Lecture 2 - Monte Carlo Methods

Dahua Lin
The Chinese University of Hong Kong

1 Monte Carlo Methods

Monte Carlo methods are a large family of computational algorithms that rely on *random sampling*. These methods are mainly used for

- Numerical integration
- Stochastic optimization
- Characterizing distributions

---

2 Motivation: Expectations in Statistical Analysis

Computing *expectation* is perhaps the most common operation in statistical analysis.

- Computing the *normalization factor* in posterior distribution:

$$p(x|y) = \frac{p(y|x)p(x)}{\int_X p(y|x')p(x')dx'} = \frac{p(x,y)}{E_X[p(X,y)]}.$$

- Computing marginalization:

$$p(x) = \int_Z p(x,z)dz = E_Z[p(x,Z)].$$

- Computing expectation of functions:

$$E_X(f(X)) = \int_X f(x)p(x)dx.$$

---

3 Computing Expectation

- Generally, *expectation* can be written as.

$$E[f(x)] = \int f\mu(dx).$$

- This formula can be expanded to different forms depending on the base measure dx:
  - For discrete space:

$$E[f(x)] = \sum_{k \in \Omega} f(k) P(x = k).$$

– For continuous space and *Riemann-integrable* function f:

$$E[f(x)] = \int_{\Omega} f(x) dx.$$

- Complexity grow exponentially as dimension of Ω increases

---

4 Monte Carlo Integration

- *(Strong) Law of Large Numbers (LLN)*: Let $X, X_1, X_2, \ldots$ be i.i.d random variables and f be a measurable function. Let $I_n(f) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(X_i)$, then

$$I_n(f) \xrightarrow{a.s.} E[f(X)], \text{ as } n \to \infty.$$

- We can use *sample mean* to *approximate* expectation:

$$E[f(X)] \simeq \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

- How many samples are enough?

---

5 Variance of Sample Mean

- By the *Central Limit Theorem (CLT)*:

$$\sqrt{n} \left(I_n(f) - E[(f(X))] \right) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2), \text{ as } n \to \infty.$$

  Here, σ_f^2 is the variance of $f(X)$.

- The variance of $f(X)$ is σ_f^2 / n. The number of samples required to attain a certain variance σ_ϵ^2 is at least $\sigma_f^2 / \sigma_\epsilon^2$.

- The variance σ_f^2 is usually a polynomial of the dimension of X, and even a constant in some cases.

---

6 Random Number Generation

- All sampling methods rely on a stream of *random numbers* to construct random samples.
- *"True"* random numbers are difficult to obtain. A more widely used approach is to use computational algorithms to produce long sequences of *apparently random* numbers, called *pseudorandom numbers*.
- The sequence of *pseudorandom numbers* is determined by a *seed*.
  - If a randomized simulation is based on a single *random stream*, it can be *exactly* reproduced by fixing the *seed* initially.

---

7 Pseudorandom Number Generators

- Linear Congruential Generator (LCG): $m = 2^{32}$ or 2^{64}.
  - C and Java's builtin.
  - Useful for simple randomized program.
  - Not good enough for serious Monte Carlo simulation.
- Mersenne Twister (MT): $m = 2^{19937} - 1$
  - Passes *Die-hard* tests, good enough for most Monte Carlo experiments
  - Provided by C++'11 or Boost
  - Default RNG for MATLAB, Numpy, Julia, and many other numerical softwares
  - Not amenable to parallel use
- Xorshift1024: $m = 2^{1024}$
  - Proposed in year 2014
  - Passes *BigCrush*
  - Incredibly simple (*5 - 6* lines of C code)

---

8 Sampling from a Discrete Distribution

Let p be the *probability mass function* over $\{1, \ldots, K\}$. Please design an algorithm to sample from this distribution and analyze its complexity.

. . .

- Linear search: $O(K)$
- Sorted search: $O(K)$, but much faster when prob. mass concentrates on a few values
- Binary search: $O(\log_2 K)$, but each step is a bit more expensive
- Can we do better?

. . .

- Huffman coding: $O(K)$ preprocessing + $O(Entropy)$ per sample
- Alias methods (by *A. J. Walker*): $O(K)$ preprocessing + $O(1)$ per sample

---

9 Transform Sampling

- Let F be the *cumulative distribution function (cdf)* of a distribution D. Let $U \sim \text{Uniform}([0,1])$, then $F^{-1}(U) \sim D$.

. . .

- **(Proof):** Let $X = F^{-1}(U)$:

$$P(X \leq t) = P(F^{-1}(U) \leq t) = P(U \leq F(t)) = F(t).$$

- How to generate a *exponentially distributed* sample ?

. . .

