω_t are independent.

Example 2.4 The simplest time series model is the first order model, which is given by

$$\begin{aligned} y_t &= \beta_t + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_t^2\right)\\ \beta_t &= \beta_{t-1} + \omega_t \quad \omega_t \sim N\left(0, W_t\right)\end{aligned}$$

where $\beta_t \in \mathbb{R}$ and $d = 1$.

A typical linear growth model is as the following:

$$
\begin{aligned}
y_t &= \beta_{1,t} + \varepsilon_t \quad \varepsilon_t \sim N\left(0, \sigma_t^2\right) \\
\beta_{1,t} &= \beta_{1,t-1} + \beta_{2,t-1} + \omega_{1,t} \\
\beta_{2,t} &= \beta_{2,t-1} + \omega_{2,t} \\
\omega_t &= (\omega_{1,t}, \omega_{2,t}) \sim N\left(0, \omega_t\right)
\end{aligned}
$$

this model is a special case of (2.3.1)-(2.3.2) with $F_t = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $G_t = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ for all t therein.

Sequential inference: Denote by y^0 the initial information available, and $y^i = \{y^0, y_1, \cdots, y_i\}$. For dynamic model, the inference is based on the updated distribution $\beta_t \mid y^t$. There are three basic operations involved here: evolution, prediction and updating. Let us demonstrate these three operations as below.

Consider at the time $t-1$, the updated distribution is

$$\beta_{t-1} \mid y^{t-1} \sim N\left(m_{t-1}, C_{t-1}\right)$$

Recall the system equation (2.3.2): $\beta_t = G_t \beta_{t-1} + \omega_t$, from which we know

$$\beta_t \mid \beta_{t-1} \sim N\left(G_t \beta_{t-1}, W_t\right)$$

By a standard calculation (not required in this course), we know

$$\beta_t \mid y^{t-1} \sim N\left(a_t, R_t\right)$$

where $a_t = G_t m_{t-1}$ and $R_t = G_t C_{t-1} G_t' + W_t$.

Assignment: Let

$$
\begin{aligned}
y \mid x &\sim N\left(\mu_1, \sigma_1^2\right) \\
z \mid y &\sim N\left(\mu_2 y, \sigma_2^2\right)
\end{aligned}
$$

where $\mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}$ and $\sigma_1^2 > 0, \sigma_2^2 > 0$. Show that

$$z \mid x \sim N\left(\mu_1\mu_2, \sigma_1^2\mu_2^2 + \sigma_2^2\right).$$

Now we would like to compute the updating distribution $\beta_t \mid y^t$.

$$\begin{aligned} p\left(\beta_t \mid y^t\right) &= p\left(\beta_t \mid y_t, y^{t-1}\right) \\ &= \frac{p\left(\beta_t, y_t, y^{t-1}\right)}{p\left(y_t, y^{t-1}\right)} = \frac{f\left(y_t \mid \beta_t\right) p\left(\beta_t \mid y^{t-1}\right) p\left(y^{t-1}\right)}{p\left(y_t, y^{t-1}\right)} \\ &= \frac{p\left(y^{t-1}\right)}{p\left(y_t, y^{t-1}\right)} f\left(y_t \mid \beta_t\right) p\left(\beta_t \mid y^{t-1}\right) \\ &\propto f\left(y_t \mid \beta_t\right) p\left(\beta_t \mid y^{t-1}\right) \end{aligned}$$

Chapter 3. Approximate method of inference

Recall $\pi(\theta)$ is the posterior distribution, we would like to draw the useful information from $\pi(\theta)$, such as posterior mean, posterior median and credibility region, which can be generally summarized as

$$J = \int t(\theta)\pi(\theta)d\theta$$

e.g. $t(\theta) = \theta$, one obtains the posterior mean, $t(\theta) = 1(\theta + C)$ and $J = 1 - \alpha$ give the credibility regime. However, it is often difficult to compute this integral, so people have to develop approximation methods. There are two type of approximation methods, one deterministic approximation and the other stochastic approximation.

3.1. Asymptotic approximation

3.1.2. Normal approximation

Let m be the mode of the distribution $\pi(\theta)$, i.e.

$$m = \underset{\theta}{\text{argmax}}\pi(\theta),$$

note that there may be more then one modes. Consider the Taylor expansion of $\log \pi(\theta)$ at $\theta = m$, we get

$$\begin{aligned}\log \pi(\theta) =& \log \pi(m) + (\nabla \log \pi(m))'(\theta - m) \\ &+ \frac{1}{2}(\theta - m)' \left(\nabla^2 \log \pi(m)\right)(\theta - m) + R(\theta). \\ =& \log \pi(m) + \frac{1}{2}(\theta - m)' \left(\nabla^2 \log \pi(m)\right)(\theta - m) + R(\theta)\end{aligned}$$

Since $\nabla \log \pi(m) = 0$. Note that $\nabla^2 \log \pi(m)$ is the Hessian matrix of $\log \pi(\theta)$ at $\theta = m$. More precisely, for $\theta = (\theta_1, \cdots, \theta_d)$,

$$\nabla \log \pi(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \log \pi(\theta) \\ \vdots \\ \frac{\partial}{\partial \theta_d} \log \pi(\theta) \end{bmatrix} \in \mathbb{R}^d,$$

$$\nabla^2 \log \pi(\theta) = \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \log \pi(\theta) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \pi(\theta) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_d} \log \pi(\theta) \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \pi(\theta) & \frac{\partial^2}{\partial \theta_2^2} \log \pi(\theta) & \cdots & \frac{\partial^2}{\partial \theta_2 \partial \theta_d} \log \pi(\theta) \\ \cdots & \cdots & & \\ \frac{\partial}{\partial \theta_1 \partial \theta_d} \log \pi(\theta) & \frac{\partial^2}{\partial \theta_2 \partial \theta_d} \log \pi(\theta) & \cdots & \frac{\partial^2}{\partial \theta_d^2} \log \pi(\theta) \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

$\nabla^2 \log \pi(\theta)$ is a symmetric matrix and $\nabla^2 \log \pi(m) = (\nabla^2 \log \pi(\theta))|_{\theta=m}$. We drop the remainder $R(\theta)$ and write

$$\begin{aligned}\log \pi(\theta) &\approx \log \pi(m) + \frac{1}{2}(\theta - m)' \left(\nabla^2 \log \pi(m)\right)(\theta - m) \\ &= \log \pi(m) - \frac{1}{2}(\theta - m)' \left(-\nabla^2 \log \pi(m)\right)(\theta - m),\end{aligned}$$

which implies

$$\pi(\theta) \propto \exp\left\{-\frac{1}{2}(\theta - m)' \left(-\nabla^2 \log \pi(m)\right)(\theta - m)\right\}$$

Denote $\Sigma = \left(-\nabla^2 \log \pi(m)\right)^{-1}$, then

$$\pi(\theta) \propto \exp\left(-\frac{1}{2}(\theta - m)'\Sigma^{-1}(\theta - m)\right).$$

Example 3.1: Let $\pi(\theta) = \frac{1}{C}\exp\left(-\frac{1}{2}(\theta + \sin\theta)^2\right)$ with $C = \int \exp\left(-\frac{1}{2}(\theta + sin\theta)^2\right) d\theta$

$$\begin{aligned}\log \pi(\theta) &= \log C - \frac{1}{2}(\theta + \sin\theta)^2. \quad m = 0 \\ (\log \pi(\theta))' &= -(\theta + \sin\theta)(1 + \cos\theta), \\ (\log \pi(\theta))'' &= (\theta + \sin\theta)\sin\theta - (1 + \cos\theta)^2 \\ (\log \pi(0))'' &= -(1 + \cos 0)^2 = -4.\end{aligned}$$

So we have the following normal approximation for $\pi(\theta)$:

$$\tilde{\pi}(\theta) \propto \exp\left(-2\theta^2\right)$$

Assignment: Let $\pi(\theta) = \frac{1}{C}\exp\left(-\frac{1}{2}(\theta + \cos\theta)^2\right)$ where $C = \int e^{-\frac{1}{2}(\theta+\cos\theta)^2} d\theta$. Derive its normal approximation.

