Winter 2021 Math 106
Topics in Applied Mathematics
Data-driven Uncertainty Quantification

Yoonsang Lee (yoonsang.lee@dartmouth.edu)

Lecture 9: Markov Chain Monte Carlo

9.1 Markov Chain

A Markov Chain Monte Carlo (MCMC) is a method producing an **ergodic Markov chain** X_t whose **stationary** distribution is the target density $p(x)$.

- Markov Chain
- Ergodic
- Stationary

9.1 Markov Chain

- A **stochastic process** $\{X_t \in \Omega, t \in T\}$ is a collection of random variables where Ω is the *state space* and T is the *index set*.
  Example. If $\{X_i\}$ is IID, it is also a stochastic process with an index $t = i$.
  Example. Let $\Omega = \{\text{sunny,cloudy,rain,snow}\}$. A typical sequence might be

$$\text{sunny,cloudy,snow,snow,snow,snow,sunny,}$$

  which is a stochastic process with a discrete index set.

9.1 Markov Chain

- A **Markov chain** is a stochastic process for which the distribution of X_t depends only on X_{t-1}

$$\mu(X_t = x | X_0, X_1, ..., X_{t-1}) = \mu(X_t = x | X_{t-1})$$

  for all t and x.

- In general,

$$p(x_1, x_2, ..., x_t) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_t|x_1, x_2, ..., x_{t-1})$$

  If X_t is a Markov chain,

$$p(x_1, x_2, ..., x_t) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_t|x_{t-1})$$

- For simplicity, we will consider the case when the state space is discrete.

9.1 Markov Chain

- The probability from $X_t = j$ to $X_{t+1} = i$

$$\mu(X_{t+1} = i | X_t = j) := p_{ij}$$

  is call the transition probability.
- The matrix P whose (i, j) element is p_{ij} is called the transition matrix.
- A Markov chain is **homogeneous** if the probability $\mu(X_{t+1} = i | X_t = j)$ does not change with time. That is

$$\mu(X_{t+1} = i | X_t = j) = \mu(X_t = i | X_{t-1} = j)$$

- It is straightforward to check that the n-th step probability

$$p_{ij}(n) := \mu(X_{t+n} = i | X_t = j) = (\mathrm{P}^n)_{ij}$$

9.1 Markov Chain

- **Chapman-Kolmogorov equations**

$$p_{ij}(n+m) = \sum_k p_{ik}(m) p_{kj}(n).$$

- Let f_0 be the initial probability. The marginal probability at the n-th step f_n such that $f_n(i) = \mu(X_n = i)$ is given by

$$f_n = \mathrm{P}^n f_0$$

because

$$f_n(i) = \mu(X_t = i) = \sum_k \mu(X_t = i | X_0 = k) \mu(X_0 = k)$$

$$= \sum_k p_{ik}(n) f_0(k)$$

9.1 Markov Chain

- The state j **reaches** the state i (or i is **accessible** from j) if $p_{ij}(n) > 0$ for some n and we write $j \to i$.
- If $j \to i$ and $i \to j$, we say i and j **communicate**.
- If all states communicate with each other, the chain is called **irreducible**.
- A set of states is **closed** if, once you enter that set you never leave.
- State i is **recurrent** or **persistent** if

$$\mu(X_n = i \text{ for some } n \geq 1 | X_0 = i) = 1.$$

  Otherwise, state i is **transient**.

9.1 Markov Chain

- Suppose $X_0 = j$. The **recurrence time** T_{ij} is defined as

$$T_{ij} = \min\{n > 0 : X_n = i\}$$

  assuming X_n ever returns to state i, otherwise $T_{ij} = \infty$.

- The **mean recurrence time** of a recurrent state i is

$$m_i = E[T_{ii}] = \sum_n n f_{ii}(n)$$

  where f_{ii} is the probability that the chain starting from state i returns to state i at the n-th step for the first time, that is,

$$f_{ii}(n) = \mu(X_1 \neq i, X_2 \neq i, ..., X_{n-1} \neq i, X_n = i | X_0 = i).$$

- A recurrent state is **null** if $m_i = \infty$. Otherwise it is called **positive** or **non-null**.

