MATH-414 – Stochastic simulation

Lecture 9: Markov Chain Monte Carlo (discrete state space)

Prof. Fabio Nobile

EPFL

Outline

Markov Chains on discrete state space

- Invariant measure
- Reversibility and detailed balance

Markov Chain Monte Carlo (MCMC) in discrete state space

- Metropolis-Hastings

Problem setting

- $\mathcal{X}$: state space
- π: target probability measure on $\mathcal{X}$, possibly known only up to a multiplicative constant (i.e. $\pi = C\tilde{\pi}$ and only $\tilde{\pi}$ is accessible).

Goals:

- sample from π
- Given $\psi : \mathcal{X} \to \mathbb{R}$ with finite first moment wrt π, compute $\mu = \mathbb{E}_\pi[\psi]$

Example 1 (Statistical physics) Given an energy function $H : \mathcal{X} \to \mathbb{R}$ and a (absolute) temperature $T > 0$, sample from Gibbs' distribution

$$\pi(x) = \frac{1}{Z} \exp\left\{-\frac{H(x)}{kT}\right\}$$

with normalizing constant Z (partition function) not easily computable.

Example 2: (Bayesian inference) Given an iid sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a parametric distribution $g_\theta(X)$

- Likelihood function: $L(\theta \mid \mathbf{X}) = \prod_{i=1}^n g_\theta(X_i)$
- Prior distribution $\pi_0(\theta)$

Goal: sample from posterior distribution $\pi(\theta) = \dfrac{1}{Z(\mathbf{X})} L(\theta \mid \mathbf{X})\pi_0(\theta)$

Idea of Markov Chain Monte Carlo (MCMC)

- Construct an ergodic Markov Chain $\{X_n,\ n \in \mathbb{N}_0\} \sim \mathrm{Markov}\,(\lambda, P)$ on $\mathcal{X}$ that has π as invariant distribution
- Approximate $\mu = \mathbb{E}_\pi[\psi]$ by ergodic estimator

$$\hat{\mu}_N^{MCMC} = \frac{1}{N} \sum_{i=1}^{N} \psi(X_i)$$

To reduce the influence of the (wrong) initial condition, we may cut out the first part of the chain

$$\hat{\mu}_{N,b}^{MCMC} = \frac{1}{N} \sum_{i=1}^{N} \psi(X_{i+b})$$

b is called the burn in time

Markov chain in discrete state space (review)

- State space $\mathcal{X} = \{y_1, \ldots, y_d\}$ with $d < \infty$ (finite) or $d = \infty$ (countable)
- Initial distribution $\lambda = (\lambda_1, \ldots, \lambda_d)$, $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$ (p.m.f on $\mathcal{X}$)
- homogeneous Markov Chain $\{X_n\}_{n\in\mathbb{N}_0} \sim \mathrm{Markov}\,(\lambda, P)$ with transition matrix P

Notation

- $\mathbb{P}_\lambda(\cdot)$, $\mathbb{E}_\lambda[\cdot]$ denote probability/expectation when $X_0 \sim \lambda$
- $\mathbb{P}_i(\cdot)$, $\mathbb{E}_i[\cdot]$ denote probability/expectation when $\lambda = \delta_{y_i}$ (i.e. $X_0 = y_i$ with probability 1)

Definition. *[Stopping time] A random variable τ is called a stopping time if the event $\{\tau \leq n\}$ depends only on $X_0, \ldots, X_n$.*

(i.e. the event $\{\tau \leq n\}$ is measurable with respect to the σ-algebra $\sigma(X_0, \ldots, X_n)$ generated by $X_0, \ldots, X_n$)

Markov property

We recall that a stochastic process $\{X_n \in \mathcal{X},\ n \in \mathbb{N}_0\}$ is a Markov chain if it satisfies

$$\mathbb{P}\left(X_{n+1} = y_{n+1} \mid X_n = y_n, X_{n-1} = y_{n-1}, \ldots, X_0 = y_0\right) = \mathbb{P}\left(X_{n+1} = y_{n+1} \mid X_n = y_n\right)$$