Example 3.2: Consider the observation $x \sim poi(\lambda)$ and a conjugate prior distribution $\lambda \sim G(\alpha, \beta)$, i.e. $p(\lambda) = \frac{1}{C}\lambda^{\alpha-1}e^{-\beta\lambda}$ with $C = \int_0^\infty \lambda^{\alpha-1}e^{-\beta\lambda} d\lambda$. According to

Chapter 2, we have

$$\begin{aligned}\pi(\lambda) &\propto p(\lambda)\ell(x)\\ &\propto \lambda^{\alpha-1}e^{-\beta\lambda}\lambda^{x}e^{-\lambda}\\ &= \lambda^{\alpha+x-1}e^{-(\beta+1)\lambda}\end{aligned}$$

Hence $\pi(\lambda)$ is $G(\alpha_1, \beta_1)$ with $\alpha_1 = \alpha + x$ and $\beta_1 = \beta + 1$. Let us consider the normal approximation of $\pi(\lambda)$. The mode is

$$m = \frac{\alpha + x - 1}{\beta + 1}$$

$$\log \pi(\lambda) = C + (\alpha + x - 1)\log\lambda - (\beta + 1)\lambda$$

$$(\log \pi(\lambda))'' = -\frac{(\alpha + x - 1)}{\lambda^2}$$

$$(\log \pi(m))'' = -(\alpha + x - 1)\cdot\left(\frac{\beta + 1}{\alpha + x - 1}\right)^2 = -\frac{(\beta + 1)^2}{\alpha + x - 1}.$$

Hence the normal approximation is

$$\tilde{\pi}(\lambda) \propto \exp\left(-\frac{1}{2}\cdot\frac{(\beta + 1)^2}{\alpha + x - 1}(\lambda - m)^2\right).$$

So the approximate posterior variance is $\frac{\alpha+\alpha-1}{(\beta+1)^2}$.

3.2. Monte Carlo integration

Our aim is to compute the integral:

$$J = \int t(\theta)\pi(\theta)d\theta.$$

If we can sample a sequence of data $\theta_1, \cdots, \theta_n$ from π, then we have the approximation $\hat{J}$ for J as the following:

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} t(\theta_i).$$

By Law of Large Number (LLN), we know $\hat{J} \to J$ a.s.s. as $n \to \infty$.

Quite often, sampling from $\pi(\theta)$ is difficult, one considers first sampling from an auxiliary distribution $q(\theta)$ and compute

$$\hat{J} = \frac{1}{n} \sum_{i=1}^{n} t\left(\theta_i\right) \frac{\pi\left(\theta_i\right)}{q\left(\theta_i\right)}, \theta_1, \cdots, \theta_n \sim q(\theta),$$

because $\int t(\theta)\pi(\theta)d\theta = \int \left[t(\theta)\frac{\pi(\theta)}{q(\theta)}\right] q(\theta)d\theta$.

Another problem in Bayesian inference is that we only know

$$\pi(\theta) \propto \ell(\theta)p(\theta).$$

Denote $\pi^*(\theta) = \ell(\theta)p(\theta)$, then $\pi(\theta) = \frac{\ell(\theta)p(\theta)}{\int \pi^*(\theta)d\theta}$. and

$$J = \frac{\int t(\theta)\pi^*(\theta)d\theta}{\int \pi^*(\theta)d\theta}.$$

Rewrite it as $J = \frac{\int \frac{t(\theta)\pi^*(\theta)}{q(\theta)} q(\theta)d\theta}{\int \frac{\pi^*(\theta)}{q(\theta)} q(\theta)d\theta}$, it is natural to approximates J as the following: draw $\theta_1, \cdots, \theta_n$ from $q(\theta)$ and compute

$$\hat{J} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{t(\theta_i)\pi^*(\theta_i)}{q(\theta_i)}}{\frac{1}{n}\sum_{i=1}^{n} \frac{\pi^*(\theta_i)}{q(\theta_i)}}$$

3.3. Bayes' theorem via the rejection method

Recall the classical accept-reject method:

We aim to draw a random quantity Z from a distribution whose probability density is f, which may be difficult to be drown. We introduce an auxiliary probability distribution g and take the following assumption:

(A) Assume that there exists some $M > 0$ such that

$$\frac{f(x)}{g(x)} \leqslant M \quad \text{for all } x.$$

Pseudo code for accept-rejection method:

- Sample a random quantity Y from the distribution g
- Sample a random quantity U from $U([0,1])$.

  If $U \leqslant \frac{f(Y)}{Mg(Y)}$, accept, i.e. $X = Y$

  If $U > \frac{f(Y)}{Mg(Y)}$, reject and repeat the procedure.

In the Bayes' theorem setting, we recall that the posterior distribution $\pi(\theta)$ satisfies

$$\pi(\theta) \propto p(\theta)\ell(\theta).$$

where $p(\theta)$ is the prior distribution and $\ell(\theta)$ is the likelihood function. Denote

$$\pi^*(\theta) = p(\theta)\ell(\theta),$$

We aim to sample random quantities from $\pi(\theta)$ by the accept-reject method.

- We choose $p(\theta)$ as the auxiliary distribution, i.e. $g(\theta)$ above,
- We choose $M = \ell_{\max} = \max_\theta \ell(\theta)$.

The corresponding accept-reject sampling is as follows:

- (1) Sample a random quantity Y from the prior distribution $p(\theta)$,

- (2) Sample a random quantity U from the uniform distribution $U([0,1])$

  (2.1) If $U \leqslant \frac{\pi^*(Y)}{\ell_{\max} p(Y)} = \frac{\ell(Y)}{\ell_{\max}}$. Accept, i.e. $X = Y$

  (2.2) If $U > \frac{\ell(Y)}{\ell_{\max}}$. Reject, i.e. do not assign a quantify to X, return to (1).

- (3) Return X.

In the practice, it is often not easy to get $\ell_{\max}$.

3.4. Bayes' theorem via weighted resampling.

Resampling is a very important and popular method for drawing random quantity from a complicated probability distribution. The procedure for drawn a random quantity from the posterior distribution is as follows:

- (1) Draw n random quantities $\theta_1, \theta_2, \cdots, \theta_n$ from a distribution $q(\theta)$, which can be easily sampled, and create

$$\omega_i = \frac{\pi(\theta_i)/q(\theta_i)}{\sum_{j=1}^{n} \pi(\theta_j)/q(\theta_j)}.$$

- (2) For these sampled $\theta_1, \theta_2, \cdots, \theta_n$, resample them according to the probability $P(\theta = \theta_i) = \omega_i$ for $i = 1, 2, \cdots, n$. (It is sampled in the following way: We split the interval $[0,1]$ into the sub-intervals $I_1 = [0, \omega_1), I_2 = [\omega_1, \omega_1 + \omega_2), \cdots, I_k = [\omega_1 + \cdots + \omega_{k-1}, \omega_1 + \cdots + \omega_k), \cdots, I_n = [\omega_1 + \cdots \omega_{n-1}, 1]$. One randomly draws a quantity u from $[0,1]$, if $u \in I_i$, then we take $\theta = \theta_i$).

Theorem 3.1 The random quantity θ which is drown according to (1) and (2) has a distribution π. as $n \to \infty$.

