6 Markov Chain Monte Carlo (MCMC)

The underlying idea in MCMC is to replace the iid samples of basic MC methods, with dependent samples from an ergodic Markov chain, whose limiting (stationary) distribution is the target distribution. This often provides essential improvements in sampling efficiency, especially for multidimensional distributions. Furthermore, MCMC provides a useful tool for global optimization. The most prominent MCMC algorithms are the Metropolis-Hastings algorithms and the Gibbs sampler, which we describe next, after a short reminder on Markov chains.

6.1 Markov Chain Basics

MCMC applies both to discrete and continuous state spaces. For simplicity we will consider in this lecture only the former. The theory of continuously-valued Markov chains is more involved technically, and we leave its consideration to the references (e.g., R&C'2004).

Finite State Space: Recall that a (discrete-time, time-homogeneous) Markov chain $(X_t, t = 0, 1, \dots)$ over a finite state space $\mathcal{X}$ is defined in terms of the transition matrix $P = (p_{ij})_{i,j\in\mathcal{X}}$, and satisfies the Markov property

$$\mathbb{P}(X_{t+1} = j | X_t = i, X_{t-1}, \dots, X_0) = \mathbb{P}(X_{t+1} = j | X_t = i) = p_{ij}$$

The p_{ij}'s are the transition probabilities, which satisfy

$$p_{ij} \geq 0, \quad \sum_j p_{ij} = 1 \;\; \forall i$$

(i.e., P is a *stochastic matrix*). The distribution of the process (X_t) is fully defined once the initial distribution $\pi^{(0)}$ of X_0 is specified.

Evolution equations: Let $\pi^{(t)}$ be the row vector that described the marginal distribution of X_t, namely $\pi^{(t)} = (\pi_i^{(t)})_{i\in\mathcal{X}}$ and $\pi_i^{(t)} = \mathbb{P}(X_t = i)$. Then

$$\pi^{(t)} = \pi^{(0)} P^t \, .$$

Similarly, the t-step transition probabilities are given by

$$\mathbb{P}(X_t = j | X_0 = i) = (P^t)_{ij} \, ,$$

Recurrence: Let $T_j = \inf(t \geq 1 : X_t = j)$ be the first time that the chain hits state i. State i is *recurrent* if $\mathbb{P}(T_j < \infty | X_0 = j) = 1$, i.e., the process returns to state j if started there with probability 1. Otherwise the state i is *transient.*

Periodicity: The period d_i of a state i is smallest integer $d \geq 1$ such that

$$\mathbb{P}(X_t = i \text{ for some } t \notin \{nd, n \geq 0\} | X_0 = i) = 0.$$

If $d_i = 1$ the state is *a-periodic.* The period of any two states that communicate is the same.

Irreducibility: State j is *accessible* from state i (written as $i \to j$) if $(P^t)_{ij} > 0$ for some $t \geq 1$. States i and j *communicate* if $i \to j$ and $j \to i$. The chain is *irreducible* if all states communicate with each other.

Ergodicity: A state is *ergodic* if it is (positive) recurrent and a-periodic. A Markov chain is said to be ergodic if all its states are ergodic.

In an irreducible finite Markov chain, all states are recurrent. Also, the periodicity and ergodicity are global properties: That is, all states share the same period (1 in the ergodic case).

Stationary distribution: A probability vector π (i.e., $\pi_i \geq 0$, $\sum_i \pi_i = 1$) is called a stationary distribution (or an invariant measure) of the chain if

$$\pi = \pi P\,.$$

More explicitly, this gives the system of equations

$$\pi_j = \sum_i \pi_i p_{ij}, \quad j \in \mathcal{X}\,.$$

These are the *global balance equations.*

A finite irreducible chain always has a stationary distribution, which is unique.

Reversible Markov chain: The chain is *reversible* if there exists a probability vector π such

$$\pi p_{ij} = \pi_j p_{ji} \quad \forall i, j\,.$$

This set of equations for π is known as the *detailed balanced equations.* By simple summation it can be verified that $\pi = P\pi$, i.e., π is a stationary distribution of the chain.

