MCMC Sampling

- Consider a distribution over (finite) S: $\pi(x) = \frac{b(x)}{Z}$
- Since this is a distribution, $Z = \sum_{x \in S} b(x)$
- We assume, we can efficiently calculate $b(x)$ for any x but computation of Z is intractable or computationally expensive
  E.g., the Boltzmann distribution: $b(x) = e^{-E(x)/KT}$
- We want $E[g(X)]$ w.r.t. distribution π (for any g)

$$E[g(X)] = \sum_{x} g(x)\, \pi(x) \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i), \quad X_1, \cdots X_n \sim \pi$$

- One way to generate samples is to design an ergodic markov chain with stationary distribution π
  – MCMC sampling

- Suppose $\{X_n\}$ is a an irreducible, aperiodic positive recurrent Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
- Then we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} g(X_m) = \sum_x g(x)\pi(x)$$

- hence, if we can design a Markov chain with a given stationary distribution, we can use that to calculate the expectation.
- We can also use the chain to generate samples from distribution π

- $\{X_n\}$: Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
  We can approximate the expectation as

$$\sum_x g(x)\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_{M+i})$$

  Where M is large enough to assume chain is in steady state

- When we take sample mean, $\frac{1}{n}\sum_{i=1}^{n} Z_i$, we want Z_i to be uncorrelated
- We can, for example, use

$$\sum_x g(x)\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(X_{M+Ki})$$

- For all these, we need to design a Markov chain with π as stationary distribution

- Let $Q = [q(i,j)]$ be the transition probability matrix of an irreducible Markov chain over S.
- Q is called the proposal distribution
- We start with arbitrary X_0 and generate $X_{n+1},\ n = 0, 1, 2, \cdots$, iteratively as follows
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i,k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \ \alpha(i,j) \\ X_n & \text{with probability } \ 1 - \alpha(i,j) \end{cases}$$

- $\alpha(i,j)$ is called the acceptance probability
- We want to choose $\alpha(i,j)$ to make X_n an ergodic Markov chain with stationary probabilities π

- The stationary distribution π satisfies (with transition probabilities P)

$$\pi(y) = \sum_x \pi(x)\ P(x,y), \ \ \forall y \in S$$

- Suppose there is a distribution $g(\cdot)$ that satisfies

$$g(y)\ P(y,x) = g(x)\ P(x,y), \ \ \forall x, y \in S$$

  This is called detailed balance
- Summing both sides above over x give

$$g(y) = \sum_x g(y)\ P(y,x) = \sum_x g(x) P(x,y), \ \ \forall y$$

- Thus if $g(\cdot)$ satisfies detailed balance, then it must be the stationary distribution
- Note that it is not necessary for a stationary distribution to satisfy detailed balance

- Any stationary distribution has to satisfy

$$\pi(y) = \sum_x \pi(x)\ P(x,y), \ \ \forall y \in S$$

- If I can find a π that satisfies

$$\pi(x)P(x,y) = \pi(y)P(y,x), \ \ \forall x, y \in S, \ x \neq y$$

  that would be the stationary distribution
- This is called detailed balance

- Recall our algorithm for generating $X_n,\ n = 0, 1, \cdots$
- Start with arbitrary X_0 and generate X_{n+1} from X_n
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i, k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \alpha(i,j) \\ X_n & \text{with probability } 1 - \alpha(i,j) \end{cases}$$

- Hence the transition probabilities for X_n are

$$\begin{aligned} P(i,j) &= q(i,j)\,\alpha(i,j), \quad i \neq j \\ P(i,i) &= q(i,i) + \sum_{j \neq i} q(i,j)\,(1 - \alpha(i,j)) \end{aligned}$$

- $\pi(i) = b(i)/Z$ is the desired stationary distribution
- So, we can try to satisfy

$$\begin{aligned} \pi(i)\ P(i,j) &= \pi(j)\ P(j,i), \ \ \forall i, j, i \neq j \\ \text{that is,}\quad b(i)q(i,j)\,\alpha(i,j) &= b(j)q(j,i)\,\alpha(j,i) \end{aligned}$$

- We want to satisfy

$$b(i)q(i,j)\ \alpha(i,j) = b(j)q(j,i)\ \alpha(j,i)$$

- Choose

$$\alpha(i,j) = \min\left(\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right)$$