Transition matrix: $P_{ij} = \mathbb{P}\left(X_n = y_j \mid X_{n-1} = y_i\right)$
(independent of n if the chain is homogeneous)

The definition of a Markov chain implies the following Markov property:

- **Weak Markov Property**. Conditional on $X_m = y_i$, $\{X_{m+n},\ n \in \mathbb{N}_0\}$ is $\mathrm{Markov}\left(\delta_{y_i}, P\right)$ and independent of $\{X_0, \ldots, X_m\}$.
- **Strong Markov property**. Let τ be a stopping time of $\{X_n\}$. Conditional on $\tau < +\infty$ and $X_\tau = y_i$, $\{X_{\tau+n},\ n \in \mathbb{N}_0\}$ is $\mathrm{Markov}\left(\delta_{y_i}, P\right)$ independent of $X_0, \ldots, X_\tau$.

n-step transition matrix and Chapman-Kolmogorov eq.

Let $\{X_n\}_{n\in\mathbb{N}_0} \sim \mathrm{Markov}\,(\lambda, P)$

n-step transition matrix: $P_{ij}^{(n)} = \mathbb{P}(X_{m+n} = y_j \mid X_m = y_i)$
(independent of m by homogeneity of the chain)

$$P_{ij}^{(n)} = \sum_{\ell} \mathbb{P}(X_{m+n} = x_j \mid X_{m+n-1} = x_\ell, X_m = x_i)\mathbb{P}(X_{m+n-1} = x_\ell \mid X_m = x_i)$$
$$= \sum_{\ell} P_{\ell j} P_{i\ell}^{(n-1)}$$

In matrix notation

$$P^{(n)} = P^{(n-1)}P = \ldots = P^n$$

More generally

Chapman-Kolmogorov eq. $\qquad P^{(n+m)} = P^{(n)}P^{(m)}$

Invariant distribution

Let $\pi_i^{n,\lambda} = \mathbb{P}_\lambda(X_n = y_i)$

Probability distribution of X_n: $\pi^{n,\lambda} = (\pi_1^{n,\lambda}, \ldots, \pi_d^{n,\lambda}) \in \mathbb{R}_+^d$ (row vector)

$$\begin{aligned}\pi_i^{n,\lambda} &= \sum_\ell \mathbb{P}(X_n = y_i \mid X_{n-1} = y_\ell)\mathbb{P}_\lambda(X_{n-1} = y_\ell) \\ &= \sum_\ell P_{\ell i}\pi_\ell^{n-1,\lambda}\end{aligned}$$

In matrix notation,

$$\pi^{n,\lambda} = \pi^{n-1,\lambda}P = \lambda P^n.$$

Let $M_1(\mathcal{X}) = \{(\mu_1, \ldots, \mu_d) \in \mathbb{R}^d : \ \mu_i \geq 0, \ \sum_i \mu_i = 1\}$ be the set of p.m.f on $\mathcal{X}$

Definition. *A probability mass function $\pi \in M_1(\mathcal{X})$ is called **invariant distribution** for P if $\pi P = \pi$.*

A Markov chain $\{X_n\}_{n\in\mathbb{N}_0} \sim \mathrm{Markov}\,(\pi, P)$ with π invariant distribution for P is called at equilibrium or stationarity

Invariant distribution

A transition matrix P defines an operator $P : M_1(\mathcal{X}) \to M_1(\mathcal{X})$ acting (on the left) on probability measures.

$$\mu = \lambda P \qquad \Leftrightarrow \qquad \mu_i = \mathbb{P}_\lambda(X_1 = y_i) = \sum_\ell \lambda_\ell P_{\ell i}$$

An invariant measure π (if it exists) is a left eigenvector (fixed point) of such operator, corresponding to the eigenvalue $\sigma = 1$

Let $\mathcal{F}(\mathcal{X}) = \{\varphi : \mathcal{X} \to \mathbb{R}\}$ be the set of real valued measurable functions on $\mathcal{X}$. An element $\varphi \in \mathcal{F}(\mathcal{X})$ can be represented as a column vector $\varphi = (\varphi_1, \dots, \varphi_d)^\top \in \mathbb{R}^d$, $\varphi_i = \varphi(y_i)$

