Computational Methods in Stochastics

CS-E5795: Summarised lecture notes **DRAFT**

All content based on the lecture notes by Prof. Riku Linna (Aalto University).
Summarised and edited by Jieming You.

. . .

Last updated: November 7, 2023

❖ Random variables and distributions

1.1 Random variables

The expression $X \leq x$ is the event that random variable X assumes a value lesser or equal to the real number x. The probability of this event is $\Pr(X \leq x)$.

The cumulative probability distribution of random variable X is defined as

$$F(x) = \Pr(X \leq x), \quad -\infty < x < \infty$$

Similarly the probability of the event of $X > x$ is $\Pr(X > x) = 1 - F(x)$.

For continuous random variable X: $\Pr(X = x) = 0 \quad \forall x$.

Probability density function of random variable X is defined as a nonnegative function

$$\Pr(a < X \leq b) = \int_a^b f(x)\, dx \qquad \text{for} -\infty < a < b < \infty$$

Consequently, PDF is the derivative of the CDF.

1.1.1 Expectation of random variables

If X is a random bariable, $Y = g(X)$ is also a random varialble. The expectation ("expected value") of Y is

$$\mathbb{E}[g(x)] = \int g(x) f_X(x)\, dx$$

1.1.2 Joint distributions, independency, and conditionality

Given two random variables X and Y, their joint distribution is defined as

$$\begin{aligned} F_{XY}(x, y) &= F(x, y) \\ &= \Pr(X \leq x \text{ and } Y \leq y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(\xi, \eta)\, d\xi d\eta \quad \forall x, y \end{aligned}$$

For jointly distributed random variables, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Random variables X and Y are called independent if $F(x, y) = F_X(x)F_Y(y)$.

For all events A and B where $\Pr(B) > 0$, the conditional probability of A given B is defined by the Bayes Theorem:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \Pr(B) > 0$$

This relates to the law of total probability

$$\Pr(A) = \sum_{i}^{\infty} \Pr(A \cap B_i) = \sum_{i}^{\infty} \Pr(A|B_i) \Pr(B_i)$$

1.2 Central limit theorem

An intuitive definition of the central limit theorem: The independent observations of any random process with fixed mean and variance (homogenous) are normally distributed with the original mean. This is called an **additive process**.

Similarly, the lognormal distribution is the result of a **multiplicatively processes**. That is, the random process is described by a product (instead of a sum in the case of the normal distribution) of i.i.d. observations.

❖ Stochastic simulation

2.1 Monte Carlo

The expected value of a random variable can be calculated by integrating over the probability density functions $\mathbb{E}[g(x)] = \int g(x) f_X(x)\, dx$. In the case of more complicated distributions and functions, it's seldom trivial to calculate the integral analytically.

Monte Carlo integration is a form of stochastic simulation, where we approximate the integral of $g(X)$ by taking random samples from $g(x)$ as $g(x_i), ..., g(x_n)$. The law of large numbers assure that the integral can be numerically approximated by

$$\begin{aligned} E(g(X)) &= \int g(x) f_X(x)\, dx \\ &\approx \frac{1}{n} \sum_{i=1}^{n} g(x_i) \quad x_i \sim f(\cdot) \end{aligned}$$

This can be applied to problems where X is also difficult to sample but we can sample the realisations of Y that has a PDF $f_Y(\cdot)$

$$\begin{aligned} E(g(X)) &= \int g(x) f_X(x)\, dx \\ &= \int \frac{g(x) f_X(x)}{f_Y(x)} f_Y(x)\, dx \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \frac{g(y_i) f_X(y_i)}{f_Y(y_i)} \quad y_i \sim f_Y(\cdot) \end{aligned}$$

2.2 Transformation methods

2.2.1 Inverse distribution

When an inverse of a CDF can be calculated, the probability density function of a random variable can be sampled from the unversed CDF with using a uniform distribution $y \sim U(0,1)$.

Example: Sampling exponential random variates using y $\sim U(0,1)$

$F(x) = 1 - e^{-\lambda x} \Rightarrow F^{-1}(y) = -\frac{1}{\lambda} \ln(1-y)$. Samples $x \sim Exp(\lambda)$ can be calculated as simply as

$$x = F^{-1}(y) \quad y \sim U(0,1)$$

2.2.2 Rejection sampling

If we want to simulate a distribution $f(x)$ with support on $[a,b]$ and an determined upper bound m such that $f(x) \leq m, \forall x \in [a,b]$, we can simulate the distribution by drawing the value for x-axis from $x \sim U(a,b)$ and then drawing

the value for y-axis from $y \sim U(0, m)$, and accept the drawn x if $y < f(x)$. The accepted X values will have a PDF f(x).

