MATH-414 – Stochastic simulation

Lecture 11-12: Markov Chain Monte Carlo

Prof. Fabio Nobile

EPFL

Outline

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 $\mu = \mathbb{E}_\pi[\psi]$

Idea of MCMC:

- Construct a Markov Chain $\{X_n,\ n \geq 0\} \sim \mathrm{Markov}(\lambda, P)$ ergodic and with π as invariant measure
- Approximate $\mu = \mathbb{E}_\pi[\psi]$ by ergodic estimator

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

b is called the burn in time (initial length of the chain disregarded in the temporal average)

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 $n = 0, 1, \ldots,$ **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 $\alpha_j^* = \sum_j \alpha(i,j) Q_{ij}$. *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

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

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

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 $i \neq j$

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

Ergodicity of the MH Markov chain

Let us assume that $\pi_i > 0$ for any i (otherwise the state i can be removed from the state space).

- Irreducibility of P is implied by the irreducibility of Q. Hence we should always consider proposals Q that are irreducible.
- Positive recurrence is verified since P admits an inveriant probability measure by construction (remember that P has an invariant probability measure if and only if it is positive recurrent)
- Aperiodicity is satisfied automatically if $\alpha_i^* < 1$ for some i (positive probability of staying in state i in the next iteration – this brakes any periodicity).

  Only if $\alpha_i^* = 1$ for all i (e.g. in Gibb's sampler) we need to check that the chain is aperiodic

Markov Chains on continuous state space

Let $\mathcal{X} \subset \mathbb{R}^d$ with Borel σ-algebra $\mathcal{B}(\mathcal{X})$ (we could even work on an arbitrary metric space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$)

Definition. *A Markov transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a function $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ s.t.*

1. *for all $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $\mathcal{X}$,*
2. *for all $A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ is measurable.*

The transition density associated to P, if it exists, is a function

$$p : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+, \qquad \int_{\mathcal{X}} p(x, y) dy = 1, \quad \forall x \in \mathcal{X}$$

such that $P(x, A) = \int_A p(x, y)\, dy$ for all $A \in \mathcal{B}(\mathcal{X})$.

Definition. *A sequence of random variables $\{X_n \in \mathcal{X}, n \geq 0\}$ is a homogeneous Markov chain with transition kernel P and initial distribution λ, in short $\{X_n\} \sim \mathrm{Markov}(\lambda, P)$ if*

- $X_0 \sim \lambda$
- $\mathbb{P}(X_{n+1} \in A \mid X_n, \ldots, X_0) = \mathbb{P}(X_{n+1} \in A \mid X_n) = P(X_n, A)$

Markov Chains on continuous state space

- *n*-step transition kernel

$$P^{(n)}(x, A) := \mathbb{P}\left(X_n \in A \mid X_0 = x\right) = \int_{\mathcal{X}} P^{(n-1)}(y, A) P(x, dy), \quad P^{(1)} = P$$

- *n*-step transition density

$$p^{(n)}(x, y) = \int_{\mathcal{X}} p^{(n-1)}(z, y) p(x, z)\, dz, \qquad p^{(1)} = p.$$

Markov transition operator

- Markov transition operator acting on measures (to the left): $\mathcal{P} : \mathcal{M}_1(\mathcal{X}) \to \mathcal{M}_1(\mathcal{X})$

$$\mu = \lambda\mathcal{P} \quad \Longrightarrow \quad \mu(A) = \int_{\mathcal{X}} P(y, A)\lambda(dy), \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

implies

$$\lambda\mathcal{P}^2(A) = \int_{\mathcal{X}} \int_{\mathcal{X}} P(x, A)P(y, dx)\lambda(dy) = \int_{\mathcal{X}} P^{(2)}(y, A)\lambda(dy)$$

and

$$\lambda\mathcal{P}^n(A) = \int_{\mathcal{X}} P^{(n)}(y, A)\lambda(dy)$$

- n-step distribution

$$\pi^{n,\lambda}(A) = \mathbb{P}_\lambda(X_n \in A) = \int_{\mathcal{X}} P^{(n)}(y, A)\lambda(dy) = \lambda\mathcal{P}^n(A)$$