- How to draw a sample from a *multivariate normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?

. . .

- **(Algorithm):**
  1. Perform Cholesky decomposition to get $\mathbf{L}$ s.t. $\mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}$.
  2. Generate a vector $\mathbf{z}$ comprised of iid values from $\mathcal{N}(0,1)$
  3. Let $\mathbf{x} = \mathbf{L}\mathbf{z}$.

---

10 Rejection Sampling

- **(Rejection sampling algorithm):** To sample from a distribution $p(x)$, which has $p(x) \leq Mq(x)$ for some $M < +\infty$:
  1. Sample $x \sim q(x)$
  2. Accept x with probability $\frac{p(x)}{Mq(x)}$.

. . .

- Expected acceptance rate:

$$P(\text{accept}) = \int \frac{p(x)}{Mq(x)} q(x) \mu(dx) = \frac{1}{M}.$$

- What are the problems of this method?

---

11 Importance Sampling

- Basic idea: generate samples from an *easier distribution* $q(x)$, which is often referred to as the *proposal distribution*, and then reweight the samples.
- Let $p(x)$ be the *target distribution* and $q(x)$ be the *proposal distribution*, which have $p \prec q$ (meaning p is absolutely continuous *w.r.t.* q), and let the *importance weight* be $w(x) = p(x)/q(x)$. Then

$$E_p[f(x)] = E_q[w(x) \cdot f(x)].$$

- We can approximate $E_p[f(x)]$ with:

$$E_p[f(x)] \simeq I_{q,n}(f) \triangleq \frac{1}{n} \sum_{i=1}^{n} w(x_i) f(x_i), \quad \text{with } x_1, \ldots, x_n \sim q$$

  By the *strong law of large numbers*, we have $I_{q,n}(f) \xrightarrow{a.s.} E_p[f]$, as $n \to \infty$.

---

12 Variance of Importance Sampling

- How to choose a good proposal distribution q?

. . .

- The variance of $w(x)f(x)$ is

$$\mathrm{var}_q\left(w(x)f(x)\right) = E_q\left[w^2(x)f^2(x)\right] - \left(E_p[f]\right)^2$$

  The 2nd term does not depend on q, while the 1st term has

$$E_q\left[w^2(x)f^2(x)\right] \geq \left(E_q\left[w(x)\left|f(x)\right|\right]\right)^2 = \left(E_p\left[\left|f(x)\right|\right]\right)^2.$$

  The lower bound is attached when $w(x)\left|f(x)\right|$ is a constant:

$$\hat{q}(x) = \frac{\left|f(x)\right| p(x)}{\int \left|f(x)\right| p(x) \mu(dx)}.$$

. . .

- The *optimal proposal distribution* is generally difficult to sample from. However, this analysis leads us to an insight: we can achieve high sampling efficiency by emphasizing regions where the value of $p(x)\left|f(x)\right|$ is high.

---

13 Adaptive Importance Sampling

- Basic idea – *Learn* to do sampling
  - choose the proposal from a tractable family: $q(x;\theta)$.

. . .

- Objective: minimize the sample mean of $f^2(x)w^2(x;\theta)$. Update the parameter θ as:

$$\theta_{t+1} \leftarrow \theta_t - \alpha \frac{1}{N} \sum_{i=1}^{N} f^2(x_i) w(x_i;\theta_t) \nabla_\theta w(x_i,\theta_t)$$

  with $x_1, \ldots, x_N \sim q(x;\theta_t)$.

---

14 Self-Normalized Weights

- In many practical cases, $w(x) \propto p(x)/q(x)$ is known only upto a normalizing constant. For such case, we can write:

$$E_p[f] = \frac{\int f(x)w(x)q(x)\mu(dx)}{\int w(x)q(x)\mu(dx)}.$$

- Hence, we may approximate $E_p[f]$ with

$$\tilde{I}_{q,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{w}_i f(x_i)$$

  Here, $\tilde{w}_i$ is called the *self-normalized weight$, given by $\tilde{w}_i = \frac{w(x_i)}{\sum_{i=1}^{n} w(x_i)}$.

- By strong law of large numbers, we have $\tilde{I}_{q,n}(f) \xrightarrow{a.s.} E_p[f]$ as $n \to \infty$.

---

15 MCMC: Motivation and Overview

- Simple strategies like *transform sampling*, *rejection sampling*, and *importance sampling* all rely on drawing *independent samples* from p or a proposal distribution q over the sample space.
  - This can become very difficult (if not impossible) for complex distributions.

. . .

- **Markov Chain Monte Carlo (MCMC)** explores the sample space through an *ergodic* Markov chain, whose equilibrium distribution is the target distribution p.
- Many sampling methods used in practice belong to MCMC:
  - Gibbs sampling
  - Metropolis-Hastings algorithm

- Slice sampling
- Reversible Jump
- ...