Proof: For any $x \in \mathbb{R}$, consider $P(\theta \leqslant x)$. It can be decomposed into

$$\begin{aligned}
P(\theta \leqslant x) &= \sum_{i=1}^{n} P\left(\theta \leqslant x, \theta = \theta_i\right) \\
&= \sum_{i=1}^{n} P\left(\theta_i \leqslant x, \theta = \theta_i\right) \\
&= \sum_{i=1}^{n} \int_{-\infty}^{x} P\left(\theta = \theta_i \mid \theta_i = y\right) P_{\theta_i}(y) dy \\
&= \sum_{i=1}^{n} \int_{-\infty}^{x} P\left(\theta = y \mid \theta_i = y\right) q(y) dy \\
&\approx \sum_{i=1}^{n} \int_{-\infty}^{x} \frac{\frac{\pi(y)}{q(y)}}{\sum_{j=1}^{n} \frac{\pi(\theta_j)}{q(\theta_j)}} q(y) dy \\
&= \frac{n \int_{-\infty}^{x} \pi(y) dy}{\sum_{j=1}^{n} \frac{\pi(\theta_j)}{q(\theta_j)}} \\
&= \frac{\int_{-\infty}^{x} \pi(y) dy}{\frac{1}{n} \sum_{j=1}^{n} \frac{\pi(\theta_j)}{q(\theta_j)}} = \int_{-\infty}^{x} \pi(y) dy
\end{aligned}$$

As $n \to \infty$, by law of large number, we know

$$\frac{1}{n} \sum_{j=1}^{n} \frac{\pi\left(\theta_j\right)}{q\left(\theta_j\right)} = \int \frac{\pi(\theta)}{q(\theta)} \cdot q(\theta) d\theta = \int_{-\infty}^{+\infty} \pi(\theta) d\theta.$$

In the Bayesian inference, we let $q(\theta)$ be the prior distribution $p(\theta)$. The procedure of sampling random quantity with distribution can be done as the following:

- (1) Sampling a sequence of i.i.d. random quantities $\theta_1, \cdots, \theta_n$, according to the probability $p(\theta)$ and define

$$\omega_i = \frac{\pi^*\left(\theta_i\right) / p\left(\theta_i\right)}{\sum_{j=1}^{n} \frac{\pi^*(\theta_j)}{p(\theta_j)}}, i = 1, 2, \cdots, n.$$

- (2) Resample $\theta_1, \cdots, \theta_n$ obtained from (1) according to the probability $P\left(\theta = \theta_i\right) = \omega_i$ for $i = 1, 2, \cdots, n$.

By the same method, we can show that

$$P(\theta \leqslant a) = \frac{\int_{-\infty}^{a} \pi^*(\theta)d\theta}{\int_{-\infty}^{+\infty} \pi^*(\theta)d\theta} = \int_{-\infty}^{a} \pi(\theta)d\theta.$$

Example 3.3. A total of animals are categories into three types with $y = (y_1, y_2, y_3)$, the three types are assigned probability $\left(\frac{2+\theta}{4}, \frac{2-2\theta}{4}, \frac{\theta}{4}\right)$ with $\theta \in [0, 1]$. Given a θ, the conditional probability is

$$l(\theta) \quad p(y \mid \theta) \propto (2+\theta)^{y_1}(2-2\theta)^{y_2}\theta^{y_3} \propto (2+\theta)^{y_1}(1-\theta)^{y_2}\theta^{y_3}$$

Suppose that the observed data are $y = (125, 38, 34)$, take $\theta \sim U([0, 1])$, then the posterior distribution is

$$\pi(\theta) \propto (2+\theta)^{125}(1-\theta)^{38}\theta^{34}$$

However, we do not know the number $C = \int_0^1 (2+\theta)^{125}(1-\theta)^{38}\theta^{34}d\theta$. We shall use resampling algorithm, take $q(\theta) = p(\theta) \sim U([0, 1])$.

- Sample n (a large number) $\theta_1, \theta_2, \cdots, \theta_n$ from $U([0, 1])$ and compute the values

$$\omega_i = \frac{\ell(\theta_i)}{\sum_{j=1}^{n} \ell(\theta_j)} = \frac{(2+\theta_i)^{125}(1-\theta_i)^{38}\theta_i^{34}}{\sum_{j=1}^{n}(2+\theta_j)^{125}(1-\theta_j)^{38}\theta_j^{34}}$$

- Taking the obtained $\{\theta_1, \cdots, \theta_n\}$ as the sample space and resample them according to the probability $\{\omega_i\}_{1 \leqslant i \leqslant n}$, i.e.

$$P(\theta = \theta_i) = \omega_i.$$

Assignment: Make a python program to realize the resampling in Example 3.3, choose $n = 10000$.

Chapter 4. Markov chains

In this chapter, we consider a special stochastic process, Markov chain, denoted by $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ with $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$.

4.1. Definition and transition probabilities

A Markov chain $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ is a stochastic process such that given the present state, the past and future states are independent, more precisely,

$$\begin{aligned} & P\left(\theta^{(n+1)} \in A \mid \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \cdots, \theta^{(0)} \in A_0\right) \\ =& P\left(\theta^{(n+1)} \in A \mid \theta^{(n)} = x\right) \end{aligned}$$

where $A_0, A_1, \cdots, A_{n-1} \subset S$ and $x \in S$, S is the state space. Generally, the conditioned probability $P\left(\theta^{(n+1)} \in A \mid \theta^{(n)} = x\right)$ depends on x, A and n. If it does not depend on n, then the Markov chain is time homogeneous.

In this lecture, without specific statement, we only consider time homogeneous Markov chain. In this case, $P\left(\theta^{(1)} \in A \mid \theta^{(0)} = x\right) = P\left(\theta^{(n+1)} \in A \mid \theta^{(n)} = x\right)$, we denote $P(x, A) = P\left(\theta^{(1)} \in A \mid \theta^{(0)} = x\right)$ for $x \in S$ and $A \subset S$. We call $P(x, A)$ a transition probability:

- For $x \in S, P(x, \cdot)$ is a probability measure over S.
- For $A \subset S, P(\cdot, A)$ is a function on S.

If S is a discrete space, we write

$$P(x, y) = P(x, \{y\}) \text{ for } x, y \in S.$$

It is easy to see that

$$\begin{aligned} P(x,y) &\geqslant 0 \quad \text{for all } x,y \in S. \\ \sum_{y\in S} P(x,y) &= 1 \quad \text{for all } x \in S \end{aligned}$$

Example 4.1: (Random walk on $\mathbb{Z}$) Consider a particle starting from the origin and moving on $\mathbb{Z}$, more precisely, the positions of the particle is as follows:

$$\begin{aligned} \theta^{(0)} &= 0, \\ \theta^{(n)} &= \theta^{(n-1)} + \omega_n, n \geqslant 1, \end{aligned}$$

where $\omega_1, \omega_2, \cdots$, are i.i.d. integer valued random variables. Denote by f the probability function of ω_1.

If $f(1) = p, f(-1) = q, f(0) = r$ with $p+q+r = 1$, then the transition probability are given by

$$P(x,y) = \begin{cases} p, & y = x+1 \\ q, & y = x-1 \\ r, & y = x \\ 0, & \text{otherwise} \end{cases}$$

In the case of discrete space $S = \{x_1, x_2, \ldots\}$, we can put all transition probabilities $P(x_i, x_j)$ together to define transition probability matrix

$$\{P(x_i, x_j)\}_{x_i \in S, x_i \in S}$$

For instance, the transition probability matrix of above random walk is

$$\begin{bmatrix} \ddots & \ddots & \ddots & & & & \\ & q & r & p & & & \\ & & q & r & p & & \\ & & & q & r & p & \\ & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

Chapman - Kolmogorov relation

Denote $P^m(x,y) = P\left(\theta^{(m)} = y \mid \theta^{(0)} = x\right)$. When $m = 2$, we see

$$\begin{aligned} P^2(x,y) &= P\left(\theta^{(2)} = y \mid \theta^{(0)} = x\right) \\ &= \sum_{x_1} P\left(\theta^{(2)} = y, \theta^{(1)} = x_1 \mid \theta^{(1)} = x\right) \\ &= \sum_{x_1} \frac{P\left(\theta^{(2)} = y, \theta^{(1)} = x_1, \theta^{(0)} = x\right)}{P\left(\theta^{(0)} = x\right)} \\ &= \sum_{x_1} \frac{P\left(\theta^{(2)} = y \mid \theta^{(1)} = x_1, \theta^{(1)} = x\right) P\left(\theta^{(1)} = x_1, \theta^{(1)} = x\right)}{P\left(\theta^{(0)} = x\right)} \\ \text{Markov Property} \quad &= \sum_{x_1} P\left(\theta^{(2)} = y \mid \theta^{(1)} = x_1\right) P\left(\theta^{(1)} = x_1 \mid \theta^{(0)} = x\right) \\ &= \sum_{x_1} P\left(x_1, y\right) P\left(x, x_1\right) \end{aligned}$$

Hence. $P^2 = P \cdot P$.