Convergence: Consider an ergodic (irreducible and a-periodic), finite Markov chain. The limit

$$\lim_{t\to\infty} \mathbb{P}(X_t = j | X_0 = i) \equiv \lim_{t\to\infty} P_{ij}^t \overset{\triangle}{=} \pi_j^\infty$$

exists for all i, j, is independent of the initial state i, and π^∞ equals the the unique stationary distribution of the chain.

Convergence Rate: The rate of convergence of the marginal distributions $\pi^{(t)}$ to the stationary one is of central importance in MCMC. At this point, we only mention that this rate essentially depends on the size $|\lambda_2|$ of the second-largest eigenvalue of P (the largest eigenvalue is always 1). Estimating λ_2 for large chains is a deep topic, which we do not consider at the moment.

Ergodic Theorem: For a finite irreducible chain (X_t) with stationary distribution π,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} h(X_t) = E_pi(h) \overset{\triangle}{=} \sum_{i\in\mathcal{X}} \pi_i h(i)$$

with probability 1.

Central Limit Theorem: For a finite irreducible chain (X_t) with stationary distribution π,

$$\frac{1}{\sqrt{N}} \left(\sum_{t=1}^{n} (h(X_t) - E_\pi(h) \right) \Rightarrow_{distr.} N(0, \sigma_h^2)$$

with $\sigma_h = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k < \infty$, $\gamma_k = \mathrm{Cov}_\pi(h(X_0), h(X_k))$.

Countable State Space: We briefly mention the differences from the finite case when the state space is countably infinite.

Recurrence: A recurrent state i is *positive recurrent* if the return time has finite mean, $\mathbb{E}(T_i | X_0 = i) < \infty$, and is otherwise *null recurrent.*
In the finite case all recurrent states are positive recurrent.

Ergodicity: A state is *ergodic* if it is *positive* recurrent and a-periodic. A Markov chain is said to be ergodic if all its states are ergodic.

In an irreducible Markov chain, the properties of positive recurrence, null recurrence, periodicity and ergodicity and global, i.e., shared by all states.

An irreducible chain has a stationary distribution *if and only if* all of its states are *positive* recurrent In that case, the stationary distribution is unique.

Convergence: Consider an irreducible, a-periodic Markov chain. Then the limit

$$\lim_{t\to\infty} \mathbb{P}(X_t = j | X_0 = i) \equiv \lim_{t\to\infty} P_{ij}^t \stackrel{\triangle}{=} \pi_j^\infty$$

exists for all i, j, and is independent of the initial state i. We note that $0 \leq \pi_j^\infty < 1$, but the sum $\sum_j \pi_j^\infty$ may be less than 1 in general. Also, $\pi_j^\infty = 0$ iff i is null recurrent.

The following conditions are equivalent:

(1) All states are positive recurrent.

(2) The limiting vector π^∞ is a probability distribution.

(3) There exists a stationary distribution π.

If these hold, the stationary distribution π is unique, and coincides with π^∞.

MCMC: We can now describe more precisely the basic idea of MCMC. To sample from a target distribution $f(x)$, we devise a Markov chain (X_t) whose stationary (and limiting) distribution is f. Then, after an initial period, the marginal distribution of X_t will be close to f. A Monte-Carlo estimate for $E_f H(X)$, for example, can be formed by

$$\hat{\ell}_N = \frac{1}{N} \sum_{t=k+1}^{N+k} X_t \,.$$

It should be observed that the samples are not independent any more, and the size N of the sample should be large relative to the so called *mixing time* of the chain.

6.2 The Metropolis-Hastings Algorithm

Suppose we wish to generate an RV X with target distribution f. Let $q(x,y) \equiv q(y|x)$ be the transition function of a Markov chain over the same state space $\mathcal{X}$ as that of X. In the following we treat x as discrete, although the algorithm is valid for continuous state as well. Thus, $q(x,y)$ is a transition matrix of the chain. Starting with some initial state X_0, the algorithm proceeds as follows for $t \geq 0$:

1. Given the current state X_t, generate $Y \sim q(X_t, \cdot)$.
2. Set
$$X_{t+1} = \begin{cases} Y & \text{with probability } \alpha(X_t, Y) \\ X_t & \text{otherwise} \end{cases}$$
where
$$\alpha(x,y) = \min\{\rho(x,y), 1\}, \quad \rho(x,y) = \frac{f(y)q(y,x)}{f(x)q(x,y)}$$

Note that the target distribution f needs to be known only up to a normalization constant.