Invarinat measure and detailed balance

- A measure π is called invariant (or stationary) if

$$\pi = \pi\mathcal{P} = \int_{\mathcal{X}} P(y,\cdot)\pi(dy)$$

If f is the density of π and p the transition density, then

$$f(x) = \int_{\mathcal{X}} p(y,x) f(y)\, dy.$$

- A chain $\{X_n\}_{n=0} \sim \mathrm{Markov}\,(\lambda, P)$ is reversible if, for any $N > 0$, the chain $\{Y_n = X_{N-n}\}_{n=0}^{N} \sim \mathrm{Markov}\,(\lambda, P)$.
- (λ, P) are said to be in detailed balance if

$$\int_A P(x,B)\lambda(dx) = \int_B P(y,A)\lambda(dy), \quad \forall A, B \in \mathcal{B}(\mathcal{X}).$$

If ℓ denotes the density of λ, then $\ell(x)p(x,y) = \ell(y)p(y,x)$

- If (P, λ) are in detailed balance, then P is reversible and λ is an invariant measure. Indeed,

$$\int_{\mathcal{X}} P(x,B)\lambda(dx) = \int_B \underbrace{P(y,\mathcal{X})}_{=1} \lambda(dy) = \lambda(B).$$

Irreducibility

Definition. *A set $A \in \mathcal{B}(\mathcal{X})$ is accessible if $\mathbb{P}_x(\sigma_A < \infty) > 0$ for all $x \in \mathcal{X}$, where $\sigma_A = \inf\{n > 0 : \ X_n \in A\}$ is the return time to the set A.*

Definition. *A Markov chain $\{X_n\}_n \sim \mathrm{Markov}\,(\lambda, P)$ is irreducible if there exists a (σ-finite) measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, called* **irreducibility measure** *such that any set $A \in \mathcal{B}(\mathcal{X})$, with $\varphi(A) > 0$ is accessible.*

The notion of irreducibility does not really depend on the measure φ as lon as one irreducibilty measure exists.

Theorem
If $\{X_n\}_n \sim \mathrm{Markov}\,(\lambda, P)$ is irreducible for some irreducibility measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, then there exists a probability measure ψ on $\mathcal{B}(\mathcal{X})$, called **maximal irreducibility measure** *such that*

- *$\{X_n\}_n$ is ψ-irreducible*
- *For any other measure φ' on $\mathcal{B}(\mathcal{X})$ for which $\{X_n\}$ is φ'-irreducible, one has $\varphi' \gg \psi$ (i.e. for all $A \in \mathcal{B}(\mathcal{X})$, $\psi(A) = 0 \implies \varphi'(A) = 0$)*
- *Any invariant measure is a maximal irreducibility measure.*

Recurrence and a-periodicity

Definition. *A Markov chain* $\{X_n\}_n \sim \mathrm{Markov}(\lambda, P)$ *is* *aperiodic* *if for any* $x \in \mathcal{X}$ *and any accessible set* $A \in \mathcal{B}(\mathcal{X})$

$$\exists n_0 \geq 0 : \qquad P^{(n)}(x, A) > 0 \quad \forall n \geq n_0.$$

Let $A \in \mathcal{B}(\mathcal{X})$ and $V_A = \sum_{n \geq 0} \mathbb{1}_{\{X_n \in A\}}$ be the number of visits to A

Definition. *A Markov chain* $\{X_n\}_n \sim \mathrm{Markov}(\lambda, P)$ *is* *recurrent* *if it is irreducible and every accessible set* A *satisfies* $\mathbb{E}_x[V_A] = \infty$*, for all* $x \in A$*.*

It is *Harris recurrent* *if it is irreducible and every accessible set* A *satisfies* $\mathbb{P}_x(V_A = \infty) = 1$*, for all* $x \in A$*;*

Harris recurrence is stronger and implies recurrence. In the discrete state case, the two notions coincide.