Assignment: Using a similar argument to show that $P^{m+1} = P^m P$. and $P^m = \underbrace{P \ldots P}_{m}$ by an induction. Moreover, $P^{m+n} = P^m P^n$ for all $m \in \mathbb{N}$ and $n \in \mathbb{N}$.

Remark: It is easy to see that $P^{m+n} = P^m P^n$ can be presented as

$$P^{m+n}(x,y) = \sum_{z \in S} P^m(x,z) P^n(z,y), \quad \forall x, y \in S.$$

For all $n \in \mathbb{N}_0$, denote by $\pi^{(n)}$ the distribution of $\theta^{(n)}$. For $y \in S$,

$$\begin{aligned}\pi^{(n)}(y) &= P\left(\theta^{(n)} = y\right) \\ &= \sum_{x \in S} P\left(\theta^{(n)} = y, \theta^{(0)} = x\right) \\ &= \sum_{x \in S} P\left(\theta^{(n)} = y \mid \theta^{(0)} = x\right) P\left(\theta^{(0)} = x\right) \\ &= \sum_{x \in S} P^n(x, y)\pi^{(0)}(x)\end{aligned}$$

for all $x, y \in S$. The above relation can be written as

$$\pi^{(n)} = \pi^{(0)} P^n. \quad (*)$$

When $S = \{x_1, \cdots, x_r\}$ is finite, $\pi^{(n)} = \left(\pi^{(n)}(x_1), \cdots, \pi^{(n)}(x_r)\right)$, then (*) can be presented as

$$\left(\pi^{(n)}(x_1), \cdots, \pi^{(n)}(x_r)\right) = \left(\pi^{(0)}(x_1), \cdots, \pi^{(0)}(x_r)\right) \begin{bmatrix} P^n(x_1, x_1) & \cdots & P^n(x_1, x_r) \\ P^n(x_2, x_1) & \cdots & P^n(x_2, x_r) \\ \cdots & \cdots & \cdots \\ P^n(x_r, x_1) & \cdots & P^n(x_r, x_r) \end{bmatrix}$$

Example 4.2. Consider a Markov chain $\left\{\theta^{(n)}\right\}_{n \in \mathbb{N}_0}$ on the state space $S = \{0, 1\}$ with initial distribution $\pi^{(0)} = \left(\pi^{(0)}(0), \pi^{(0)}(1)\right)$ and transition probability matrix $P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$. Let us compute P^n, the n-step transition probability matrix. To this end, we compute the eigenvalue of P.

$$\text{Consider } \det(\lambda I - P) = \det \begin{bmatrix} \lambda - 1 + p & -p \\ -q & \lambda - 1 + q \end{bmatrix} = (\lambda - 1 + p)(\lambda - 1 + q) - pq$$

$$= \lambda^2 - (2 - p - q)\lambda + (1 - p)(1 - q) - pq = \lambda^2 - (2 - p - q)\lambda + 1 - p - q.$$

Let $\det(\lambda I - P) = 0$, we obtain

$$\lambda_1 = 1 \quad \lambda_2 = 1 - p - q$$

It is easy to see that

$$P \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} = \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1-p-q \end{bmatrix}$$

$$\Rightarrow \quad P = \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1-p-q \end{bmatrix} \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix}^{-1}$$

$$P^n = \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-p-q)^n \end{bmatrix} \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-p-q)^n \end{bmatrix} \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{-1}{p+q} & \frac{1}{p+q} \end{bmatrix}$$

If $0 < p + q < 2$. As $n \to \infty$,

$$P^n \to \begin{bmatrix} 1 & -p \\ 1 & q \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{-1}{p+q} & \frac{1}{p+q} \end{bmatrix} = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}.$$

Given $\pi^{(0)} = \left(\pi^{(0)}(0), \pi^{(0)}(1)\right)$, we see that

$$\pi^{(n)} = \pi^{(0)} P^n$$

As $n \to \infty, \pi^{(n)} = \pi^{(0)} \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix} = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$.

If a Markov chain starts from $x \in S$, i.e. $\theta^{(0)} = x$, We denote

$$P_x(\cdot) = P\left(\cdot \mid \theta^{(0)} = x\right).$$

For instance, let $B \subset S$, we have

$$P_x\left(\theta^{(n)} \in B\right) = P\left(\theta^{(n)} \in B \mid \theta^{(0)} = x\right).$$

Definition (First hitting time) Let $A \subset S$, define

$$T_A = \min\left\{n \geqslant 1 : \theta^{(n)} \in A\right\}$$

T_A is the moment that the Markov chain $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ first visits the set A. It is called the first hitting time of A . T_A is a random variable. If the Markov chain can never hit A, we define $T_A = \infty$.

Example 4.2 (continued): Let $\theta^{(0)} = 0$, we now compute $P_0\left(T_0 = n\right)$.

Observe

$$\begin{aligned}
P_0\left(T_0 = n\right) &= P_0\left(\theta^{(n)} = 0, \theta^{(n-1)} = 1, \theta^{(n-2)} = 1, \cdots, \theta^{(1)} = 1\right) \\
&= P(0,1)\underbrace{p(1,1)\cdots p(1,1)}_{n-2}p(1,0) \\
&= p(1-q)^{n-2}\cdot q \quad P_0\left(T_0 = 1\right) = P(0,0) = 1-p. \\
P_0\left(T_0 < \infty\right) &= \sum_{n=1}^{\infty} P\left(T_0 = n\right) \\
&= (1-p) + \sum_{n=2}^{\infty} p(1-q)^{n-2}q = 1 - p + pq\frac{1}{q} = 1
\end{aligned}$$

This means that the Markov chain starting 0 and will come back to 0 in finite time with probability 1.

4.2. Decomposition of state space

Let $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ be a M.C. on the state space S. Given $y \in S$, recall the hitting time T_y of y is defined by

$$T_y = \inf\left\{n \geqslant 1 : \theta^{(n)} = y\right\}$$

Let us define two quantities:

(1) $\rho_{xy} = P_x\left(T_y < \infty\right)$, it is the probability that the Markov chain starting from x hits the state y in finite time.

(2) The number of the visits of the M.C. to y, ie.

$$N(y) = \#\left\{n \geqslant 1 : \theta^{(n)} = y\right\} = \sum_{k=1}^{\infty} 1\left(\theta^{(k)} = y\right)$$

There are two useful relations:

$$\mathbb{E}\left[T_y \mid \theta^{(0)} = x\right] = \sum_{n=1}^{\infty} P_x\left(T_y > n\right),$$

$$\mathbb{E}\left[N(y) \mid \theta^{(0)} = x\right] = \sum_{n=1}^{\infty} P^n(x, y).$$

Let us show the second relation. $\mathbb{E}\left[N(y) \mid \theta^{(0)} = x\right] = \mathbb{E}\left[\sum_{n=1}^{\infty} 1\left(\theta^{(n)} = y\right) \mid \theta^{(0)} = x\right]$

$$= \sum_{n=1}^{\infty} \mathbb{E}\left[1\left(\theta^{(n)} = y\right) \mid \theta^{(n)} = x\right] = \sum_{n=1}^{\infty} P\left(\theta^{(n)} = y \mid \theta^{(0)} = x\right) = \sum_{n=1}^{\infty} P^n(x, y).$$

A state $y \in S$ is said to be **recurrent** if

$$\rho_{yy} = P_y\left(T_y < \infty\right) = 1.$$

This means that the M.C. starting from y will return to y in finite time with probability 1.