9.1 Markov Chain

- The **period** of state i, $d(i)$, is $gcd\{n : p_{ii}(n) > 0\}$. Note that it is gcd (the greatest common divisor), not the minimum value.
- State i is **periodic** if $d(i) > 1$ and **aperiodic** if $d(i) = 1$.
- **Definition.** A state is **ergodic** if it is recurrent, non-null, and aperiodic. A chain is ergodic if all its states are ergodic.
- **Definition.** A distribution π is **stationary** (or **invariant**) if

$$\pi = \mathrm{P}\pi.$$

- **Definiton.** We say that a chain has a **limiting distribution** if $\pi_i = \lim_{n\to\infty} p_{ij}(n)$ exists and is independent of j.
- **Theorem.** An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distribution exists and is equal to π. If g is any bounded function, then, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} g(X_n) \to E_\pi[g]$$

9.1 Markov Chain

- **Example.** Let
$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$
Let $\pi = (1/3, 1/3, 1/3)$. Then π is a stationary distribution of P.
- **Example.** Let $\Omega = \{1,2,3,4,5,6\}$. Let
$$P = \begin{pmatrix} 1/2 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/2 & 3/4 & 1/4 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/4 & 1/2 & 1/2 \end{pmatrix}$$
  - $C_1 = \{1,2\}$ and $C_2 = \{5,6\}$ are irreducible closed sets.
  - States 3 and 4 are transient (note the path $3 \to 4 \to 6$).
  - All states are aperiodic because $p_{ii}(1) > 0$.
  - 1,2,5 and 6 are ergodic.

9.1 Markov Chain

Exercise. Let P be a matrix of transition probabilities of a homogeneous ergodic Markov chain on a finite state space such that $p_{ij} = p_{ji}$. Find its stationary distribution.

9.1 Markov Chain

Exercise. Let P be a matrix of transition probabilities of a homogeneous ergodic Markov chain on a finite state space such that $p_{ij} = p_{ji}$. Find its stationary distribution.

- A distribution π satisfies **detailed balance** if

$$p_{ij}\pi_j = p_{ji}\pi_i$$

- If π satisfies detailed balance, it is a stationary distribution.

9.1 Markov Chain

Exercise. Let P be a matrix of transition probabilities of a homogeneous ergodic Markov chain on a finite state space such that $p_{ij} = p_{ji}$. Find its stationary distribution.

- A distribution π satisfies **detailed balance** if

$$p_{ij}\pi_j = p_{ji}\pi_i$$

- If π satisfies detailed balance, it is a stationary distribution.
  Idea of Proof. We want to show $P\pi = \pi$. The i-th component of $P\pi$ is

$$\sum_j p_{ij}\pi_j = \sum_j p_{ji}\pi_i = \pi_i \sum_j p_{ji} = \pi_i.$$

9.1 Markov Chain

Exercise. Let P be a matrix of transition probabilities of a homogeneous ergodic Markov chain on a finite state space such that $p_{ij} = p_{ji}$. Find its stationary distribution.

- A distribution π satisfies **detailed balance** if

$$p_{ij}\pi_j = p_{ji}\pi_i$$

- If π satisfies detailed balance, it is a stationary distribution.

For the uniform distribution $\pi_j = 1/n$, it satisfies the detailed balance condition

$$p_{ij}\pi_j = p_{ji}\pi_j = p_{ji}\pi_i.$$

Thus the uniform distribution is the stationary distribution.

9.1 Markov Chain

Exercise. Consider a homogeneous Markov chain on the finite state space $\Omega = \{1, 2, ..., r\}$. Assume that all the elements of the transition matrix are positive. Prove that for any $k \geq 0$ and any $x^0, x^1, ..., x^k \in \Omega$,

$$\mu(\text{there is } n \text{ such that } x_n = x^0, x_{n+1} = x^1, ..., x_{n+k} = x^k) = 1$$

Exercise. For a homogeneous Markov chain on a finite state space Ω with transition matrix P and initial distribution π_0, find

$$\mu(x_n = x^1 | x_0 = x^2, x_{2n} = x^3)$$

where $x^1, x^2, x^3 \in \Omega$.

9.2 MCMC: The Metropolis-Hastings Algorithm

Goal of MCMC: We want to draw a sample from a density $f(x)$. MCMC generates a Markov chain whose stationary density is the target density.

The Metropolis-Hastings Algorithm. Let $q(y|x)$ be a proposal density where it is easy to draw a sample from $q(y|x)$.

1. Choose X_0 arbitrarily.
2. Suppose we have generated $X_0, X_1, ..., X_t$. To generate X_{t+1},
3. Generate a proposal value Y from $q(y|X_t)$.
4. Evaluate $r := r(Y, X_t)$ where

$$r(y,x) = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$

5. Set

$$X_{t+1} = \begin{cases} Y & \text{with probability r} \\ X_t & \text{with probability 1-r} \end{cases}$$

9.2 MCMC: The Metropolis-Hastings Algorithm

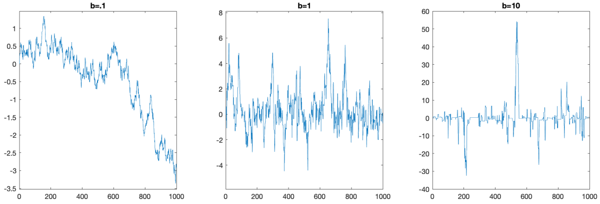
- **Remark.** If the proposal density is symmetric, that is, $q(y|x) = q(x|y)$,

$$r = \left\{\min \frac{f(y)}{q(x)}, 1\right\}$$

- A common choice for $q(y|x)$ is a normal $N(0, b^2)$ for some $b > 0$.
- Matlab/Python code

9.2 MCMC: The Metropolis-Hastings Algorithm

Example. MCMC samples for a Cauchy density $p(x) = \frac{1}{\pi(1+x^2)}$ using $b = 10, 1$, and $.1$.