---

16 Markov Processes

- Intuitively, a *Markov process* is a stochastic process for which the future depends only on the present and not on the past.
- A sequence of random variables $X_0, X_1, X_2, \ldots$ defined on a measurable space Ω is called a *(discrete-time) Markov process* if it satisfies the *Markov property*:

$$P(X_{t+1} \in A | X_0 = x_0, \ldots, X_t = x_t) = P(X_{t+1} \in A | X_t = x_t)$$

  Here, A is an arbitrary measurable subset of Ω.

. . .

- We first review the formulation and properties of *Markov chains* under a simple setting, where Ω is a countable space. We will later extend the analysis to more general spaces.

---

17 Homogeneous Markov Chains

- A *homogeneous Markov chain* on a countable state space Ω, denoted by $Markov(\pi_0, P)$ is characterized by an initial distribution π_0 and a *transition probability matrix (TPM)*, denoted by P, such that
  - $\Pr(X_0 = x) = \pi_0(x)$, and
  - $\Pr(X_{t+1} = y | X_t = x) = P(x, y)$.
- The TPM P is non-negative and has $\sum_{y \in \Omega} P(x, y) = 1, \forall x \in \Omega$. Such a matrix is called a *stochastic matrix*.

. . .

- Let π_t be the distribution of X_t, then:

$$\pi_{t+1}(y) = \sum_{x \in \Omega} \pi_t(x) P(x, y),$$

  or simply $\pi_{t+1} = \pi_t P$.

---

18 Multi-step Transition Probabilities

- Consider two transition steps:

$$\Pr(X_2 = y | X_0 = x) = \sum_{z \in \Omega} P(x,z)P(z,y) = P^2(x,y)$$

. . .

- More generally, $\Pr(X_{t+m} = y | X_t = x) = P^m(x,y)$.
- Let π_t be the distribution of X_t, then $\pi_{t+m} = \pi_t P^m$.

---

19 Classes of States

- A state y is said to be *accessible* from state x, or x *leads to* y, denoted by $x \to y$, if $P^n(x,y) > 0$ for some n.
- States x and y are said to *communicate* with each other, denoted by $x \leftrightarrow y$, if $x \to y$ and $y \to x$.
- $x \leftrightarrow y$ is an *equivalence relation* on Ω, which partitions Ω into *communicating classes*, where states within the same class communicate with each other.

. . .

- The Markov chain is said to be *irreducible* if it forms a single communicating class, or in other words, all states communicate with any other states.

---

20 Periodicity of Markov Chains

- The *period* of a state x is defined as

$$\text{period}(x) \triangleq \gcd\{n : P^m(x,x) > 0\}.$$

- A state x is said to be *aperiodic* if $\text{period}(x) = 1$.
- *Period* is a class property: if $x \leftrightarrow y$, then $\text{period}(x) = \text{period}(y)$.
- A Markov chain is called *aperiodic*, if all states are *aperiodic*.
- An *irreducible Markov chain* is *aperiodic*, if there exists an *aperiodic* state.
- Lazyness breaks periodicity: $\alpha I + (1-\alpha)P$.

---

21 Recurrence of Markov Chains

- Suppose the chain is initially at state x, the *first return time* to state x is defined to be

$$\tau_x = \inf \{t \geq 1 : X_t = x | X_0 = x\}.$$

  Note that τ_x is a random variable.

- We also define $f_{xx}^{(n)} = \Pr(\tau_x = n)$, the probability that the chain returns to x *for the first time* after n steps.

- A state x is said to be *recurrent* if it is guaranteed to have a *finite hitting time*, as

$$\Pr(\tau_x < \infty) = \sum_{n=1}^{\infty} f_{xx}^{(n)} = 1.$$

  Otherwise, x is said to be *transient*.

---

22 Recurrence of Markov Chains (cont'd)

- A state x is *recurrent* if and only if the chain returns to x *infinitely often*:

$$\sum_{n=1}^{\infty} P^n(x, x) = \infty.$$

- *Recurrence* is a class property: if $x \leftrightarrow y$ and x is *recurrent*, then y is also *recurrent*.

- Every *finite* communicating class is *recurrent*.

  - An irreducible finite Markov chain is *recurrent*.

---

23 Invariant Distributions

Consider a Markov chain with TPM P on Ω:

- A distribution π over Ω is called an *invariant distribution* (or *stationary distribution*) if $\pi P = \pi$.

- *Invariant distribution* is NOT necessarily *existent* and *unique*.