A transition matrix P defines an operator $P : \mathcal{F}(\mathcal{X}) \to \mathcal{F}(\mathcal{X})$ acting (on the right) on functions.

$$g = P\varphi \qquad \Leftrightarrow \qquad g_i = \mathbb{E}_i[\varphi(X_1)] = \sum_\ell \varphi(y_\ell)\mathbb{P}_i(X_1 = y_\ell) = \sum_\ell P_{i\ell}\varphi_\ell$$

Maximal eignevalue

Take $\varphi = (1, \dots, 1)^\top \in \mathcal{F}(\mathcal{X})$

$$(P\varphi)_i = \sum_\ell P_{i\ell} \underbrace{\varphi_\ell}_{=1} = 1 = \varphi_i$$

Hence $\varphi = (1, \dots, 1)$ is a right eigenvector and an eigenvalue $\sigma = 1$ always exists. Moreover, it is always the largest one

Lemma

Given a stochastic matrix $P \in \mathbb{R}^{d\times d}$, let (σ, v) be a left eigenpair of P, i.e. $vP = \sigma v$, with $\|v\|_{\ell^1} = \sum_j |v_j| < \infty$. Then $|\sigma| \leq 1$.

Proof.

$$|\sigma v_i| = |\sum_j v_j P_{ji}| \leq \sum_j |v_j| P_{ji}.$$

$$\implies \qquad |\sigma| \sum_i |v_i| \leq \sum_i \sum_j |v_j| P_{ji} = \sum_j |v_j| \underbrace{\sum_i P_{ji}}_{=1} = \sum_j |v_j|$$

which implies $|\sigma| \leq 1$.

Power iterations

Consider a Markov chain $\{X_n\}_{n\in\mathbb{N}_0} \sim \mathrm{Markov}\,(\lambda, P)$ and the distribution $\pi^{n,\lambda}$ of X_n

The equation $\pi^{n,\lambda} = \lambda P^n$ corresponds to power iterations hence we expect $\pi^{n,\lambda}$ to converge to the left eigenvector corresponding to the largest eigenvalue $\sigma = 1$ (at least in the finite dimensional case, provided $\sigma = 1$ is simple)

The convergence will be related to the *spectral gap* $1 - \beta$ with $\beta = \max_{2,\dots,d} |\sigma_i(P)|$ the second largest eigenvalue.

Time reversal chain

Let $\{X_n,\ n=0,\dots,N\} \sim \mathrm{Markov}(\pi, P)$ be a Markov chain at equilibrium with π invariant for P

Time reversal chain: $\{Y_n = X_{N-n},\ n=0,\dots,N\}$

The time reversal chain $\{Y_n,\ n=0,\dots,N\}$ is also a Markov chain

$$\{Y_n,\ n=0,\dots,N\} \sim \mathrm{Markov}\left(\pi, \hat{P}\right), \qquad \hat{P}_{ij} = P_{ji}\frac{\pi_j}{\pi_i}$$

Indeed

$$\begin{aligned}
&\mathbb{P}\left(Y_n = y_{i_n} \mid Y_0 = y_{i_0}, \dots, Y_{n-1} = y_{i_{n-1}}\right) \\
&\qquad = \mathbb{P}\left(X_{N-n} = y_{i_n} \mid X_N = y_{i_0}, \dots, X_{N-n+1} = y_{i_{n-1}}\right) \\
&\qquad = \frac{\mathbb{P}\left(X_{N-n} = y_{i_n}, \dots, X_N = y_{i_0}\right)}{\mathbb{P}\left(X_{N-n+1} = y_{i_{n-1}}, \dots, X_N = y_{i_0}\right)} \\
&\qquad = \frac{P_{i_1 i_0} P_{i_2 i_1} \dots P_{i_n i_{n-1}} \mathbb{P}\left(X_{N-n} = y_{i_n}\right)}{P_{i_1 i_0} P_{i_2 i_1} \dots P_{i_{n-1} i_{n-2}} \mathbb{P}\left(X_{N-n+1} = y_{i_{n-1}}\right)} \\
&\qquad = P_{i_n i_{n-1}} \frac{\pi_{i_n}}{\pi_{i_{n-1}}} =: \hat{P}_{i_{n-1}, i_n}
\end{aligned}$$