Definition. *A Markov chain* $\{X_n\}_n \sim \mathrm{Markov}(\lambda, P)$ *is* *positive* *(recurrent) if it has an invariant probability measure.*

Theorem

An irreducible, recurrent Markov kernel P *admits a non-zero invariant measure, unique up to a multiplicative constant.*

If an irreducible Markov kernel P *has an invariant probability measure (i.e. is positive), then it is recurrent.*

Convergence of Markov chains

Theorem

Let $\{X_n\}_n$ be irreducible, poisitive, Harris recurrent, and aperiodic, with (unique) invariant distribution π. Then $\lim_{n\to\infty} \|\lambda \mathcal{P}^n - \pi\|_{TV} = 0$ *for any $\lambda \in \mathcal{M}_1(\mathcal{X})$.*

Theorem

Let $\{X_n\}_n$ be irreducible and positive, with invariant probability distribution π and let $\psi \in \mathcal{F}(\mathcal{X})$ be a π-integrable function with $\mathbb{E}_\pi[\psi] < \infty$. Then, for any $\lambda \in \mathcal{M}_1(\mathcal{X})$

$$\mathbb{P}_\lambda\left(\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \psi(X_j) = \mathbb{E}_\pi[\psi]\right) = 1.$$

Central limit theorems can be established as well.

Metropolis-Hastings algorithm on continuous state space

- State space $\mathcal{X} \subset \mathbb{R}^d$, target distribution π with density f.
- Take a proposal Markov transition kernel $Q : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ with transition density $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ satisfying

$$Q(x, A) = \int_A q(x,y)dy, \qquad x \in \mathcal{X}, \ \ A \in \mathcal{B}(\mathcal{X})$$

  and $q(x,y) = 0 \ \Leftrightarrow q(y,x) = 0$.

- Define acceptance rate $\alpha : \mathcal{X} \times \mathcal{X} \to [0,1]$

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

  (if $q(x,y) = 0$ simply set $\alpha(x,y) = 0$)

- Given X_n
  - generate proposal state $\boldsymbol{Y}_{n+1} \sim Q(X_n, \cdot)$
  - with probabiliy $\alpha(X_n, Y_{n+1})$ accept the move and set $X_{n+1} = Y_{n+1}$. Otherwise, set $X_{n+1} = X_n$

Metropolis-Hastings algorithm

Algorithm: Metropolis-Hastings.

Given: λ (initial measure), q (proposal density), f (target density)

Generate $X_0 \sim \lambda$
for $n = 0, 1, \ldots,$ **do**
  Generate $Y_{n+1} \sim q(X_n, \cdot)$ `// proposal state`
  Generate $U \sim \mathcal{U}(0, 1)$
  if $U \leq \alpha(X_n, Y_{n+1})$ **then**
    set $X_{n+1} = Y_{n+1}$ `// accept proposal`
  else
    set $X_{n+1} = X_n$ `// reject proposal`
  end
end

Transition kernel of Metropolis-Hastings algorithm

- For $x \in \mathcal{X}$, overall acceptance probability of accepting the move being in x:
$$\alpha^*(x) = \int_{\mathcal{X}} \alpha(x,y) q(x,y)\, dy$$

- Transition density of Metropolis-Hastings algorithm:
$$p(x,y) = \alpha(x,y)q(x,y) + (1-\alpha^*(x))\delta_x(y), \quad x,y \in \mathcal{X},$$
where $\delta_x(y)$ is a Dirac mass in x.

- Markov transition kernel of Metropolis-Hastings algorithm:
$$P(x,A) = \int_A \alpha(x,y)q(x,y)dy + (1-\alpha^*(x))\mathbb{1}_A(x), \quad A \in \mathcal{B}(\mathcal{X}).$$

Detailed balance condition

Lemma

The transition kernel P of the Metropolis-Hastings algorithm is in detailed balance with the probability density f. Hence f is invariant probability density for P.