- Under certain condition (*ergodicity*), there exists a unique invariant distribution π. In such cases, π is often called an *equilibrium distribution*.

---

24 Positive Recurrence

- The *expected return time* of a state x is defined to be $m_x \triangleq E[T_x]$.
- When x is transient, $m_x = \infty$. If x is recurrent, m_x is NOT necessarily finite.
- A *recurrent state* x is called *positive recurrent* if $m_x < \infty$. Otherwise, it is called *null recurrent*.

. . .

- *(Existence of Invariant distributions)* For an *irreducible* Markov chain, if some state is *positive recurrent*, then all states are *positive recurrent* and the chain has an *invariance distribution* π given by $\pi(x) = 1/m_x$.

---

25 Ergodic Markov Chains

- An *irreducible*, *aperiodic*, and *positive recurrent* Markov chain is called an *ergodic Markov chain*, or simply *ergodic chain*.
- A finite Markov chain is *ergodic* if and only if it is irreducible and aperiodic.
- A Markov chain is *ergodic* if it is aperiodic and there exist N such that any state can be reached from any other state within N steps with positive probability.

. . .

- *(Convergence to equilibrium)* Let P be the transition probability matrix of an *ergodic* Markov chain, then there exists a unique *invariant distribution* π. Then with any initial distribution, $\Pr(X_n = x) \to \pi(x)$ as $n \to \infty$ for all $x \in \Omega$. Particularly, $P^n(x', x) \to \pi(x)$ as $n \to \infty$ for all $x, y \in \Omega$.

---

26 Ergodic Theorem

- The *ergodic theorem* relates *time mean* to *space mean*:
- Let $Markov(\pi_0, P)$ be an *ergodic Markov chain* over Ω with equilibrium distribution π, and f be a measurable function on Ω, then

$$\frac{1}{n}\sum_{t=0}^{n} f(X_t) \xrightarrow{a.s.} E_\pi[f], \quad \text{as } n \to \infty.$$

  More generally, we have for any positive integer m:

$$\frac{1}{n}\sum_{t=0}^{n} f(X_{\tau+tm}) \xrightarrow{a.s.} E_\pi[f], \quad \text{as } n \to \infty.$$

- The *ergodic theorem* is the theoretical foundation for MCMC.

---

27 Mixing

- Let μ and ν be probability measures over a measurable space $(\Omega, \mathcal{S})$, then the *total variation distance* between them is defined as

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{S}} |\mu(S) - \nu(S)| .$$

  If Ω is *countable*, we have

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| .$$

- The *total variation distance* is a metric.

. . .

- The time required by a Markov chain to get close to the equilibrium distribution is measured by the *mixing time*, defined as $t_{mix}(\epsilon) = \inf\{t : d(t) \leq \epsilon\}$, and in particular $t_{mix} \triangleq t_{mix}(1/4)$.

---

28 Bounds of Mixing Time

There are various ways to bound the mixing time, taking into account different factors:

- Counting Bound
- Diameter Bound
- Spectral Analysis
- Conductance

---

29 Simple Bounds on Mixing Time

- *(Counting Bound)*: if the chain can only transit to a limited number of states from each state, it may take quite a long time to cover the explore the entire space. Consider an ergodic finite Markov chain over Ω whose equilibrium distribution is uniform, then

$$t_{mix}(\epsilon) \geq \frac{\log(|\Omega|(1-\epsilon))}{\log \Delta}.$$

  Here, Δ is the maximum outgoing degree of each state.

. . .

- *(Diameter Bound)*: For an ergodic finite Markov chain, we have $t_{mix}(\epsilon) \geq L/2$ for any $\epsilon < 1/2$, where L is the *diameter* of the state graph.

---

30 Spectral Representation

Let P be a *stochastic matrix* over a finite space Ω with $|\Omega| = N$:

- The *spectral radius* of P, namely the maximum absolute value of all eigenvalues, is $\rho(P) = 1$.

. . .

Furthermore, if P is *ergodic* and *reversible* with equilibrium distribution π:

- Consider an inner product space $\left(\mathbb{R}^{\Omega}, \langle \cdot, \cdot \rangle_{\pi}\right)$ with

$$\langle f, g \rangle_{\pi} \triangleq E_{\pi}[f(x)g(x)] = \sum_{x \in \Omega} \pi(x) f(x) g(x).$$

- All eigenvalues of P are real values, given by $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$. Let f_i be the right eigenvector associated with λ_i. Then the left eigenvector is $\pi_i \circ f_i$ *(element-wise product), and P^t can be represented as

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{i=1}^{N} f_i(x) f_i(y) \lambda_i^t = 1 + \sum_{i=2}^{N} f_i(x) f_i(y) \lambda_i^t.$$