- Note that one of $\alpha(i,j),\ \alpha(j,i)$ is 1

$$\begin{aligned} \text{suppose} \ \ \alpha(i,j) &= \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)} \ < \ 1 \\ \Rightarrow \ \ \pi(i)\ q(i,j)\ \alpha(i,j) &= \pi(j)\ q(j,i) \\ &= \pi(j)\ q(j,i)\ \alpha(j,i) \end{aligned}$$

- Note that $\pi(i)$ above can be replaced by $b(i)$

Metropolis-Hastings Algorithm

- Start with arbitrary X_0 and generate X_{n+1} from X_n
  - If $X_n = i$, we generate Y with $Pr[Y = k] = q(i, k)$
  - Let the generated value for Y be j. Set

$$X_{n+1} = \begin{cases} j & \text{with probability } \alpha(i,j) \\ X_n & \text{with probability } 1 - \alpha(i,j) \end{cases}$$

  Where $Q = [q(i, j)]$ is the transition probabilities of an irreducible chain and

$$\alpha(i,j) = \min\left(\frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}, 1\right)$$

- Then $\{X_n\}$ would be an irreducible, aperiodic chain with stationary distribution π.
- Q is called the proposal chain and $\alpha(i, j)$ is called acceptance probabilities

- Consider Boltzmann distribution: $b(x) = e^{-E(x)/KT}$
- Take proposal to be uniform: from any state, we go to all other states with equal probabilities
- Then,

$$\alpha(x, y) = \min\left(\frac{b(y)}{b(x)}, 1\right) = \min\left(e^{-(E(y)-E(x))/KT}, 1\right)$$

- In state x you generate a random new state y.
  If $E(y) \leq E(x)$ you always go there;
  if $E(y) > E(x)$, accept with probability $e^{-(E(y)-E(x))/KT}$
- An interesting way to simulate Boltzmann distribution
- We could have chosen Q to be 'uniform over neighbours'

- Suppose $E : S \rightarrow \Re$ is some function.
- We want to find $x \in S$ where E is *globally* minimized.
- A gradient descent type method tries to find a locally minimizing direction and hence gives only a ‘local’ minimum.
- The Metropolis-Hastings algorithm gives another view point on how such optimization problems can be handled.
- We can think of E as the energy function in a Boltzmann distribution

- Let $b(x) = e^{-E(x)/T}$ where T is a parameter called 'temparature'
- $\{X_n\}$ be Markov chain with stationary dist $\pi(x) = \frac{b(x)}{Z}$
- We can find relative occupation of different states by the chain by collecting statistics during steady state
- We know

$$\frac{\pi(x_1)}{\pi(x_2)} = \frac{b(x_1)}{b(x_2)} = e^{-(E(x_1)-E(x_2))/T}$$

- We spend more time in global minimum
  We can increase the relative fraction of time spent in global minimum by decreasing T (There is a price to pay!)
- Gives rise to interesting optimization technique called simulated annealing

- In most applications of MCMC, $x \in S$ is a vector.
- One normally changes one component at a time. That is how neighbours can be defined
- A special case of proposal distribution is the conditional distribution.
- Suppose $X = (X_1, \cdots, X_N)$. To propose a value for X_i, we use $f_{X_i|X_{-i}}$
- Here the conditional distribution is calculated using the target π as the joint distribution.
- With such a proposal distribution, one can show that $\alpha(i,j)$ is always 1
- This is known as Gibbs sampling

Random process

- A random process or a stochastic process is a collection of random variables: $\{X_t, \ t \in T\}$
- Markov chain is an example. Here $T = \{0, 1, \cdots\}$
- We call T the index set.
- Normally, T is either (a subset of) set of integers or an interval on real line.
- We think of the index t as time
- Thus a random process can represent the time-evolution of the state of a system
- We assume T is infinite
- The index need not necessarily represent time. It can represent, for example, space coordinates.

- A random process: $\{X_t,\ t \in T\}$
- The set T can be countable e.g., $T = \{0, 1, 2, \cdots\}$
- Or, T can be continuous e.g., $T = [0, \infty)$
- These are termed **discrete-time** or **continuous-time** processes
- The random variables, X_t, may be discrete or continuous
- These are termed **discrete-state** or **continuous-state** processes
- The Markov chain we considered is a discrete-time discrete-state process

- A random process: $\{X_t,\ t \in T\}$
- We can think of this as a mapping: $X : \Omega \times T \to \Re$
- Thus, $X(\omega, \cdot)$ is a real-valued function over T.
- So, we can think of the process also as a collection of time functions.
- X can be thought of as a map that associates with each $\omega \in \Omega$ a real-valued function on T.
- These functions are called sample paths or paths of the process
- We can view the random process as a collection of random variables, or as a collection of functions
- We will denote the random variables as X_t or $X(t)$