This algorithm induces a stationary Markov chain with transition function

$$p(x,y) = \begin{cases} q(x,y)\alpha(x,y) & \text{if } x \neq y \\ 1 - \sum_{z \neq x} q(x,z)\alpha(x,z) & \text{if } x = y \end{cases}$$

It is now easy to verify that the following holds:

$$f(x)p(x,y) = f(y)p(y,x), \quad x, y \in \mathcal{X}$$

These can be seen to be the detailed balance equations for the Markov chain with transition matrix $p(x,y)$, which imply that f is a stationary distribution for that chain.

Now, if $q(x,y)$ is chosen so that the chain is irreducible and aperiodic, then f is its limiting distribution. This trivially holds if $q(x,y) > 0$ for all x, y, but the latter is certainly not a necessary condition.

The transition function $q(x,y)$ is called the proposal distribution/function/kernel, and $\alpha(x,y)$ the acceptance probability. The original Metropolis algorithm was suggested in 1953 with symmetric proposal functions, $f(x,y) = f(y,x)$, and extended by Hastings (1970) to the non-symmetric case.

Since the algorithm is of the acceptance-rejection type, its efficiency is related to the acceptance probability $\alpha(x, y)$. To obtain $\rho(x, y)$ close to 1, one would ideally like $q(x, y)$ to be close to target distribution $f(y)$ (for all x). Consider, for illustration, some particular choices for q.

1. Crude MC: Let $q(x, y) = f(y)$. Then $\alpha \equiv 1$, and the scheme reduces to iid sampling from f.
2. Let $q(x, y) = g(y)$, independent of x. Then, at each stage, Y is sampled from g, and
$$\alpha(x, y) = \min\{\frac{f(y)g(x)}{f(x)g(y)}, 1\} .$$
The scheme is similar to the acceptance-rejection method from lecture 3. However, here the samples are not independent since α depends on the previous sample X_t.

Some general guidelines for choosing the proposal distribution $q(x, y)$ are summarized below:

1. The induced Markov chain should be irreducible, with short mixing time, to allow full coverage of the state space $\mathcal{X}$.
2. Low correlation between adjacent samples X_t and X_{t+1}. This can be partitioned into the following two requirements, which unfortunately are *contradictory.*
   a. Low correlation between X_{t+1} and Y.
   b. LArge acceptance probability α.

We proceed with some more useful examples.

Uniform Sampling: Suppose we wish to sample uniformly from a discrete set S. A general template is as follows. Define a neighborhood structure on S, namely, for each $x \in S$ define its neighbors N_x, so that each state can be reached from any other by moving through neighbors only. The proposal distribution $q(x, y)$ simply chooses each neighbor of x with equal probability, namely $q(x, y) = 1/n_x$ for $y \in N_x$, where $n_x = |N_x|$. Since the target distribution f is constant here, the acceptance probability is

$$\alpha(x, y) = \min\{\frac{n_x}{n_y}, 1\} .$$

The Random Walk Sampler: Suppose that, for a given x, Y is obtained by $Y = x + Z$, where Z and RV with symmetric distribution around 0 (e.g., Gaussian with mean 0 in the continuous case). Noting that the proposal function $q(x, y)$ is symmetric in this case, we obtain

$$\alpha(x, y) = \min\{\frac{f(y)}{f(x)}, 1\},$$

similarly to the original Metropolis algorithm.

The Hit-And-Run Sampler: The basic version of this sampler is useful for sampling (possibly uniformly) from an open set S in $\mathbb{R}^m$. From a given point x, draw a random direction d in $\mathbb{R}^m$, and choose Y randomly and uniformly from the intersection of the line segment $L = \{x + \lambda d, \lambda \in \mathbb{R}\}$ with S (e.g., using rejection sampling). Noting that $q(x, y)$ is again symmetric, the acceptance probability $\alpha(x, y)$ is the same as in RW sampler above, and in particular it equals 1 for sampling uniformly from S.