---

31 Spectral Gap and Relaxation Time

- Let $\lambda_* \triangleq \sup\{|\lambda_i| : 2 \leq i \leq |\Omega|\}$. Then the *spectral gap* is defined to be $\gamma \triangleq 1 - \lambda_2$ and the *absolute spectral gap* is defined to be $\gamma_* \triangleq 1 - \lambda_*$. Then:

$$\left| \frac{P^t(x,y)}{\pi(y)} - 1 \right| \leq \frac{\lambda_*^t}{\sqrt{\pi(x)\pi(y)}} \leq \frac{\lambda_*^t}{\pi_{min}} \leq \frac{e^{-\gamma_* t}}{\pi_{min}}.$$

  Here, $\pi_{min} = \min_{x \in \Omega} \pi(x)$.

- The *relaxation time* of a *Markov chain* is defined to be $t_{rel} = 1/\gamma_*$, then

$$\log\left(\frac{1}{2\epsilon}\right)(t_{rel} - 1) \leq t_{mix}(\epsilon) \leq \log\left(\frac{1}{\epsilon \pi_{min}}\right) t_{rel}.$$

- Generally, the goal to design a *rapidly mixing* reversible Markov chain is to minimize λ_*, or in other words, maximize the γ_*.

---

32 Conductance

Consider an ergodic Markov chain on a finite space Ω with transition probability matrix P and equilibrium distribution π:

- The *ergodic flow* from a subset A to another subset B is defined as

$$Q(A, B) \triangleq \sum_{x \in A, y \in B} \pi(x) P(x, y).$$

- The *conductance* of a Markov chain is defined as

$$\Phi_* = \inf_{S:\pi(S)\leq\frac{1}{2}} \frac{Q(S, S^c)}{\pi(S)}.$$

. . .

- *(Jerrum and Sinclair (1989))* The spectral gap is bounded by

$$\frac{1}{2}\Phi_*^2 \leq \gamma \leq 2\Phi_*.$$

---

33 Exercise 1

Consider an ergodic finite chain P with $\gamma_* < \gamma$. To improve the mixing time, one can add a little bit lazyness as $P' = (1-\alpha)P + \alpha I$. Please solve the optimal value of α that maximizes the *absolute spectral gap* γ_*.

---

34 Exercise 2

Consider a 2×2 stochastic matrix P, given by $P(x, y) = 1 - p$ when $x \neq y$.

- Please specify the condition under which P is ergodic.
- What is the equilibrium distribution when P is ergodic?

. . .

- Solve the optimal value of p that maximizes the absolute spectral gap.

---

35 Exercise 3

- Consider a random walk on a circle of length n, where n is even. At each state x, it walks to either of its neighbor with probability p, and stays at x with probability $1 - 2p$.
- Please specify the condition under which P is ergodic.
- What is the equilibrium distribution when P is ergodic?

. . .

- Solve the optimal value of p that maximizes the *conductance*.

. . .

- If we allow the chain to jump from x to its *opposite* state with probability q and thus the probability of staying at x is $1 - 2p - q$. What's the conductance now?
- Solve the optimal setting of p and q.

---

36 General Markov Chains

Next, we extend the formulation of *Markov chain* from *countable space* to general *measurable space*.

- First, the *Markov property* remains.

. . .

- Generally, a *homogeneous Markov chain* over a measurable space $(\Omega, \mathcal{S})$ is characterized by an *initial measure* π_0 and a *transition probability kernel* $P : \Omega \times \mathcal{S} \to [0, 1]$, denoted by $Markov(\pi_0, P)$. Here, Ω is called the *state space*; and:

$$\Pr(X_{t+1} \in A | X_t = x) = P(x, A).$$

- P must be a *stochastic kernel*:
  - Given $x \in \Omega$, $P_x : A \mapsto P(x, A)$ is a *probability measure* over $(\Omega, \mathcal{S})$.
  - Given a measurable subset $A \in \mathcal{S}$, $P(\cdot, A) : x \mapsto P(x, A)$ is a measurable function.
- When Ω is a countable space, P reduces to a *stochastic matrix*.

---

37 General Markov Chains (cont'd)

- Suppose the distribution of X_t is π_t, then

$$\pi_{t+1}(A) = \int_\Omega P(x, A)\pi_t(dx)$$

  Again, we can simply write this as $\pi_{t+1} = \pi_t P$.

- Composition of stochastic kernels P and Q remains a stochastic kernel, denoted by $Q \circ P$, which is given by:

$$(Q \circ P)(x, A) = \int_\Omega P(x, dy)Q(y, A).$$

- Recursive composition of P for m times results in a stochastic kernel denoted by P^m, and we have $P^{n+m} = P^n \circ P^m$ and $\mu_{t+m} = \mu_t P^m$.