- A finite collection of random variables is completely specified by its joint distribution
- How do we characterize a random process?
- We need to specify joint distribution of $X_{t_1}, X_{t_2}, \cdots X_{t_n}$ for all n and all $t_1, t_2, \cdots t_n \in T$..
- One can show this completely specifies the process.
- As we saw, for a Markov chain, π_0 and P together specify all such joint distributions

Distributions of a random process

- A random process: $\{X_t,\ t \in T\}$ or $X : \Omega \times T \to \Re$
- The first order distribution function of X is

$$F_X(x;t) = Pr[X_t \leq x] = F_{X_t}(x)$$

- The second order distribution function of X is

$$F_X(x_1, x_2; t_1, t_2) = Pr[X_{t_1} \leq x_1,\ X_{t_2} \leq x_2]$$

- The n^{th} order distribution function of X is

$$F_X(x_1, \cdots, x_n; t_1, \cdots t_n) = Pr[X_{t_i} \leq x_i,\ i = 1, \cdots, n]$$

- When it is a discrete-state process, all X_t would be discrete random variables
- We can specify distributions through mass functions:

$$f_X(x;t) = Pr[X_t = x] = f_{X_t}(x)$$

$$f_X(x_1, x_2; t_1, t_2) = Pr[X_{t_1} = x_1,\ X_{t_2} = x_2]$$

$$f_X(x_1, \cdots, x_n; t_1, \cdots t_n) = Pr[X_{t_i} = x_i,\ i = 1, \cdots, n]$$

- If all X_t are continuous random variables and if all distributions have density functions, then we denote joint density of $X_{t_1}, \cdots, X_{t_n}$ by $f_X(x_1, \cdots, x_n; t_1, \cdots t_n)$

- Specifying the n^{th} order distributions for all n separately is not feasible.
- Hence one needs some assumptions on the model so that these are specified implicitly.
- One example is the Markovian assumption.
- As we saw, in a Markov chain, the transition probabilities and initial state probabilities would determine all the distributions
- Another such useful assumption is what is called a process with independent increments

- A random process $\{X(t),\ t \in T\}$ is said to be a process with independent increments if

  for all $t_1 < t_2 \leq t_3 < t_4$, *the random variables* $X(t_2) - X(t_1)$ *and* $X(t_4) - X(t_3)$ *are independent*

- Note that this also implies, e.g., $X(t_1)$ is independent of $X(t_2) - X(t_1)$ for all $t_1 < t_2$.
- Now suppose this is a discrete-state process.
- Then we can write n^{th} order pmf's as

$$
\begin{aligned}
& Pr[X(t_1) = x_1, X(t_2) = x_2, \cdots X(t_n) = x_n] \\
& \quad = Pr[X(t_1) = x_1, X(t_2) - X(t_1) = x_2 - x_1, \cdots] \\
& \quad = Pr[X(t_1) = x_1]\ Pr[X(t_2) - X(t_1) = x_2 - x_1] \ \cdots \\
& \qquad\qquad \cdots Pr[X(t_n) - X(t_{n-1}) = x_n - x_{n-1}]
\end{aligned}
$$

- We only need up to second order distributions

- Let $\{X(t),\ t \in T\}$ be a discrete-state process with independent increments
- Then we specify $f_X(x;t)$ and another function

$$g(x_1, x_2; t_1, t_2) = Pr[X(t_2) - X(t_1) = x_2 - x_1]$$

- Now we can get all distributions as

$$\begin{aligned} f_X(x_1, \cdots, x_n; t_1, \cdots t_n) \\ &= Pr[X(t_i) = x_i,\ i = 1, \cdots, n] \\ &= f_X(x_1; t_1) \prod_{i=1}^{n-1} Pr[X(t_{i+1}) - X(t_i) = x_{i+1} - x_i] \\ &= f_X(x_1; t_1) \prod_{i=1}^{n-1} g(x_i, x_{i+1}; t_i, t_{i+1}) \end{aligned}$$

- Given a random process $\{X(t),\ t \in T\}$
- Its mean or mean function is defined by

$$\eta_X(t) = E[X(t)],\ \ t \in T$$

- We define the autocorrelation of the process by

$$R_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

- We define the autocovariance of the process by

$$\begin{aligned} C_X(t_1, t_2) &= E\left[(X(t_1) - E[X(t_1)])(X(t_2) - E[X(t_2)])\right] \\ &= R_X(t_1, t_2) - \eta_X(t_1)\eta_X(t_2) \end{aligned}$$