A state $y \in S$ is said to be **transient** if

$$\rho_{yy} = P_y\left(T_y < \infty\right) < 1.$$

This means that $P_y\left(T_y = \infty\right) > 0$, ie. the Markov chain starting from y will never return to y in finite time with positive probability.

A recurrent state y can be further classified as

- positive recurrent $\mathbb{E}\left[T_y \mid \theta^{(0)} = y\right] < \infty \quad \& \quad \rho_{yy} = 1.$
- null recurrent: $\mathbb{E}\left[T_y \mid \theta^{(0)} = y\right] = \infty \quad \& \quad \rho_{yy} = 1.$

For a recurrent state $y \in S$, the Markov chain will revisit it infinitely many times, more precisely, $P_y(N(y) = \infty) = 1$.

For two states $x, y \in S$, if $\rho_{xy} = P_x\left(T_y < \infty\right) > 0$, we denote $x \to y$ (' x conmuminates to y). $C \subset S$ said to be **irreducible** if $x \to y$ for every pair $x, y \in C$. A M.C. is irreducible if S is irreducible. The set $C \subseteq S$ is said to be closed if $\rho_{xy} = 0$ for all $x \in C$ and $y \notin C$.

An important claim: $P_{xy} > 0 \Leftrightarrow$ There exists some $n \in \mathbb{N}$ s.t. $P^n(x, y) > 0$.

Proof. " $\Rightarrow$ since $\rho_{xy} = P_x\left(T_y < \infty\right) = \sum_{n=1}^{\infty} P_x\left(T_y = n\right) > 0$ There must be some $n_0 \in \mathbb{N}$ so that $P_x\left(T_y = n_0\right) > 0$, i.e.

$$P_x\left(\theta^{(1)} \neq y, \cdots \theta^{(n_0-1)} \neq y, \theta^{(n_0)} = y\right) > 0,$$

this immediately implies $P_x\left(\theta^{(n_0)} = y\right) > 0$, i.e. $P^{n_0}(x, y) > 0$.

" $\Leftarrow$ " Observe $\left\{\theta^{(n)} = y\right\} \subset \left\{T_y < \infty\right\}$ for some $n \in \mathbb{N}$.

$$P_x\left(T_y < \infty\right) > P_x\left(\theta^{(n)} = y\right) = P^n(x, y) > 0.$$

□

Example 4.3 (a) $S = \{0, 1, 2\}$, the (one-step) transition probability matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

The Markov chain is irreducible. To this end, from the form of P, we only need to show there exist some $n \in \mathbb{N}, m \in \mathbb{N}$ so that $P^n(0,2) > 0$ and $P^m(2,0) > 0$. Obvious

$$\begin{aligned} P^2(0,2) &= P\left(\theta^{(2)} = 2 \mid \theta^{(0)} = 0\right) \\ &\geqslant P\left(\theta^{(2)} = 2, \theta^{(1)} = 1 \mid \theta^{(0)} = 0\right) \\ &= \frac{P\left(\theta^{(2)} = 2 \mid \theta^{(1)} = 1, \theta^{(0)} = 0\right) P\left(\theta^{(1)} = 1, \theta^{(0)} = 0\right)}{P\left(\theta^{(0)} = 0\right)} \\ &= P\left(\theta^{(2)} = 2 \mid \theta^{(1)} = 1\right) P\left(\theta^{(1)} = 1 | \theta^{(0)} = 0\right) \\ &= P(1,2)P(0,1) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Similarly, $P^2(2,0) \geqslant \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$.

Assignment: Show $P^2(2,0) \geqslant \frac{1}{12}$.

4.3. Stationary distributions

A fundamental problem for M.C. in the context of simulation is the study of its stationary distributions (or invariant measures). Let $\left\{\theta^{(n)}\right\}_{n \in \mathbb{N}_0}$ be a time homogeneous M.C. on the state space S, with transition probability matrix $\{P(x,y)\}_{x \in S, y \in S}$. A distribution $\pi = \{\pi(x)\}_{x \in S}$ is stationary if $\pi(y) = \sum_{x \in S} \pi(x) P(x,y), \forall y \in S$. This relation can be represented as

$$\pi = \pi P.$$

If S has finite elements, e.g. $S = \{x_1, \cdots, x_m\}$, then

$$[\pi(x_1) \cdots \pi(x_m)] = [\pi(x_1) \cdots \pi(x_m)] \begin{bmatrix} P(x_1,x_1) & \cdots & P(x_1,x_m) \\ P(x_2,x_1) & \cdots & P(x_2,x_m) \\ \cdots & \cdots & \cdots \\ P(x_m,x_1) & \cdots & P(x_m,x_m) \end{bmatrix}.$$

Example 4.2 (continued): $S = \{0, 1\}$. The transition probability is $\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$ $\pi = \begin{bmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{bmatrix}$ is a stationary probability.

Example 4.3 (A prototype of Gibbs sampler) Let the state space be $S = \{0, 1\}^2$, i.e. $S = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. For $\theta \in S, \theta = (\theta_1, \theta_2)$ with $\theta_i = 0$ or 1 for $i = 1, 2$. For a prob. distribution π on S, we represented as

	0	1
0	π_{00}	π_{01}
1	π_{10}	π_{11}

$\pi(i, j) = p\left(\theta_1 = i, \theta_2 = j\right), i \in \{0, 1\}$ and $j \in \{0, 1\}$.

- For the first component θ_1, given $\theta_2 = j$, the conditional probability θ_1 is

$$\pi_1(0 \mid j) = \frac{\pi_{0j}}{\pi_{+j}} \quad \pi_1(1 \mid j) = \frac{\pi_{1j}}{\pi_{+j}}$$

  where $\pi_{+j} = \pi_{0j} + \pi_{1j}$.

- For the second component θ_2, given $\theta_1 = i$, the conditional probability θ_2 is

$$\pi_2(0 \mid i) = \frac{\pi_{i0}}{\pi_{i+}} \quad \pi_2(1 \mid i) = \frac{\pi_{i1}}{\pi_{i+}}$$

  where $\pi_{i+} = \pi_{i0} + \pi_{i1}$.

- Design a transition probability

$$P((i, j), (k, l)) = P\left(\theta^{(n)} = (k, l) \mid \theta^{(n-1)} = (i, j)\right) = \frac{\pi_{kl}}{\pi_{k+}} \cdot \frac{\pi_{kj}}{\pi_{+j}}$$

$\{P((i,j),(k,l))\}_{i,j,k,l\in\{0,1\}}$ is a 4×4 matrix.

$$P=[P_1\ P_2\ P_3\ P_4]=\begin{bmatrix} \frac{\pi_{00}\pi_{00}}{\pi_{0+}\pi_{+0}} & \times & \times & \times \\ \frac{\pi_{00}\pi_{01}}{\pi_{0+}\pi_{+1}} & \times & \times & \times \\ \frac{\pi_{00}\pi_{00}}{\pi_{0+}\pi_{+0}} & \times & \times & \times \\ \frac{\pi_{00}\pi_{01}}{\pi_{0+}\pi_{+1}} & \times & \times & \times \end{bmatrix}$$

we have $\pi P=\pi$ with $[\pi_{00}\ \pi_{01}\ \pi_{10}\ \pi_{11}]$.