- $b = .1$ forces the chain to take small steps. Thus the chain doesn't explore much of the sample space.
- $b = 10$ causes the proposals to often be far in the tails, making r small. Thus we reject the proposal and keep the current value.
- If the sample looks like the target distribution, the chain is called "mixing well".
- Constructing a chain that mixes well is an art.

9.2 MCMC: The Metropolis-Hastings Algorithm

We need to restate the detailed balance condition in the continuous case

- Let $p(y, x)$ be the transition probability density from x to y.
- A probability density $f(y)$ is **stationary** if

$$f(y) = \int p(y, x) f(x) dx$$

- **Detailed balance**

$$p(y, x) f(x) = p(x, y) f(y)$$

- If f satisfies detailed balance, f is a stationary distribution because

$$\int p(y, x) f(x) dx = \int p(x, y) f(y) dx = f(y) \int p(x, y) dx = f(y)$$

9.2 MCMC: The Metropolis-Hastings Algorithm

We are going to show that the MCMC algorithm satisfies the detailed balance condition.

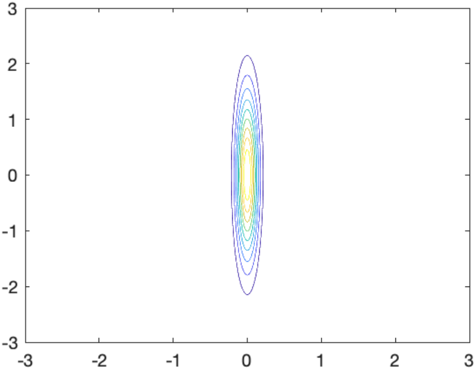
- Consider two points x and y.
- Either

$$f(x)q(y|x) < f(y)q(x|y) \quad \text{or} \quad f(x)q(y|x) > f(y)q(x|y)$$

- Assume that $f(x)q(y|x) > f(y)q(x|y)$ (if not, switch x and y).
- $r(y, x) = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\} = \frac{f(y)q(x|y)}{f(x)q(y|x)}$ and $r(x, y) = 1$.
- $p(y, x) = q(y|x)r(y, x) = q(y|x)\frac{f(y)q(x|y)}{f(x)q(y|x)} = \frac{f(y)}{f(x)}q(x|y)$. That is, $p(y, x)f(x) = f(y)q(x|y)$.
- Also $p(x, y) = q(x|y)r(x, y) = q(x|y)$. That is, $p(x, y)f(y) = q(x|y)f(y)$.

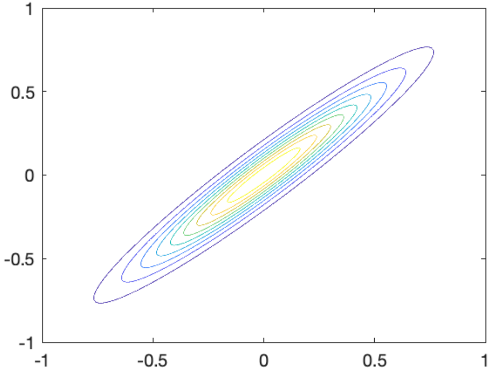
9.3 Examples

- $f(x,y) \sim \exp\left(-\frac{100x^2}{2} - \frac{y^2}{2}\right)$.



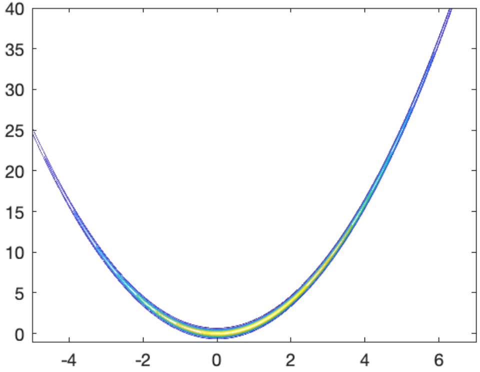
9.3 Examples

- $f(x,y) \sim \exp\left(-\frac{100(x-y)^2}{2} - \frac{(x+y)^2}{2}\right)$.