EPFL

Reversibility and detailed balance

Definition. *Let P be a stochastic matrix, π a distribution on $\mathcal{X}$ and $\{X_n\} \sim \mathrm{Markov}(\pi, P)$ a Markov chain. We say that $\{X_n\}_{n\geq 0}$ is* ***reversible*** *if for all $N \geq 1$, $\{X_{N-n}\}_{n=0}^{N} \sim \mathrm{Markov}(\pi, P)$.*

(Equivalent to say that the joint distributions of $(X_0, \ldots, X_N)$ and $(X_N, \ldots, X_0)$ are the same for any $N \geq 0$)

If $\{X_n\} \sim \mathrm{Markov}(\pi, P)$ is reversible, in particular $\boldsymbol{X}_n \sim \pi$, $\forall n$ and π is invariant.

Definition. *A stochastic matrix P and a probability distribution λ on $\mathcal{X}$ are said to be in detailed balance if $\lambda_i P_{ij} = \lambda_j P_{ji}$ for all i, j.*

Reversibility and detailed balance

Lemma

Let P be a stochastic matrix and π a distribution on $\mathcal{X}$. (P, π) are in detailed balance if and only if π is invariant for P and $\{X_n\} \sim \mathrm{Markov}\,(\pi, P)$ is reversible.

Proof.

$(\Longrightarrow)$ Suppose (P, π) are in detailed balance. Then

$$(\pi P)_i = \sum_j \pi_j P_{ji} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i$$

hence π is an invariant distribution.

Moreover, the detailed balance condition directly implies $\hat{P} = P$, hence the chain $\{X_n\} \sim \mathrm{Markov}\,(\pi, P)$ is reversible.

$(\Longleftarrow)$ if π is invariant for P and $\{X_n\} \sim \mathrm{Markov}\,(\pi, P)$ is reversible, by definition $\hat{P} = P$ which is equivalent to the detailed balance condition. □

Reversibility and detailed balance

The reversibility (detailed balance) condition corresponds to a symmetry property of the transition matrix.

Let (π, P) be in detailed balance and define the Hilbert space $\ell^2_\pi = \{\varphi : \mathcal{X} \to \mathbb{R} : \sum_i \varphi_i^2 \pi_i < +\infty\}$ with inner product $(\varphi, \psi)_\pi = \sum_i \varphi_i \psi_i \pi_i$

Then P is symmetric (self adjoint) w.r.t. the inner product $(\cdot, \cdot)_\pi$

$$(P\varphi, \psi)_\pi = \sum_i \pi_i (P\varphi)_i \psi_i = \sum_{i,j} \pi_i P_{ij} \varphi_j \psi_i = \sum_{i,j} \pi_j P_{ji} \psi_i \varphi_j = (\varphi, P\psi)_\pi$$

This implies that the eigenvalues of P are all real and included in $[-1, 1]$.

MCMC in discrete state space

Given

- discrete state space $\mathcal{X} = \{y_1, \ldots, y_d\}$ (with d finite or not)
- probability mass function *pi* on $\mathcal{X}$ (possibly un-normalized)

Goal: Construct a Markov chain $\{X_n\}_{n\in\mathbb{N}_0} \sim \mathrm{Markov}\,(\lambda, P)$ which has π as (unique) invariant distribution.

Idea: rely on detailed balance condition: find P such that $\pi_i P_{ij} = \pi_j P_{ji}$

This will guarantee that π is invariant for P.