Let us here only verify that $\pi P_1=\pi_{00}$

$$\begin{bmatrix} \pi_{00} & \pi_{01} & \pi_{10} & \pi_{11} \end{bmatrix}\begin{bmatrix} \frac{\pi_{00}\pi_{00}}{\pi_{0+}\pi_{+0}} \\ \frac{\pi_{00}\pi_{01}}{\pi_{0+}+\pi_{+1}} \\ \frac{\pi_{00}\pi_{00}}{\pi_{0+}\pi_{+0}} \\ \frac{\pi_{00}\pi_{01}}{\pi_{0+}\pi_{+1}} \end{bmatrix}=\frac{\pi_{00}^3+\pi_{10}\pi_{00}^2}{\pi_{0+}\pi_{+0}}+\frac{\pi_{00}\pi_{01}^2+\pi_{00}\pi_{01}\pi_{11}^2}{\pi_0+\pi_{+1}}$$

$$=\frac{\pi_{00}^2}{\pi_{0+}}+\frac{\pi_{00}\pi_{01}}{\pi_{0+}}=\pi_{00}$$

Example 4.4. Consider

$$P=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Solving $\pi P=\pi$, we obtain $\pi(0)=\pi(1)=\frac{1}{2}$. On the other hand

$$P\cdot P=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\quad PPP=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$P^{2m}=\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\quad P^{2m+1}=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\quad \text{for all } m\in\mathbb{N}.$$

4.4. Reversible chains

Let $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ be a time homogeneous M.C. with transition probability $P(x,y)$ and stationary distribution π. Let us consider the M.C. from 0 to n in the reverse order, i.e. $\theta^{(n)}, \theta^{(n-1)}, \cdots, \theta^{(0)}$.

$$
\begin{aligned}
&P\left(\theta^{(k)}=y \mid \theta^{(k+1)}=x, \theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k}\right) \\
=&\frac{P\left(\theta^{(k)}=y, \theta^{(k+1)}=x, \theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k}\right)}{P\left(\theta^{(k+1)}=x, \theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k}\right)} \\
=&\frac{P\left(\theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k} \mid \theta^{(k+1)}=x, \theta^{(k)}=y\right) P\left(\theta^{(k+1)}=x, \theta^{(k)}=y\right)}{P\left(\theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k} \mid \theta^{(k+1)}=x\right) P\left(\theta^{(k+1)}=x\right)} \\
=&\frac{P\left(\theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k} \mid \theta^{(k+1)}=x\right) P\left(\theta^{(k+1)}=x \mid \theta^{(k)}=y\right) P\left(\theta^{(k)}=y\right)}{P\left(\theta^{(k+2)}=x_2, \cdots, \theta^{(n)}=x_{n-k} \mid \theta^{(k+1)}=x\right) P\left(\theta^{(k+1)}=x\right)} \\
=&\frac{P\left(\theta^{(k+1)}=x \mid \theta^{(k)}=y\right) P\left(\theta^{(k)}=y\right)}{P\left(\theta^{(k+1)}=x\right)} \\
=&P\left(\theta^{(k)}=y \mid \theta^{(k+1)}=x\right)=\frac{P(y,x)\pi^{(k)}(y)}{\pi^{(k+1)}(x)}.
\end{aligned}
$$

Hence, $\theta^{(n)}, \theta^{(n-1)}, \cdots, \theta^{(1)}, \theta^{(0)}$ is also a M.C. If $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ is a time homogeneous M.C. with stationary distribution π, then the reverse M.C. is also time homogeneous with transition probability

$$P^*(x,y)=\frac{\pi(y)P(y,x)}{\pi(x)}$$

If $P^*(x,y)=p(x,y)$ for all $x\in S$ and $y\in S$, then we have

$$(*) \qquad \pi(x)p(x,y)=\pi(y)p(y,x). \qquad \text{for all } x,y\in S$$

We call $(*)$ as the detailed balance condition. The corresponding M.C. is called time reversible M.C.

Example 4.5 Metropolis algorithm (1953. Metropolis)

Consider a given distribution $\{p_x : x \in S\}$ with $p_x \geqslant 0$ for all $x \in S$ and $\sum_{x\in S} p_x = 1$. Let $\left\{\theta^{(n)}\right\}$ be an irreducible M.C. with transition probability matrix $\{Q(x,y) : x \in S, y \in S\}$ such that $Q(x,y) = Q(y,x)$ for all $x \in S, y \in S$. Now let us modify the M.C. $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ by injecting a censorship:

(1) Given $\theta^{(n)} = x$, the M.C. transits from x to y according to $Q(x,y)$, i.e. $\theta^{(n+1)} = y$ with probability $Q(x,y)$.

(2) The proposed value for $\theta^{(n+1)}$ is accepted with probability $\min\{1, p_y/p_x\}$, and reject otherwise, leaving the chain staging at the state x.

Combining the two steps (1) & (2), we can figure out a transition probability $P(x,y)$.

- for $y \neq x$,
$$\begin{aligned} P(x,y) &= P\left(\theta^{(n+1)} = y,\ \text{accept } y \mid \theta^{(n)} = x\right) \\ &= P\left(\theta^{(n+1)} = y \mid \theta^{(n)} = x\right) P(\text{accept } y) \\ &= Q(x,y)\min\left\{1, \frac{p_y}{p_x}\right\}. \end{aligned}$$
the censorship is independent of $\theta^{(n)} = x$.

- for $y = x$
$$\begin{aligned} P(x,x) &= P\left(\theta^{(n+1)} = x, \text{accept } x \mid \theta^{(n)} = x\right) \\ &\quad + P\left(\theta^{(n+1)} \neq x, \text{reject } y \mid \theta^{(n)} = x\right) \\ &= P\left(\theta^{(n+1)} = x \mid \theta^{(n)} = x\right) P\left(\text{accept } x\right) + \sum_{y\neq x} P\left(\theta^{(n+1)} = y,\ \text{reject } y \mid \theta^{(n)} = x\right) \\ &= Q(x,x) + \sum_{y\neq x} P\left(\theta^{(n+1)} = y \mid \theta^{(n)} = x\right) P(\text{reject } y) \\ &= Q(x,x) + \sum_{y\neq x} Q(x,y)\left(1 - \min\left\{1, \frac{p_y}{p_x}\right\}\right) \end{aligned}$$

We can define a new Markov chain $\left\{\hat{\theta}^{(n)}\right\}_{n\in\mathbb{N}_0}$ who transition prob. is $P(x,y)$. We claim that $\left\{\theta^{(n)}\right\}_{n\in\mathbb{N}_0}$ is an reversible M.C with stationary distribution $\{p_x\}$. To show this, we only need to prove that

$$p_x P(x,y) = p_y P(y,x) \quad \forall x \in S, y \in S.$$

Indeed, for $x \neq y$, suppose $p_y \geqslant p_x$, then

$$P(x,y) = Q(x,y)$$

and thus

$$\begin{aligned} p_x p(x,y) = p_x Q(x,y) &= Q(y,x) \min\left\{1, \frac{p_x}{p_y}\right\} p_y \\ &= p(y,x) p_y. \end{aligned}$$

As $p_x \geqslant p_y$, by a similar argument, we have

$$p_x P(x,y) = p_y P(y,x).$$

A remark about Markov Chain on continuous state space S, e.g. $S = \mathbb{R}$ or $S = \mathbb{R}^d$. We shall replace the prob. mass function $\{\pi(x) : x \in S\}$ with $\sum_{x\in S} \pi(x) = 1$ by the prob. density function $\pi(x)$ with $\int_S \pi(x)dx = 1$. The transition prob. matrix $\{P(x,y)\}_{x\in S, y\in S}$ with $\sum_{y\in S} P(x,y) = 1$ with the transition prob. kernel $p(x,y)$ with $\int_S p(x,y)dy = 1$. The stationary measure π has the property $\pi(y) = \int \pi(x)p(x,y)dx$. The detailed balance condition is

$$\pi(x)p(x,y) = \pi(y)p(y,x) \quad \forall x, y \in S.$$

Chapter 5 Gibbs Sampling

Gibbs sampling is a powerful tool for sampling posterior distribution and image processing. It was first proposed by German & German (1984) in image processing,

and later further developed by Gelfand & Smith (1990), Robert (2001).

5.1 Definition and properties

Assume that the distribution of interest is $\pi(\theta)$ with $\theta = (\theta_1, \cdots, \theta_d)'$, each θ_i can be a scalar, a vector or a matrix, e.g. θ_i is the color of the i-th pixel of an image. We denote $\pi_i(\theta_i) = \pi(\theta_i \mid \theta_{-i})$ where $\theta_{-i} = (\theta_1, \cdots, \theta_{i-1}, \theta_{i+1}, \cdots \theta_d)$ for $i = 1, 2, \cdots, d$.