9.3 Examples

- (The Rosenbrock density) $f(x,y) \sim \exp\left(-\frac{100(y-x^2)^2+(1-x)^2}{20}\right)$.



9.4 Affine invariant MCMC sampler (Goodman and Weare, '10)

- An affine transformation between affine spaces is a function that preserves points, straight lines and planes.
- In a compact form, an affine transformation has the following structure
$$y = Ax + b.$$
- The examples in 9.4 can be transformed into a simple problem using affine transformations (exercise).
- Let us assume that there is an affine transformation that makes a sampling in the transformed space is easy (we assume only existence).

9.4 Affine invariant MCMC sampler (Goodman and Weare, '10)

- In an abstract form, an MCMC sampler has the following structure
$$x_{t+1} = R(x_t, f, \xi)$$
for a function R where f is the target density and ξ is random variables (like the uniform density random variable to accpet/reject an proposal value).
- If R is an efficient sampling method in the transformed space, the method need to preserve the affine transformation, that is,
$$Ax_{t+1} + b = R(Ax_t + b, f, \xi)$$
- An MCMC sampler with affine invariance still requires a tuning parameter but the parameter is independent of the sample space dimension.

9.4 Affine invariant MCMC sampler (Goodman and Weare, '10)

Walk-move MCMC sampler with affine invariance.

- Instead of running one Markov chain $\{x_t\}$, there are K Markov chains $\{x_t^k\}, k = 1, 2, ..., K$. Each chain is called 'walker'.
- For a given k at time t, let $X_t^{k'} = \{x_t^1, x_t^2, ..., x_t^{k-1}, x_t^{k+1}, ..., x_t^K\}$, that is the other chain values at the same time except x_k^t.
- x_{t+1}^k is $x_t^k + W$ with an acceptance probability $\min\{1, \frac{f(x_t^k+W)}{f(x_t^k)}\}$ where

$$W = \sum_{j \neq k} Z_j(x_t^j - \bar{x}_t),$$

Z_j is standard normal, and $\bar{t}$ is the mean of $X_t^{k'}$. Note that the covariance of W is the covariance of $X_t^{k'}$.

9.4 Affine invariant MCMC sampler (Goodman and Weare, '10)

- Python implementation: 'emcee' (or MCMC hammer) at http://dfm.io/emcee/current
- Matlab implementation: 'gwmcmc' at https://www.mathworks.com/matlabcentral/fileexchange/49820-ensemble-mcmc-sampler
- R implementation: https://rdrr.io/github/SandaD/MCMCEnsembleSampler /man/s.m.mcmc.html

9.5 Optimization

Goal. Find $\theta \in \Omega$ that maximizes $J(\theta)$.

Deterministic methods

- Steepest descent, gradient descent, etc.
- Convexity is important.
- Find only local extrema if it is not convex.

Monte Carlo optimization

- Assume that $J(\theta)$ is nonnegative (if not, take $\tilde{J}(\theta) = e^{J(\theta)}$).
- Draw a sample $\{\theta_t\}$ from $J(\theta)$.
- Choose θ^* such that $J(\theta^*) = max_{\{\theta_t\}} J(\theta)$.

Additionally,

9.5 Optimization

Goal. Find $\theta \in \Omega$ that maximizes $J(\theta)$.

Deterministic methods

- Steepest descent, gradient descent, etc.
- Convexity is important.
- Find only local extrema if it is not convex.

Monte Carlo optimization

- Assume that $J(\theta)$ is nonnegative (if not, take $\tilde{J}(\theta) = e^{J(\theta)}$).
- Draw a sample $\{\theta_t\}$ from $J(\theta)$.
- Choose θ^* such that $J(\theta^*) = max_{\{\theta_t\}} J(\theta)$.

Additionally,

- Use θ^* as an initial value for a deterministic optimization method.
- Use $J(\theta)^n$ instead of $J(\theta)$.

9.5 Optimization

Example. Find $\theta \in [0, 1]$ that maximizes

$$J(\theta) = \cos(7\theta) + \sin(20\theta)^2.$$

Homework

1. Modify the one-dimensional MCMC code provided in the lecture for sampling in *n*-dimensional spaces.
2. Use your MCMC code to generate a sample of size $10,000$ from a two-dimensional Gaussian with mean zero and a covariance matrix $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. Use (-4,4) as an initial value for your chain. Please specify your proposal density for sampling and related parameters.
3. Draw an empirical density of your sample.
4. Compare it with the true density. There are many different ways to answer this question. Try as many as possible.
5. Draw an empirical density using only the last half of your sample chain. Also compare this with the true density.
6. Compare the two empirical densities using the whole chain (3) and the half last chain (5). Discuss their accuracy.
7. Find $(x, y) \in [-1, 1]^2$ that **minimizes** $J(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(sin(10x)x) + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y)$