The hit-and-run sampler allows to reach across the entire set in one step, and has many applications along with provably excellent performance (in terms of convergence times) for sampling from convex bodies.

6.3 The Gibbs Sampler

The Gibbs Sampler, introduced by Geman and Geman (1984) in the context of random fields, is particularly useful for sampling from multivariate distributions. The idea is to sample sequentially from *conditional* distributions. The algorithm is useful if sampling from the conditional distributions is simpler than sampling from the joint distribution.

The algorithm is simple. We want to sample from $f(X) = f(x_1, \ldots, x_n)$. Let $f_i(x_i|x_{1:i-1}, x_{i+1:n})$ denote the conditional distribution of x_i given the other components. We start with an initial vector X_0. Then, at each stage $t \geq 0$, we start with $X_t = (x_1, \ldots, x_n)$ and compute $X_{t+1} = (y_1, \ldots, y_n)$, using one of the following procedures.

Systematic Gibbs Sampler:

- Start with $X_t = (x_1, \ldots, x_n)$
- For $i = 1$ to n, draw y_i from $f_i(\cdot|y_{1:i-1}, x_{i+1:n})$
- Set $X_{t+1} = (y_1, \ldots, y_n)$

Random Sweep Gibbs Sampler:

- Start with $X_t = (x_1, \ldots, x_n)$
- Generate a random permutation σ of $\{1, \ldots, n\}$

- For $i = 1$ to n, draw $y_{\sigma(i)}$ from $f_{\sigma(i)}(\cdot | y_{\sigma(1):\sigma(i-1)}, x_{\sigma(i+1):\sigma(n)})$
- Set $X_{t+1} = (y_1, \ldots, y_n)$

In either case $f(X)$ is a stationary distribution of the resulting Markov chain (X_t). The random sweep sampler has the additional property that the chain is reversible, which has certain benefits in the analysis. In the following we shall focus for concreteness on the basic (systematic) scheme. We also restrict the discussion to the case of a *finite* state space.

Lemma 6.1 *$f(X)$ is a stationary distribution of the resulting Markov chain (X_t).*

Proof: We need to show that

$$\sum_{x_{1:n}} f(x_{1:n}) P(y_{1:n}|x_{1:n}) = f(y_{1:n})$$

for every $y_{1:n}$, where P is the transition matrix implied by the (systematic) Gibbs sampler. Now,

$$\begin{aligned}
&\sum_{x_{1:n}} f(x_{1:n}) P(y_{1:n}|x_{1:n}) \\
&= \sum_{x_{1:n}} f(x_{1:n}) f_1(y_1|x_{2:n}) f_2(y_2|y_1, x_{3:n}) \cdots f_n(y_n|y_{2:n}) \\
&= \sum_{x_{2:n}} f(x_{2:n}) f_1(y_1|x_{2:n}) f_2(y_2|y_1, x_{3:n}) \cdots f_n(y_n|y_{2:n}) \\
&= \sum_{x_{2:n}} f(y_1, x_{2:n}) f_2(y_2|y_1, x_{3:n}) \cdots f_n(y_n|y_{2:n}) \\
&= \sum_{x_{3:n}} f(y_1, x_{3:n}) f_2(y_2|y_1, x_{3:n}) \cdots f_n(y_n|y_{2:n}) \\
&= \sum_{x_{3:n}} f(y_1, y_2, x_{3:n}) f_3(y_3|y_{1:2}, x_{4:n}) \cdots f_n(y_n|y_{2:n}) \\
&= \cdots = f(y_{1:n})
\end{aligned}$$

□

Consequently, if the Markov chain is ergodic (irreducible and aperiodic), then $f(X)$ is its limiting distribution. A simple sufficient condition for both is the following *positivity condition*: $f(x_1, \ldots, x_n) > 0$ whenever $f_i(x_i) > 0$ for all i.

Illustration: Apply the Gibbs samples to f uniform on
(i) $\mathcal{X} = \{(0,1), (1,0)\}$, *(ii)* $\mathcal{X} = \{(0,1), (1,0), (1,1)\}$

6.4 **Examples** (in class)

M-H:

- Bayesian Probit Model (from: Johanson and Evers, Lecture Notes on Monte Carlo Methods, 2010, Example 5.2).