There are a lot of situations in which the distribution π is hard to be sampled, while the distribution π_i is easy. Gibbs sampling provides a way to sample distribution π through successive generating π_i. It can be described in the following way:

- Set an initial value $\theta^{(0)} = \left(\theta_1^{(0)}, \cdots, \theta_d^{(0)}\right)$
- For $j \geqslant 1$, update $\theta^{(j-1)} = \left(\theta_1^{(j-1)}, \cdots, \theta_d^{(j-1)}\right)$ through the following successive generation of Values:

$$
\begin{aligned}
&\theta_1^{(j)} \sim \pi\left(\theta_1 \mid \theta_2^{(j-1)} \cdots \theta_d^{(j-1)}\right) \\
&\theta_2^{(j)} \sim \pi\left(\theta_2 \mid \theta_1^{(j)}, \theta_3^{(j-1)} \cdots \theta_d^{(j-1)}\right) \\
&\theta_3^{(j)} \sim \pi\left(\theta_3 \mid \theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)} \cdots \theta_d^{(j-1)}\right) \\
&\cdots \\
&\theta_d^{(j)} \sim \pi\left(\theta_d \mid \theta_1^{(j)} \cdots \theta_{d-1}^{(j)}\right)
\end{aligned}
$$

Denote $\theta^{(j)} = \left(\theta_1^{(j)} \cdots \theta_d^{(j)}\right)$ for $j \geqslant 0$, it is a M.C. One needs to show the convergence of this M.C. $\left\{\theta^{(i)}\right\}_{j \in \mathbb{N}_0}$.

Example 5.1: (An example in chapter 4 revisited) Example 4.3 A prototype of Gibbs sampler: Let the state space be $S = \{0,1\}^2$, ie. $S = \{(0,0),(0,1),(1,0),(1,1)\}$ For $\theta \in S, \theta = (\theta_1, \theta_2)$ with $\theta_i = 0$ or 1 for $i = 1, 2$. For a prob. distribution π on S, we represented as

	0	1
0	π_{00}	π_{01}
1	π_{10}	π_{11}

$$\pi(i,j) = p\left(\theta_1 = i, \theta_2 = j\right), \; i \in \{0,1\} \text{ and } j \in \{0,1\}.$$

- For the first component θ_1, given $\theta_2 = j$, the conditional probability θ_1 is

$$\pi_1(0 \mid j) = \frac{\pi_{0j}}{\pi_{+j}} \quad \pi_1(1 \mid j) = \frac{\pi_{1j}}{\pi_{+j}}$$

  where $\pi_{+j} = \pi_{0j} + \pi_{1j}$.

- For the second component θ_2, given $\theta_1 = i$, the conditional probability θ_2 is

$$\pi_2(0 \mid i) = \frac{\pi_{i0}}{\pi_{i+}} \quad \pi_2(1 \mid i) = \frac{\pi_{i1}}{\pi_{i+}}$$

  where $\pi_{i+} = \pi_{i0} + \pi_{i1}$.

- Design a transition probability

$$\begin{aligned} P((i,j),(k,l)) &= P\left(\theta^{(n)} = (k,l) \mid \theta^{(n-1)} = (i,j)\right) \\ &= \frac{\pi_{kl}}{\pi_{k+}} \cdot \frac{\pi_{kj}}{\pi_{+j}} \end{aligned}$$

$$\theta^{(n-1)} = (i,j) : \text{First fix } \theta_2^{(n-1)} = j \to \theta_1^{(n)} = k \mid \theta_2^{(n-1)} = j$$

$$\text{Fix } \theta_1^{(n)} = k \to \theta_2^{(n)} = l \mid \theta_1^{(n)} = k$$

Example 5.2 (Carlin, Gelfand, Smith, 1992) Let $y_1, \cdots, y_n$ be a sample from a poisson distribution for which there is a suspicion of a change point m among

the observation process where the means change, $m = 1, \cdots n$. Given the change point m, the observation distributions are $y_i \mid \lambda \sim \text{Poi}(\lambda), i = 1, 2, \cdots, m$ and $y_i \mid \phi \sim \text{Poi}(\phi), i = m+1, \cdots, n$. The model is completed with independent prior distributions $\lambda \sim G(\alpha, \beta), \phi \sim G(\nu, \delta)$, and the change point m uniformly distributed over $\{1, 2, \cdots, n\}$, where α, β, ν and δ are known constants. The posterior density is

$$\begin{aligned}
&\pi(\lambda, \phi, m) \propto f\left(y_1, \cdots, y_n \mid \lambda, \phi, m\right) p(\lambda, \phi, m) \\
&= \left(\prod_{i=1}^{m} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda}\right) \left(\prod_{j=m+1}^{n} \frac{\phi^{y_j}}{y_j!} e^{-\phi}\right) \frac{1}{c_{\alpha,\beta}} \lambda^{\alpha-1} e^{-\beta\lambda} \frac{1}{C_{\nu,\delta}} \phi^{\nu-1} e^{-\delta\phi} \frac{1}{n} \\
&\propto \left(\prod_{i=1}^{m} \lambda^{y_i} e^{-\lambda}\right) \left(\prod_{j=m+1}^{n} \phi^{y_j} e^{-\phi}\right) \lambda^{\alpha-1} e^{-\beta\lambda} \phi^{\nu-1} e^{-\delta\phi} \\
&= \lambda^{\alpha+S_m-1} e^{-(\beta+m)\lambda} \phi^{\nu+S_n-S_m-1} e^{-(\delta+n-m)\phi},
\end{aligned}$$

where $S_k = y_1 + \cdots + y_k$. It is easy to see that given ϕ, m, λ satisfies a Gamma distribution $G\left(\alpha + S_m, \beta + m\right)$, i.e. $\pi(\lambda \mid \phi, m) = G\left(\alpha + S_m, \beta + m\right)$. Similarly, $\pi(\phi \mid \lambda, m) = G\left(\nu + S_n - S_m, \delta + n - m\right)$

$$\pi(m \mid \phi, \lambda) = \frac{\lambda^{\alpha+S_m-1} e^{-(\beta+m)\lambda} \phi^{\nu+S_n-S_m-1} e^{-(\delta+n-m)\phi}}{\sum_{l=1}^{n} \lambda^{\alpha+S_l-1} e^{-(\beta+l)\lambda} \phi^{\nu+S_n-S_l-1} e^{-(\delta+n-l)\phi}}$$

All these conditional distributions are easily sampled. So we can use Gibbs sampling to sample distribution $\pi(\lambda, \phi, m)$. The algorithm is as the following

1. Set the initial value $\left(\lambda^{(0)}, \phi^{(0)}, m^{(0)}\right)$

2. Obtain a new value $\left(\lambda^{(j)}, \phi^{(j)}, m^{(j)}\right)$ through the following successive generation of values

$$\begin{aligned}
&\lambda^{(j)} \sim G\left(\alpha + S_{m^{(j-1)}}, \beta + m^{(j-1)}\right) \\
&\phi^{(j)} \sim G\left(\nu + S_n - S_{m^{(j-1)}}, \delta + n - m^{(j-1)}\right) \\
&m \sim \pi\left(m \mid \lambda^{(j)}, \phi^{(j)}\right)
\end{aligned}$$

3. Change counter j to $j+1$ and return to step 2 until convergence is reached.

Chapter 6 Metropolis-Hastings algoithms

This algorithm stems from the papers by Metropolis et al. (1953) and Hastings (1970). The original paper by Metropolis et al. (1953) was published on Journal of Chemical Physics.