Proof Consider first the part of the kernel that has a density

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

Hence

$$\begin{aligned}
\int_B P(x,A)f(x)\,dx &= \int_B \left(\int_A \alpha(x,y)q(x,y)\,dy\right) f(x)\,dx + \int_B (1-\alpha^*(x))\mathbb{1}_A(x)f(x)\,dx \\
&= \int_B \int_A f(y)\alpha(y,x)q(y,x)\,dy\,dx + \int_{A\cap B}(1-\alpha^*(x))f(x)\,dx \\
&= \int_A \left(\int_B \alpha(y,x)q(y,x)\,dx\right) f(y)\,dy + \int_A (1-\alpha^*(y))\mathbb{1}_B(y)f(y)\,dy \\
&= \int_A P(y,B)f(y)\,dy.
\end{aligned}$$

EPFL

On the convergence of the Metropolis-Hastings algorithm

- π-irreducibility. We have to check that each set $A \in \mathcal{B}(\mathcal{X})$ with $\pi(A) > 0$ is accessible.

  This will be guaranteed if the proposal density satisfies for instance $q(x, y) > 0, \ \forall x, y \in \mathcal{X}$

- Recurrence: this is guaranteed automatically by the fact that the chain has an invariant probability measure, π, hence it is positive. However, for convergence in total variation, we have to check Harris recurrence, which can be complicated.

- a-periodicity: This is guaranteed if $\alpha^*(x) < 1$ for any $x \in \mathcal{X}$. Indeed, in this case, by definition of accessible set,

$$\forall x \in \mathcal{X} \text{ and } A \in \mathcal{B}(\mathcal{X}) \text{ accessible}, \quad \exists n_0 : \ P^{(n_0)}(x, A) > 0.$$

  Hence

$$P^{(n_0+1)}(x, A) = \int_{\mathcal{X}} P(y, A) P^{(n_0)}(x, dy) \geq \int_A P(y, A) P^{(n_0)}(x, dy)$$
$$\geq \int_A (1 - \alpha^*(y)) P^{(n_0)}(x, dy) > 0.$$

  Iterating the argument, we see that $P^{(n)}(x, A) > 0$ for any $n \geq n_0$

Independence sampler

Idea: take proposal density $q(x, y) = g(y)$ independent of current state x, where $g : \mathcal{X} \to \mathbb{R}_+$ is a probability density function on $\mathcal{X}$ that dominates f (i.e. $g(x) = 0 \;\Rightarrow\; f(x) = 0$)

Algorithm: Independence sampler Metropolis-Hastings

Given: $X_0 \sim \lambda$, $\text{supp}(\lambda) \subset \text{supp}(f)$

for $n = 0, 1, \ldots,$ **do**

  Generate $Y_{n+1} \sim g$

  Compute $\alpha(X_n, Y_{n+1}) = \min\left\{\frac{f(Y_{n+1})}{f(X_n)} \frac{g(X_n)}{g(Y_{n+1})}, 1\right\}$

  Generate $U \sim \mathcal{U}(0,1)$ and set

$$X_{n+1} = Y_{n+1}, \text{ if } U \leq \alpha(X_n, Y_{n+1}), \qquad X_{n+1} = X_n, \text{ otherwise}$$

end

Similar to Acceptance-Rejection sampling but:

- Whenever the proposal is rejected, the current state is repeated in the chain, contrary to AR ($\rightsquigarrow$ induces correlation in the sequence)
- No need to estimate the constant $C = \sup_{x \in \mathcal{X}} g(x)/f(x)$