---

38 Example: Random Walk in $\mathbb{R}^n$

$$X_{t+1} = X_t + B_t, \quad \text{with } B_t \sim \mathcal{N}(0, \sigma^2 I).$$

- For this case, the stochastic kernel is given by $P_x = P_{\mathcal{N}(x, \sigma^2 I)}$.

---

39 Occupation Time, Return Time, and Hitting Time

For general space, talking about the probability of *hitting a point* makes no sense. Instead, we should talk about sets:

. . .

Given a Markov train (X_t) over $(\Omega, \mathcal{S})$, and $A \in \mathcal{S}$:

- The *occupation time* of A is defined to be $\eta_A \triangleq \sum_{t=1}^{\infty} 1(X_t \in A)$.
- The *return time* of A is defined to be $\tau_A \triangleq \inf\{t \geq 1 : X_t \in A\}$.
- The *hitting time* of A is defined to be $\sigma_A \triangleq \inf\{t \geq 0 : X_t \in A\}$.
- η_A, τ_A and σ_A are all random variables.

---

40 φ-irreducibility

- Define $L(x, A) \triangleq P_x(\tau_A < \infty) = P_x(X \text{ ever enters } A)$.
- $L(x, A)$ has

$$L(x, A) = P(x, A) + \int_{A^c} P(x, dy) L(y, A).$$

- Given a positive measure φ over $(\Omega, \mathcal{S})$, a markov chain is called *φ-irreducible* if $L(x, A) > 0 \ \forall x \in \Omega$ whenever A is *φ-positive*, *i.e.* $\varphi(A) > 0$.
- Intuitively, it means that for any φ-positive set A, there is positive chance that the chain enters A within finite time, no matter where it begins.

---

41 φ-irreducibility (cont'd)

- A Markov chain over Ω is φ-irreducible if and only if either of the following statement holds:
  - $\varphi(A) > 0 \Rightarrow \forall x \in \Omega,\ E_x(\eta_A) > 0$
  - $\varphi(A) > 0 \Rightarrow \forall x \in \Omega,\ \exists t \in \mathbb{N}^+,\ P^t(x, A) > 0$

. . .

- Typical spaces usually come with *natural measure* φ:
  - The *natural measure* for countable space is the *counting measure*. In this case, the notion of *varphi*-irreducibility coincides with the one introduced earlier.
  - The *natural measure* for $\mathbb{R}$, $\mathbb{R}^n$, or a finite-dimensional manifold is the *Lebesgue measure*
  - When φ is implicitly indicated, we simple call the chain *irreducible*.

---

42 φ-irreducibility (cont'd)

The following theorem provides an easier way to verify *irreducibility* for separable space.

- A set S is called *φ-communicating* if for every $x \in S$ and every φ-positive subset $A \subset S$ (*i.e.* A is measurable and $\varphi(A) > 0$), $x \to A$.
  - The space Ω is *φ-irreducible* if and only if Ω is *φ-communicating.*

. . .

- *(Irreducibility Theorem)* Let P be a *stochastic kernel* over a *measurable* state space $(\Omega, \mathcal{S})$, then P is *μ-irreducible* if the following conditions are satisfied:
  - Ω is *separable* (meaning Ω has a countable dense set) and *connected*;
  - Every non-empty open subset A is φ-positive;
  - Every $x \in \Omega$ has a φ-communicating neighborhood.

. . .

- **Note:** This irreducibility theorem does *NOT* apply to countable space. Why?

---

43 Transience and Recurrence

Given a Markov chain (X_t) over (Ω, S), and $A \in \mathcal{S}$:

- A is called *transient* if $E_x[\eta_A] < \infty$ for every $x \in A$.
- A is called *uniformly transient* if there exists $M > 0$ such that $U(x, A) < M$ for every $x \in A$.
- A is called *recurrent* if $E_x[\eta_A] = \infty$ for every $x \in A$.

. . .

- Consider an φ-irreducible chain, then either:
  - Every φ-positive subset is recurrent, then we call the chain *recurrent*
  - Ω is covered by countably many uniformly transient sets, then we call the chain *transient.*

---

44 Harris Recurrence

- A set A is called *Harris recurrent* if

$$P_x(\eta_A = \infty) = 1, \ \forall x \in A,$$

  which means any chain starts within A visits A infinitely often.

- A Markov chain is called *Harris recurrent* if it is φ-irreducible (with maximal irreducibility measure) and every φ-positive subset is Harris recurrent.

- Most MCMC samplers are *Harris recurrent*.

. . .