Intuition: We assume that the acceptance probability is proportional to the probability density when using rejection sampling.

Envelope method is an extension of the rejection sampling method that is applicable to distributions with infinite support. We use another distribution $ah(x)$ to "cover/envelope" the distribution $f(x)$ to be sampled. First, we will draw a point along the y-axis $y \sim h(\cdot)$ and another point along the y-axis $u \sim U(0, ah(y))$ with a greater range. We will accept y as the simulated value of the target distribution if $u < f(y)$.

❖ Conditionality and Markov Processes

The conditional expected value in the discrete case was defined as

$$\Pr(X = x) = \sum_{y=0}^{\infty} P_{X|Y}(x|y)p_Y(y)$$

Then, the conditional expected value of $g(X)$ given $Y = y$ is

$$\mathbb{E}[g(X)|Y = y] = \sum_{x} g(x)p_{X|Y}(x|y) \quad \text{if } p_Y(y) > 0$$

Likelihood $L(\theta|\mathbf{x})$ gives the likelihood of parameter θ given the observed data $\mathbf{x}$.

$$L(\theta|\mathbf{x}) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f}(\mathbf{x_i}|\theta)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ is the most likely θ given the available data $\mathbf{x}$.

3.1 Markov Chains properties

The one-step transition probability in is defined as

$$P_{ij}^{n,n+1} = \Pr\{X_{n+1} = j|X_n = i\}$$

Transition matrix $\mathbf{P} \in \mathbb{R}^{\mathbf{i}\times\mathbf{j}}$ stores the transition probabilities from state i to state j. The transition probabilities $i \to j$ form a probability mass distribution and

the transition probabilities together holds the property of being equal to 1 when summed.

The n-step transition probability matrices is

$$\mathbf{P^{(n)}} = \left[\mathbf{P_{ij}^{(n)}}\right]$$

The probability of initially being in state j is $\Pr(X_0 = j) = p_j$. Thus, the probability of being at state k at time n can be expressed as

$$\Pr(X_n = j | X_0 = i) = \sum_j \mathbf{P_j P_{jk}^n}$$

A distribution is said to be a stationary distribution of the homogeneous Markov chain if

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

π is a row vector denoting the probabilities of being at each state $\pi^{(t)} = \left(\pi^{(t)}(x_1), ..., \pi^{(t)}(x_j)\right)$. If a Markov chains reaches a stationary distribution, it retains for all future time.

The stationary distribution can be solved from

$$\pi(I - P) = 0$$

3.2 Markov Chains in continuous time

In continuous time, a transition kernel can be written $p(x, x', t')$. We can write a transition matrix for each t'.

The transition matrix Q can be seen as a "flow" probability matrix which is the time derivative of the transition matrix P(t').$Q = \frac{d}{dt'} P(t')|_{t'=0}$

When solving for stationary distribution π, use $\pi Q = 0$ and $\sum_i \pi_i = 1$.

Example: Discrete event simulation

1. Initialize the process at $t = 0$ with initial state = i
2. Obtain the flow parameter q_{ii}
3. Simulate the time to the next event t' as $Exp(-q_{ii})$
4. Assign $t := t + t'$ and state = j

Example: Immigration-death process

Population grows by 1 with constant hazard λ. Each individual dies independently with constant hazard μ. The possible transition are

1. $P\{\ X(t+dt) = x+1 | X(t) = x\ \} = \lambda dt$
2. $P\{\ X(t+dt) = x-1 | X(t) = x\ \} = x\mu dt$
3. $P\{\ X(t+dt) = x | X(t) = x\ \} = 1 - (\lambda + x\mu)dt$

These equations define a homogeneous Markov process with infinite state-space S. The stationary distribution of this process is a Poisson distribution with mean λ/μ.

In an inhomogeneous Poisson process, the hazard λ is changing w.r.t. time $\lambda(t)$. Cumulative hazard is defined as the total hazard within a timespan

$$\Lambda(t) = \int_0^t \lambda(s)\, ds$$

The number of events in the interval $(s, t]$: $0 < s < t$ is $Po(\Lambda(t) - \Lambda(s))$.