Stationary Processes

- A random process $\{X(t),\ t \in T\}$ is said to be stationary if

  for all n*, for all* $t_1, \cdots, t_n$*, for all* $x_1, \cdots x_n$ *and for all* τ *we have*

$$F_X(x_1, \cdots, x_n\ ;\ t_1, \cdots, t_n) = F_X(x_1, \cdots, x_n\ ;\ t_1+\tau, \cdots, t_n+\tau)$$

- For a stationary process, the distributions are unaffected by translation of the time axis.
- This is a rather stringent condition and is often referred to as strict-sense stationarity

- A homogeneous Markov chain started in its stationary distribution is a stationary process
- As we know, if π_0 is the stationary distribution then π_n is same for all n.
- This, along with the Markov condition would imply that shift of time origin does not affect the distributions

$$
\begin{aligned}
Pr[X_n = x_0, X_{n+1} = x_1, \cdots X_{n+m} = x_m] \\
&= \pi_n(x_0) P(x_0, x_1) \cdots P(x_{m-1}, x_m) \\
&= \pi_0(x_0) P(x_0, x_1) \cdots P(x_{m-1}, x_m) \\
&= Pr[X_0 = x_0, X_1 = x_1, \cdots X_m = x_m]
\end{aligned}
$$

- Suppose $\{X(t),\ t \in T\}$ is (strict-sense) stationary
- Then the first order distribution is independent of time

$$F_X(x;t) = F_X(x;t+\tau),\ \forall x, t, \tau \quad \Rightarrow \quad \text{e.g.,} \quad F_X(x;t) = F_X(x;0)$$

- This implies $\eta_X(t) = \eta_X$, a constant
- The second order distribution has to satisfy

$$F_X(x_1, x_2; t, t+\tau) = F_X(x_1, x_2; 0, \tau),\ \forall x_1, x_2, t, \tau$$

  Hence $F_X(x_1, x_2; t_1, t_2)$ can depend only on $t_1 - t_2$
- This implies

$$R_X(t, t+\tau) = E[X(t)X(t+\tau)] = R_X(\tau)$$

  Autocorrelation depends only on the time difference

- The process $\{X(t),\ t \in T\}$ is said to be wide-sense stationary if

$$\begin{aligned} F_X(x;t) &= F_X(x;t+\tau),\ \forall x,t,\tau \\ F_X(x_1,x_2;t_1,t_2) &= F_X(x_1,x_2;t_1+\tau,t_2+\tau) \end{aligned}$$

- The process is wide-sense stationary if the first and second order distributions are invariant to translation of time origin

- Let $\{X(t),\ t \in T\}$ be wide-sense stationary. Then

1. $\eta_X(t) = \eta_X$, a constant
2. $R_X(t_1, t_2)$ depends only on $t_1 - t_2$

- In many engineering applications, we call a process wide-sense stationary if the above two hold.
- In this course we take the above as the definition of wide-sense stationary process
- When the process is wide-sense stationary, we write autocorrelation as

$$R_X(\tau) = E[X(t)X(t+\tau)]$$

Ergodicity

- Suppose $X(n)$ is a discrete-time discrete-state process (like a Markov chain)
- Suppose it is wide-sense stationary.
  Then $E[X(n)]$ does not depend on n
- Ergodicity is the question of

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(i) \overset{?}{=} E[X(n)] = \eta_X$$

- We proved that this is true for an irreducible, aperiodic, positive recurrent Markov chain (with a finite state space)
- The question is : do 'time-averages' converge to 'ensemble-averages'
- The process is wide-sense stationary and hence all $X(n)$ have the same distribution; but they need not be independent or uncorrelated (e.g., Markov chain)

- Ergodicity is a question of whether time-averages converge to ensemble-averages?

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(i) \ \overset{?}{=}\ E[X(n)] = \eta_X$$

  Or, more generally

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(X(i)) \ \overset{?}{=}\ E[g(X(n))]$$

  For a continuous time process we can write this as

$$\lim_{\tau\to\infty} \frac{1}{2\tau} \int_{-\tau}^{\tau} X(t)\, dt \ \overset{?}{=}\ E[X(t)] = \eta_X$$

- Essentially if there is no long-term correlation in the process this may hold.
- One sufficient condition could be that covariance between $X(t)$ and $X(t+\tau)$ decreases fast with increasing τ.