- *Harris recurrence* implies *recurrence*.
  - Note: $P_x(\eta_A = \infty) = 1 \Rightarrow E_x[\eta_A] = \infty$, but the converse is generally not true.

---

45 Invariant Measures

- A measure π is called an *invariant measure w.r.t.* the stochastic kernel P if $\pi = \pi P$, *i.e.*

$$\pi(A) = \int_\Omega \pi(dx) P(x, A), \ \forall A \in \mathcal{S}.$$

- A *recurrent* Markov chain admits a unique *invariant measure* π (up to a scale constant).
  - **Note:** This measure π can be finite or infinite.

---

46 Positive Chains

- A Markov chain is called *positive* if it is irreducible and admits an *invariant probability measure* π.
- If a Markov chain is *positive* then it is *recurrent*, and thus it admits a *unique* invariant probability measure.
- If a Markov chain is *Harris positive* and *recurrent*, then it is called a *positive Harris chain*.

. . .

- The study of the *existence* of π requires more sophisticated analysis that involves *petite sets*, *sub-invariance*, and *atoms*.
  - We are not going into these details, as in MCMC practice, existence of π is usually not an issue.

---

47 Subsampled Chains

Suppose P is a stochastic kernel and q is a probability vector over $\mathbb{N}$.

- Then P_q defined as below is also a stochastic kernel:

$$P_q(x, A) = \sum_{k=0}^{\infty} q_k P^k(x, A)$$

  The chain with kernel P_q is called a *subsampled chain* with q.
- When $q_n = 1(n = m)$, $P_q = P^m$.
- If π is invariant *w.r.t.* P, then π is also invariant *w.r.t.* P_q.

---

48 Birkhoff Ergodic Theorem

- A stochastic process (X_t) is called a *stationary process* if

$$P(X_{t_1+\tau}, \ldots, X_{t_k+\tau}) = P(X_{t_1}, \ldots, X_{t_k})$$

- A Markov chain is *stationary* if it has an invariant probability measure and that is also its initial distribution.

- *(Birkhoff Ergodic Theorem)* Every *irreducible stationary* Markov chain (X_t) is *ergodic*, that is, for any real-valued measurable function f:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} E_\pi[f(X)], \quad \text{as } n \to \infty.$$

  where π is the invariant probability measure.

- If $Markov(\pi, P)$ is a stationary Markov chain, then $Markov(\pi, P_q)$ is also a stationary Markov chain.

---

49 Convergence of Measures

- For a *positive Harris chain* $Markov(\pi_0, P)$ with invariant probability measure π_*, and $\pi_t = \pi_0 P^t$, we have

$$\|\pi_t - \pi_*\|_{TV} \to 0, \quad \text{as } t \to \infty.$$

- As an immediate corollary, we have

$$\|P^t(x, \cdot) - \pi_*\|_{TV} \to 0, \quad \text{as } t \to \infty.$$

. . .

- The chain is called *geometric ergodic* if there exist a nonnegative function M and $\rho \in (0, 1)$ such that

$$\|P^t(x, \cdot) - \pi_*\| \leq M(x)\rho^t, \ \forall x \in \Omega.$$

---

50 Markov Chain Monte Carlo

(Markov Chain Monte Carlo): To sample from a *target distribution* π:

1. We first construct a Markov chain with *transition probability kernel* P such that $\pi P = \pi$.
2. Then we simulate the chain, usually in two stages:
   - *(Burning stage)* simulate the chain and ignore all samples, until it gets close enough to the *equilibrium distribution*

- *(Sampling stage)* collect samples $x_1, \ldots, x_n$ from a subsampled chain P^m or P_q.

3. Approximate the expectation of the function f of interest using the sample mean, as

$$E_\pi[f] \simeq \frac{1}{n} f(x_i).$$

---

51 Detailed Balance and Reversible Chains

Most Markov chains in MCMC practice falls in a special family: *reversible chains*

- A distribution π over a *countable space* is said to be *in detailed balance* with P if $\pi(x)P(x,y) = \pi(y)P(y,x)$.
  - *Detailed balance* implies *invariance*.
  - The converse is not true.

. . .

- An irreducible Markov chain with TPM P and an invariant distribution π is called *reversible* if π is in *detailed balance* with P.
- Consider an irreducible Markov chain (X_t) $Markov(\pi, P)$ where π is in *detailed balance* with P, its *reversal* $(\hat{X}_t)$ is given by $(\pi, \hat{P})$ with $\hat{P}(x,y) \triangleq \pi(y)P(y,x)/\pi(x)$. Then

$$\Pr(X_0 = x_0, \ldots, X_n = x_n) = \Pr(\hat{X}_0 = x_n, \ldots, \hat{X}_n = x_0).$$

---

52 Reversible Chains on General Spaces

Over a general measurable space $(\Omega, \mathcal{S})$:

- A stochastic kernel P is called *reversible w.r.t.* a probability measure π if

$$\int\int f(x,y)\pi(dx)P(x,dy) = \int\int f(y,x)\pi(dx)P(x,dy)$$

  for any bounded measurable function $f : \Omega \times \Omega \to \mathbb{R}$.