Example: Inhomogeneous Poisson process

Consider an inhomogeneous Poisson process with rate parameter $\lambda(t) = \lambda t$ for some $t > 0$. The cumulative hazard of the process is

$$\Lambda(t) = \int \lambda t\ dt = \frac{1}{2}\lambda t^2$$

Thus, the number of events $N_t \sim Po(\Lambda(t)) = Po(\frac{1}{2}\lambda t^2)$. The number of events in the interval $(s, t]$ is then $Po(\frac{1}{2}\lambda(t^2 - s^2))$.

3.3 Sampling an inhomogeneous Poisson process

The lectures present two algorithms for sampling an inhomogeneous Poisson process on $(0, T]$ with rate parameter $\lambda(t)$ given that we find an upper bound U_λ satisfying $0 \leq \lambda(t) \leq U_\lambda$.

3.3.1 The "revised" algorithm

In the revised algorithm, we sample the inhomogeneous Poisson process by simulating the number of observation happening at each timestep. This ensures

that all sampled events are in order.

Example: The "revised" algorithm

1. Set $x_0 := 0$
2. Sample $m \sim Po(U_\lambda T)$
3. For $i := 1, ..., m$
   (a) sample $u \sim U(0, 1)$
   (b) Set $x_i := x_{i-1} + (T - x_{i-1})(1 - u^{1/(m-i+1)})$
   (c) Sample $y \sim U(0, U_\lambda)$
   (d) Accept x_i if $y \leq \lambda(x_i)$

3.3.2 Thinning method

In the thinning method, we use the exponential distribution to sample the inter-event times for the x-axis. The point for the y-coordinate is sampled from a $U(0, U_\lambda)$ and accepted if within Ω.

Example: Thinning method

1. Set $t_0 := 0$
2. Set $i := 1$
3. Repeat
   (a) Sample $t \sim Exp(U_\lambda)$
   (b) Set $t := t_{i-1} + t$
   (c) Stop if $t_i > T$
   (d) Sample from $u \sim U(0, U_\lambda)$
   (e) Keep t_i if u $\leq \lambda(t_i)$
   (f) Set $i := i + 1$

❖ MCMC and Bayesian Inference

Definition 4.1. A Markov chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an ergodic Markov chain $X^{(t)}$ whose

stationary distribution is f.

When we measure an outcome $X = x$, we are interested in the conditional probability of the hypothesis given the realisation of X, which is $P(H_i|X = x)$.

We use the occurrence $X = x$ to calculate the posterior belief using our prior knowledge using the Bayes Theorem.

$$P(H_i|X = x) = \frac{P(X = x|H_i)P(H_i)}{\sum_j P(X = x|H_j)P(H_j)}$$

To generalise the idea for also continuous cases, the posterior probability $\pi(\theta|X = x)$ can be expressed as

$$\pi(\theta|X = x) = \frac{\pi(\theta)L(\theta; x)}{\int_\theta P(X = x|\theta')\pi(\theta')\, d\theta'}$$

Because the denominator is just a constant of proportionality, we can express the posterior simply being proportional to the product of prior and the likelihood

$$\text{Posterior} \sim \text{Prior} \times \text{Likelihood}$$

It is, however, not a trivial calculation for most of the non-trivial problems. We need a way to compute posterior densities without any analytical integrations (Hence the MCMC methods...)

4.1 Gibbs Sampler

Gibbs sampler is an MCMC method for simulating multivariate posterior distributions using the parameters' full conditional distributions.

For a bivariate density $\pi(x, y)$, we can use the law of total probability to decompose it into conditional probabilities w.r.t. x and y

$$\textbf{1)} \quad \pi(x, y) = \pi(x)\pi(y|x)$$
$$\textbf{2)} \quad \pi(x, y) = \pi(y)\pi(x|y)$$

In order to sample from $\pi(x, y)$, we would

1. First draw $x \sim \pi(x_o)$ and then $y \sim \pi(y_0|x)$
2. Then we draw $y \sim \pi(y_0)$ and obtain $x' \sim \pi(x|y)$
3. Finally, obtain $y' \sim \pi(y|x')$

Now we have drawn a point $(x', y') \sim \pi(x, y)$ using the full conditionals using the method used by the Gibbs sampler.

4.2 Metropolis-Hastings algorithm

Usually, the closed forms of the full conditional distributions are hard to or impossible to compute. Another MCMC algorithm is called Metropolis-Hastings.

In Metropolis-Hastings, we simulate the target distribution $\pi(\theta)$ using a proposal distribution $q(\cdot, \theta)$. The sampled value will be either accepted or rejected, based on an acceptance probability.