Gibbs:

1. Auto-exponential model
2. Ising model
3. Poisson change-point model (Ibid., Example 4.1)
4. Data augmentation - mixture of Gaussians (Ibid., Section 4.5)

6.5 Simulated Annealing

Simulated Annealing is a powerful heuristic for global optimization. It was introduced by Kirkpatrick, Gelatt and Vecci (1983), and has since found numerous applications.

The basic method is closely related to MCMC sampling, and the term MCMC-optimization is often used synonymously.

Suppose we wish to minimize a function $C(x)$, $x \in \mathcal{X}$, where $\mathcal{X}$ may be a discrete or continuous multi-dimensional space. It is *not* assumed that the function C belongs to a specific "easy" class such as convex functions. In order to apply MCMC, we consider the following Boltzman distribution over $\mathcal{X}$:

$$f_T(x) \propto \exp(-C(x)/T)$$

Here $T > 0$ is considers a *temperature coefficient*, that determines the steepness of f, from uniform distribution (for $T \to \infty$) to one that is fully concentrated on the minima of C (as $T \to 0$). Note that the normalization coefficient for f_T, which is hard to compute, is not needed.

For a fixed T, the idea is simply to obtain sequential samples from f_T using MCMC sampling. This can be seen as a random walk over $\mathcal{X}$, which focuses on areas of high f_T (low cost $C(x)$).

In a typical application of the M-H algorithm, the proposal distribution from a point x to one of its neighbors $y \in N_x$ may be uniform, with acceptance probability

$$\alpha(x, y) = \min\{1, \rho(x, y)\}, \quad \rho(x, y) = \frac{n_x f_T(y)}{n_y f_T(x)} = \frac{n_x}{n_y} \exp(-(C(y) - C(x))/T)$$

Thus, if $C(y) \leq C(x)$ then y is always accepted, while for $C(y) > C(x)$ the probability of accepting y is non-zero (which allows to escape from local minima) but decreasing in the difference $C(x) - C(y)$.

Simulated Annealing starts with relatively large T, which allows to quickly reach "interesting regions" of $\mathcal{X}$, and proceed to gradually reduce T, which allows a finer resolution. The method or rate of reduction of T is called the *cooling schedule* or *annealing schedule.* Some common choices are the following:

- Originally a geometric decrease in the number of cooling phases was considered, namely $T' = \gamma T$ for some $\gamma \in (0, 1)$, where each temperature level is held fixed for a certain number of samples that allow the MCMC sampling to take place effectively.

- Later analysis established convergence in probability to the global minimum when the temperature T_k decreases proportionally to $1/\log k$, where k is the step size.
- Practically the latter decay rate is often too slow, and a decay proportional to $1/k$ may be more useful. However, the most effective choice depends on the application.

We proceed with two illustrative examples (rough outline only):

6.5.1 The Travelling Salesperson Problem (TSP)

Weighted graph $G = (V, E)$ with $n + 1$ nodes, starting node '0'
Cost w_{ij} for each edge ij.
A *tour* (or Hamiltonian Cycle) is a path that visits each node exactly once and then returns to the start.
The problem is to find the shortest tour. This problem is known to be NP-hard.
Assuming (wlog) that the graph is fully connected, there are $n!$ tours.
Each tour can be represented as a permutation $x \equiv \sigma$ of $1 \dots n$, with cost

$$C(x) = w_{0x_1} + w_{x_1x_2} + \cdots + w_{x_n0}$$

To apply MCMC, we need to define neighborhood structure on the space of permutations, namely the allowed transitions from a given x.
The simplest options in called *2-opt*: Choose randomly two different indices in $1 \dots n$, and reverse their position in the permutation.

6.5.2 The Max-cut problem

A *cut* in an (undirected) graph (V, E) is a partition of V into two disjoint sets V_1, V_2.
The *cut set* is the set of edges that connect V_1 and V_2.
The max-cut problem: find a cut with maximal cut-set.
This problem is NP-complete. (Note: The min-cut problem is polynomial-time)
We can also define a weighted version (with positive weights).

A possible neighborhood structure here: choose a random node and move to the other group.