- Suppose both π and P_x are absolutely continuous *w.r.t.* a base measure μ, that is, $\pi(dx) = \pi(x)\mu(dx)$ and $P(x,dy) = P_x(dy) = p_x(y)\mu(dy)$, then the chain is *reversible* if and only if

$$\pi(x)p_x(y) = \pi(y)p_y(x), \ a.e.$$

  which is called the *detailed balance*.

- If P is *reversible w.r.t.* π, then π is an *invariant* to P.

---

53 Metropolis-Hastings: An Overview

- In MCMC practice, the *target distribution* π is usually known up to an *unnormalized density* h, such that $\pi(x) = h(x)/c$, and the *normalizing constant* c is often intractable to compute.

. . .

- The *Metropolis-Hastings algorithm (M-H algorithm)* is a classical and popular approach to MCMC sampling. It works as follows:
  1. It is associated with a *proposal kernel* Q.
  2. At each iteration, a *candidate* is generated from Q_x given the current state x.
  3. With a certain acceptance ratio, which depends on both Q and h, the *candidate* is *accepted*.
- The acceptance ratio is determined in a way that maintains *detailed balance*, so the resultant chain is *reversible w.r.t.* π.

. . .

- Many sampling algorithms, notably *Gibbs sampling* and *Slice sampling*, are special cases of the M-H algorithm.

---

54 Metropolis Algorithm

- The *Metropolis algorithm* is a precursor (and a special case) of the M-H algorithm, which requires the designer to provide a *symmetric kernel* Q, *i.e.* $q_x(y) = q_y(x)$, where q_x is the density of Q_x.
  - **Note:** π is not necessarily invariant to Q
  - *Gaussian random walk* is a symmetric kernel.

. . .

- At each iteration, with current state x:
  1. Generate a candidate y from Q_x
  2. Accept the candidate with *acceptance ratio* $a(x, y) = \min\{h(y)/h(x), 1\}$.

. . .

- The *Metropolis update* satisfies *detailed balance*. Why?

---

55 Metropolis-Hastings Algorithm

- The *Metropolis-Hastings algorithm* requires a *proposal kernel* Q,
  - Q is NOT necessarily *symmetric* and does NOT necessarily admit π as an invariant measure.

. . .

- At each iteration, with current status x:
  - Generate a candidate y from Q_x
  - Accept the candidate with *acceptance ratio* $a(x, y) = \min\{r(x, y), 1\}$, with

$$r(x, y) = \frac{h(y)q_y(x)}{h(x)q_x(y)}.$$

. . .

- The *Metropolis-Hastings update* satisfies *detailed balance*. Why?

---

56 Gibbs Sampling

- The *Gibbs sampler* was introduced by Geman and Geman (1984) from sampling from a Markov random field over images, and popularized by Gelfand and Smith (1990).
- Each state x is comprised of multiple components $(x^{(1)}, \ldots, x^{(n)})$.
- At each iteration, following a permutation σ over $(1, \ldots, n)$:
  - For $i = 1, \ldots, n$, let $j = \sigma(i)$, update x by re-drawing $x^{(j)}$ conditioned on all other components:

$$x^{(j)} \sim \pi_{/j}\left(\cdot | x^{(1)}, \ldots, x^{(j-1)}, x^{(j+1)}, \ldots, x^{(n)}\right).$$

  - At each iteration, one can use either a *random scan* or a *fixed scan*.
  - Different schedules can be used at different iterations to scan the components.
- The *Gibbs update* is a special case of *M-H update*, and thus satisfies *detailed-balance*. Why?

---

57 Combination of MCMC Kernels

Let $K_1, \ldots, K_m$ be *stochastic kernels* with the same invariant probability measure π:

- *(Mixture of kernels)*: Let q be a probability vector, then $K := \sum_{i=1}^{m} q_k K_k$ remains a *stochastic kernel* with invariant probability measure π.
  - Furthermore, if $K_1, \ldots, K_m$ are all reversible, then K is reversible.

. . .

- *(Composition of kernels)*: $K = K_m \circ \cdots \circ K_1$ is also a *stochastic kernel* with invariant probability measure π.
  - **Note:** K is generally not *reversible* even when $K_1, \ldots, K_m$ are all reversible, except when $K_1 = \cdots = K_m$.