6.1 Definition and properties

Let us recall Metropolis algorithm (1953). Let $\{\pi(x)\}_{x\in S}$ be a prob. distri. which is hard to be sampled directly. Metropolis proposed a Markov chain $\left\{\tilde{\theta}_n\right\}_{n\in\mathbb{N}_0}$, which is reversible and has π as its limiting distribution. Metropolis' procedure is as follows:
(1) Introduce a Markov chain $\{\theta_n\}_{n\in\mathbb{N}_0}$, whose transition probability matrix is $\{Q(x,y)\}_{x\in S,y\in S}$ such that $\theta(x,y)=\theta(y,x), x\in S, y\in S$.
(2) Suppose that $\tilde{\theta}_n = x$, it jumps to $\tilde{\theta}_{n+1}$ according to the following rule:

- (2.1) x first transits to y with a prob. $Q(x,y)$.
- (2.2) this transition is accepted with a prob. $\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$, and rejected with a prob. $1-\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$, ie. the chain staying at x.

(3) Replace n by $n+1$, continuing step (2).

We have learnt that the transition probability $\tilde{Q}(x,y)$ for $\left\{\hat{\theta}_n\right\}_{n\in\mathbf{N}_0}$ is computed

as the following:

$$\begin{aligned}
\text{As } y \neq x, \tilde{Q}(x,y) &= P\left(\theta_{n+1} = y, y \text{ accepted } \mid \theta_n = x\right) \\
&= \frac{P\left(\theta_{n+1} = y, y \text{ accepted }, \theta_n = x\right)}{P\left(\theta_n = x\right)} \\
&= \frac{P\left(y \text{ accepted } \mid \theta_{n+1} = y, \theta_n = x\right) P\left(\theta_{n+1} = y, \theta_n = x\right)}{P\left(\theta_n = x\right)} \\
&= \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} \frac{P\left(\theta_{n+1} = y, \theta_n = x\right)}{P\left(\theta_n = x\right)} \\
&= \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} Q(x,y).
\end{aligned}$$

$$\begin{aligned}
\text{As } y = x, \tilde{Q}(x,x) &= Q(x,x) + \sum_{y \neq x} P\left(\theta_{n+1} = y, y \text{ rejected} \mid \theta_n = x\right) \\
&= Q(x,x) + \sum_{y \neq x} \frac{P\left(\theta_{n+1} = y, y \text{ rejected, } \theta_n = x\right)}{P\left(\theta_n = x\right)} \\
&= Q(x,x) + \sum_{y \neq x} \frac{P\left(y \text{ rejected } \mid \theta_{n+1} = y, \theta_n = x\right) P\left(\theta_{n+1} = y, \theta_n = x\right)}{P\left(\theta_n = x\right)} \\
&= Q(x,x) + \sum_{y \neq x} \left(1 - \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}\right) Q(x,y)
\end{aligned}$$

It is very easy to verify that the following detailed balance condition holds:

$$\pi(x)\tilde{Q}(x,y) = \pi(y)\tilde{Q}(y,x) \quad \forall x, y \in S.$$

Indeed, if $x = y$, the relation is trivially true.

$$\begin{aligned}
\text{if } x \neq y, \pi(x)Q(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\} &= \min\{\pi(x)Q(x,y), \pi(x)Q(x,y)\} \\
(*) = \min\{\pi(x)Q(x,y), \pi(y)Q(y,x)\} &= \pi(y)Q(y,x)\min\left\{1, \frac{\pi(x)Q(x,y)}{\pi(y)Q(y,x)}\right\} \\
(*) &= \pi(y)Q(y,x)\min\left\{1, \frac{\pi(x)}{\pi(y)}\right\}
\end{aligned}$$

where * used $Q(x,y) = Q(y,x)$.

Metropdis-Hastings algorithm is a generalization of Metropolis algorithm, first proposed by Hastings, which remove the symmetry condition $Q(x,y) = Q(y,x)$ and modify the accept rate $\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$ with the new one $\min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}$.

More precisely, the procedure of MH algorithm is

(1) The auxiliary Markov chain $\{\theta_n\}_{n\in\mathbb{N}}$ has a transition prob. $Q(x,y)$, and can be easily runned in practice. The condition $Q(x,y) = Q(y,x)$ it NOT necessary.

(2) Suppose $\tilde{\theta}_n = x$, it jumps to $\tilde{\theta}_{n+1}$ according to the following rule:

- (2.1) it first transits to a state y according to the prob. $Q(x,y)$.
- (2.2) this transition is accepted according to the accept rate $\min\left\{1, \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}\right\}$, and is rejected with $1 - \min\left\{1, \frac{\pi(y)\theta(y,x)}{\pi(x)\theta(x,y)}\right\}$ i.e. the chain staying at x.

(3) Replace n with $n+1$, continue the step 2.

Example 6.1. Consider the following mixture of Gaussian model

$$\pi(\theta, p) = pf_N\left(\theta; \mu_1, \Sigma_1\right) + (1-p)f_N\left(\theta; \mu_2, \Sigma_2\right)$$

where $p \in (0,1)$, $f_N\left(\theta; \mu_1, \Sigma_1\right) = \frac{1}{2\pi\sqrt{\det(\Sigma_1)}} \exp\left(-\frac{1}{2}\left(\theta-\mu_1\right)^\top \Sigma_1^{-1}\left(\theta-\mu_1\right)\right)$

$$f_N\left(\theta, \mu_2, \Sigma_2\right) = \frac{1}{2\pi\sqrt{\det\left(\Sigma_2\right)}} \exp\left(-\frac{1}{2}\left(\theta-\mu_2\right)^\top \Sigma_2^{-1}\left(\theta-\mu_2\right)\right)$$

$$\mu_1 = \begin{pmatrix} 4 \\ 5 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0.7 \\ 3.5 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1.0 & 0.7 \\ -0.7 & 1.0 \end{pmatrix},$$

p is usually an unknown number. For simplicity, we assume $p = 0.7$. Usually we can sample a random variable $\chi \sim \text{Ber}(p), X_1 \sim f_N\left(\theta; \mu_1, \Sigma_1\right), X_2 \sim f_N\left(\theta; \mu_2, \Sigma_2\right)$,

we take the combination

$$X = \chi X_1 + (1-\chi)X_2$$
$$X \sim p f_N\left(\theta; \mu_1, \Sigma_1\right) + (1-p) f_N\left(\theta; \mu_2, \Sigma_2\right).$$

We now use MH algorithm to sample $\pi(\theta)$. We choose the Markov chain whose transition prob. is $q(x,y) = f_N\left(y; x, \gamma I_2\right) = \frac{1}{2\pi\sqrt{\gamma}} \exp\left(-\frac{1}{2\gamma}|y-x|^2\right)$. The algorithm reads as

- 1. Initial $\theta^{(0)} \in \mathbb{R}^2$;
- 2. For $j \geqslant 1$, sample $\tilde{\theta}^{(j)}$ according to the transition prob. $q\left(\theta^{(j-1)}, \tilde{\theta}^{(j)}\right)$. Compute $r_j = \min\left\{1, \frac{\pi\left(\tilde{\theta}^{(j)}\right) q\left(\tilde{\theta}^{(j)}, \theta^{(j-1)}\right)}{\pi\left(\theta^{(j-1)}\right) q\left(\theta^{(j-1)}, \tilde{\theta}^{(j)}\right)}\right\} = \min\left\{1, \frac{\pi\left(\tilde{\theta}^{(j)}\right)}{\pi\left(\theta^{(j-1)}\right)}\right\}$. Sample a $U \sim U([0,1])$, if $U \leqslant r_j$, accept $\tilde{\theta}^{(j)}$, i.e. $\theta^{(j)} = \tilde{\theta}^{(j)}$, if $U > r_j$ reject $\tilde{\theta}^{(j)}$, i.e. $\theta^{(j)} = \theta^{(j-1)}$.
- 3. Replace j by $j+1$, repeat step 2.

The parameter γ plays an important role in the algorithm;

- (1) γ is too small, the chain has difficulty moving across the parameter space.
- (2) γ is too large, the acceptance rate is small, leading to a computationally inefficient sampling scheme.