- Define

$$\eta_\tau = \frac{1}{2\tau} \int_{-\tau}^{\tau} X(t)\, dt \quad (\tau > 0)$$

- For each τ, η_τ is a rv. We write η for η_X.
- We say the process is mean-ergodic if

$$\eta_\tau \xrightarrow{P} \eta, \quad \text{as} \;\; \tau \to \infty$$

- That is, if

$$\lim_{\tau \to \infty} Pr\left[|\eta_\tau - \eta| > \epsilon\right] = 0, \;\; \forall \epsilon > 0$$

- Note that $E[\eta_\tau] = \eta, \; \forall \tau$.
- Hence it is enough if we show

$$\sigma_\tau^2 \triangleq E\left[(\eta_\tau - \eta)^2\right] \; \to \; 0, \quad \text{as} \quad \tau \to \infty$$

- Let $C_X(t_1, t_2)$ be the autocovariance of the process

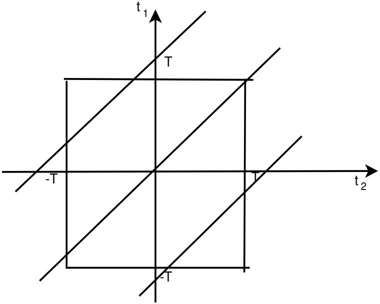
$$C_X(t_1, t_2) = E[(X(t_1) - \eta)(X(t_2) - \eta)]$$

- Assuming wide-sense stationarity, $C_X(t_1, t_2) = C_X(t_1 - t_2)$
- We can get σ_τ^2 as

$$\begin{aligned}
\sigma_\tau^2 &= E\left[(\eta_\tau - \eta)^2\right] \\
&= E\left[\frac{1}{2\tau}\int_{-\tau}^{\tau}(X(t) - \eta)\;dt\;\frac{1}{2\tau}\int_{-\tau}^{\tau}(X(t') - \eta)\;dt'\right] \\
&= \frac{1}{4\tau^2}\int_{-\tau}^{\tau}\int_{-\tau}^{\tau} E[(X(t) - \eta)(X(t') - \eta)]\;dt\;dt' \\
&= \frac{1}{4\tau^2}\int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t - t')\;dt\;dt'
\end{aligned}$$

$$\text{Let} \quad I = \int_{-\tau}^{\tau} \int_{-\tau}^{\tau} C_X(t_1 - t_2) \, dt_2 \, dt_1$$

- Let $z = t_1 - t_2$. We want to change the integration to be over t_2 and z



- Easy to see z goes from -2τ to 2τ
  When $z \geq 0$, for a given z, t_2 goes from $-\tau$ to $\tau - z$
  When $z < 0$, for a given z, t_2 goes from $-\tau - z$ to τ

- Now we get

$$
\begin{aligned}
I &= \int_{-\tau}^{\tau}\int_{-\tau}^{\tau} C_X(t_1 - t_2)\, dt_2\, dt_1 \\
&= \int_{-2\tau}^{0}\int_{-\tau-z}^{\tau} C_X(z)\, dt_2\, dz \;+\; \int_{0}^{2\tau}\int_{-\tau}^{\tau-z} C_X(z)\, dt_2\, dz \\
&= \int_{-2\tau}^{0} C_X(z)\,(\tau - (-\tau - z))\, dz \;+\; \int_{0}^{2\tau} C_X(z)\,(\tau - z - (-\tau))\, dz \\
&= \int_{-2\tau}^{0} C_X(z)\,(2\tau + z)\, dz \;+\; \int_{0}^{2\tau} C_X(z)\,(2\tau - z)\, dz \\
&= \int_{-2\tau}^{2\tau} C_X(z)\,(2\tau - |z|)\, dz
\end{aligned}
$$

- Now we get σ_τ^2 as

$$\begin{aligned}
\sigma_\tau^2 &= \frac{1}{4\tau^2} \int_{-\tau}^{\tau} \int_{-\tau}^{\tau} C_X(t - t')\ dt\ dt' \\
&= \frac{1}{4\tau^2} \int_{-2\tau}^{2\tau} C_X(z)\ (2\tau - |z|)\ dz \\
&= \frac{1}{2\tau} \int_{-2\tau}^{2\tau} C_X(z)\ \left(1 - \frac{|z|}{2\tau}\right)\ dz
\end{aligned}$$

- Hence, a sufficient condition for $\sigma_\tau^2 \to 0$ is

$$\int_{-\infty}^{\infty} |C_X(z)|\ dz\ <\ \infty$$

- This is a sufficient condition for the process being mean-ergodic