Metropolis-Hastings algorithm – discrete state space

- Take a stochastic matrix Q s.t.

$$Q_{ij} = 0 \iff Q_{ji} = 0$$

- For any $i, j \in \{1, \dots, d\}$, define the acceptance probability

$$\alpha(i,j) = \min\left\{1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right\} \text{ if } Q_{ij} \neq 0, \qquad \alpha(i,j) = 0, \text{ if } Q_{ij} = 0.$$

- Given X_n
  - generate proposal state $\tilde{X}_{n+1} \sim Q_{X_n,:}$
  - with probabiliy $\alpha(X_n, \tilde{X}_{n+1})$ accept the move and set $X_{n+1} = \tilde{X}_{n+1}$. Otherwise, set $X_{n+1} = X_n$

No need to know the normalization constant – only the ratio $\frac{\pi_{\tilde{X}_{n+1}}}{\pi_{X_n}}$ needs to be evaluated

If Q is symmetric, then $\alpha(i,j) = \min\{1, \frac{\pi_j}{\pi_i}\}$ – moves to higher probability states are always accepted.

Metropolis-Hastings algorithm – discrete state space

Algorithm: Metropolis-Hastings

Given: λ (initial distribution), Q (proposal), π (target distribution)

Generate $X_0 \sim \lambda$
for $n = 0, 1, \dots,$ **do**
  Generate candidate new state $\tilde{X}_{n+1} \sim Q_{X_n,:}$
  Generate $U \sim \mathcal{U}([0,1])$
  if $U \leq \alpha(X_n, \tilde{X}_{n+1})$ **then**
    set $X_{n+1} = \tilde{X}_{n+1}$ // $\tilde{X}_n$ accepted with prob. $\alpha(X_n, \tilde{X}_{n+1})$
  else
    set $X_{n+1} = X_n$ // $\tilde{X}_n$ rejected with prob. $1 - \alpha(X_n, \tilde{X}_{n+1})$
  end
end

Transition matrix of the Metropolis-Hastings algorithm

Lemma

Let $\alpha_i^* = \sum_j \alpha(i,j) Q_{ij}$*. The transition matrix of the chain produced by the Metropolis-Hastings algorithm is given by*

$$P_{ij} = \alpha(i,j) Q_{ij} + (1 - \alpha_i^*)\delta_{ij}.$$

Proof

for $i \neq j$

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}\left(\tilde{X}_{n+1} = j, X_{n+1} = \tilde{X}_{n+1} \mid X_n = i\right)$$
$$= \mathbb{P}\left(X_{n+1} = \tilde{X}_{n+1} \mid \tilde{X}_{n+1} = j, X_n = i\right) \mathbb{P}\left(\tilde{X}_{n+1} = j \mid X_n = i\right) = \alpha(i,j) Q_{ij}$$

for $i = j$

$$P_{ii} = \mathbb{P}(X_{n+1} = i \mid X_n = i)$$
$$= \mathbb{P}\left(\tilde{X}_{n+1} = i, X_{n+1} = \tilde{X}_{n+1} \mid X_n = i\right) + \mathbb{P}\left(X_{n+1} \neq \tilde{X}_{n+1} \mid X_n = i\right)$$
$$= \alpha(i,i) Q_{ii} + \sum_j \mathbb{P}\left(\tilde{X}_{n+1} = j, X_{n+1} \neq \tilde{X}_{n+1} \mid X_n = i\right)$$
$$= \alpha(i,i) Q_{ii} + \sum_j (1 - \alpha(i,j)) Q_{ij} = \alpha(i,i) Q_{ii} + (1 - \alpha_i^*)$$

α_i^* represents the probability of accepting a new state when being in state i.

Detailed balance condition

Lemma

The transition matrix P of the Metropolis-Hasting algorithm is in detailed balance with π. Hence, the chain produced by MH is reversible and has π as invariant distribution.

Proof

We have to show that

$$\pi_i P_{ij} = \pi_j P_{ji}, \quad \forall i, j$$

Obvious if $i = j$. For $i \neq j$

$$\begin{aligned} \pi_i P_{ij} &= \pi_i \alpha(i,j) Q_{ij} = \pi_i Q_{ij} \min\left\{1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right\} \\ &= \min\left\{\pi_i Q_{ij}, \pi_j Q_{ji}\right\} \\ &= \underbrace{\min\left\{\frac{\pi_i Q_{ij}}{\pi_j Q_{ji}}, 1\right\}}_{\alpha(j,i)} \pi_j Q_{ji} = \pi_j P_{ji} \end{aligned}$$

EPFL