Example: The Metropolis-Hastings algorithm

1. Initiate the counter j and parameter θ
2. Generate a proposal value $\phi \sim q(\theta^{j-1}, \phi)$
3. Evaluate the acceptance probability $\alpha = \min\left\{1, \frac{\pi(\theta)q(\phi,\theta)}{\pi(\phi)q(\theta,\phi)}\right\}$
4. Step $\theta^{(j)} = \phi$ with probability α, otherwise $\theta^{(j)} = \theta^{(j-1)}$

When calculating the acceptance probability α, it's worth noting that if the proposal distribution is symmetric, the probabilities $q(\phi, \theta) = q(\theta, \phi)$ and they will cancel out.

Any distribution $q(\cdot)$ can be used as the proposal distribution, but the choice of the proposal distribution will greatly affect the acceptance rate and distance covered each MH-step. A usual choice is a normal distribution

$$q(\cdot) \sim \mathrm{N}(\mu, 1)$$

where the mean parameter is chose as $\mu = \theta$. When using a normal distribution, MH-algorithm samples the target distribution like a random walker with the distance covered equaling to $\sqrt{k}$ where k is the number of steps.

❖ Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo method uses Hamiltonian dynamics (from mechanics) combined with Metropolis sampling to construct an MCMC method.

5.1 Hamiltonian dynamics

The Hamiltonian function

$$H(q,p) = \underbrace{U(q)}_{\substack{\text{potential}\\ \text{energy}}} + \underbrace{K(p)}_{\substack{\text{kinetic}\\ \text{energy}}} = E_{tot}$$

Gives the total and constant energy of the system, which consists of the potential energy and the kinetic energy.

In order to understand how the model's behaviour change over time t, we define the Hamilton's equations

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \qquad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

The kinetic energy's definition is borrowed from the mechanics and can be modelled as

$$K(p) = \frac{p^T M^{-1} p}{2}$$

where M is a positive-definite matrix representing the weight of the "particle".

Hamiltonian stays invariant as the energy of the system stays constant.

5.2 Discretisation of Hamilton's equations

Hamilton's equations can be solved numerically using different discretisation methods, leapfrog integration being the most used one.

Example: The leapfrog method

Propagate p_i in half-steps.

1. $p_i(t + \epsilon/2) = p_i(t) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t))$
2. $q_i(t + \epsilon) = q_i(t) + \epsilon \frac{p_i(t+\epsilon/2)}{m_i}$
3. $p_i(t + \epsilon) = p_i(t + \epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_i}(q(t + \epsilon))$

In short, the target of HMC is to **translate the density function of the target distribution to the potential energy function** and introduce momentum variables to go with the original variable(s) of interest. Finally, using a Markov Chain for each iteration, resample the momentum and **perform a Metropolis update with a proposal found using Hamiltonian dynamics**.

Imagine a probability distribution with a form of a skate pool. The depth of the pool equals to the probability density of the distribution. Using HMC, we will simulate the shape of the skate pool by placing a hockey puck on the ground, and kick it around the skate pool. The momentum of the puck is modelled by the kinetic energy and the position of the puck by the parameters of the potential

energy function. The points where the puck lands after each kick is equals the proposal in the HMC sampling.

Coming back to the formal definition, the density function can be translated to potential energy function accordingly:

$$\begin{aligned} H(q,p) &= -\log[\pi(p|q)] - \log[\pi(q)] \\ &\equiv K(q,p) + V(q) \end{aligned}$$

Example: A single iteration of HMC

1. Initiate $q^* = q_0$
2. Sample $p_0 \sim N(0,1)$; $p^* := p_0$
3. Leapfrog integration

   Half step for momentum
   $p^* := p^* - (\epsilon/2) \cdot dU(q^*)/dq$

   Full step for position
   $q^* := q^* - \epsilon \cdot p^*$

   Half step for momentum
   $p^* := p^* - (\epsilon/2) \cdot dU(q^*)/dq$
4. Negate the momentum at the end $p^* := -p^*$
5. Evaluate the change in potential and kinetic energy

   $U_0 = U(q_0) \qquad K_0 = \frac{1}{2}p_0^2$

   $U^* = U(q^*) \qquad K^* = \frac{1}{2}p^{*2}$
6. Accept or reject proposal

   $u \sim U(0,1)$

   if $u < \exp(U_0 - U^* + K_0 - K^*)$ then $q^* = q^*$

   else $q^* = q^0$

If the target distribution doesn't have the full support $(-\infty, \infty)$, we can introduce constraints (boundaries) by treating them as fully elastic collisions, i.e. the energies are mirrored with respect to the boundaries.