Convergence of Independence sampler

Lemma

Let $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ *be a Markov transition kernel with invariant measure* π*. If there exists* $\epsilon \in (0,1)$ *and a probability measure* ν *on* $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ *such that*

$$P(x, A) \geq \epsilon \nu(A), \quad \forall x \in \mathcal{X}, \; A \in \mathcal{B}(\mathcal{X}), \tag{1}$$

then

$$\|\pi^{n,\lambda} - \pi\|_{TV} \leq 2(1-\epsilon)^n. \tag{2}$$

- The condition (1) is called *uniform minorizing condition*.
- An exponential convergence of the type (2) is called *geometric ergodicity*

Proof

Consider two coupled chains $\{X_n\} \sim \mathrm{Markov}(\lambda, P)$ and $\{Y_n\} \sim \mathrm{Markov}(\pi, P)$ constructed using the following algorithm. (Rk. $\{Y_n\}$ is at stationarity)

Let $X_0 \sim \lambda,\ Y_0 \sim \pi$
for $n = 0, 1, \ldots,$ **do**
  Draw $Z_n \sim \mathrm{Be}(\epsilon)$, $\mathbb{P}(Z_n = 1) = \epsilon$, $\mathbb{P}(Z_n = 0) = 1 - \epsilon$
  if $Z_n = 1$ **then**
    draw $W \sim \nu$ and set $X_{n+1} = Y_{n+1} = W$
  else
    draw $X_{n+1} \sim \frac{P(X_n, \cdot) - \epsilon\nu(\cdot)}{1-\epsilon}$ and $Y_{n+1} \sim \frac{P(Y_n, \cdot) - \epsilon\nu(\cdot)}{1-\epsilon}$ independently
  end
end

- It is easy to verify that $\{X_n\} \sim \mathrm{Markov}(\lambda, P)$ and $\{Y_n\} \sim \mathrm{Markov}(\pi, P)$.
- Let $T = \inf\{n \geq 0 : Z_n = 1\}$, which satisfies $\mathbb{P}(T \geq n) = (1-\epsilon)^n$.
- After T, the two chains have the same distribution $X_n \sim Y_n$, $n > T$.

$$
\begin{aligned}
\|\pi^{n,\lambda} - \pi\|_{\mathrm{TV}} &= 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\pi^{n,\lambda}(A) - \pi(A)| = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\mathbb{P}(X_n \in A) - \mathbb{P}(Y_n \in A)| \\
&= 2 \sup_A |\mathbb{P}(X_n \in A, T < n) + \mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T < n) - \mathbb{P}(Y_n \in A, T \geq n)| \\
&= 2 \sup_A |\mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T \geq n)| \\
&= 2 \sup_A |\mathbb{P}(X_n \in A, Y_n \notin A, T \geq n) - \mathbb{P}(X_n \notin A, Y_n \in A, T \geq n)| \leq 2\mathbb{P}(T \geq n).
\end{aligned}
$$

Convergence of Independence sampler

Theorem

If there exists $M < +\infty$ such that $f(x) \leq Mg(x)$ for all $x \in \mathcal{X}$, then the chain generated by the independence sampler is uniformly ergodic and

$$\|\pi^{n,\lambda} - \pi\|_{TV} \leq 2\left(1 - \frac{C}{M}\right)^n, \quad \textit{for any } \lambda, \textit{ with } C = \int_{\mathcal{X}} f(x)\,dx.$$

Proof: If f is not normalized, let $\tilde{f} = f/C$, $C = \int_{\mathcal{X}} f$. Notice that

$$\alpha(x,y)q(x,y) = g(y)\min\left\{\frac{f(y)}{f(x)}\frac{g(x)}{g(y)}, 1\right\} = f(y)\min\left\{\frac{g(x)}{f(x)}, \frac{g(y)}{f(y)}\right\} \geq \frac{1}{M}f(y).$$

It follows that for any $A \in \mathcal{B}(\mathcal{X})$,

$$P(x,A) = \int_A \alpha(x,y)q(x,y)dy + (1-\alpha^*(x))\mathbb{1}_A(x) \geq \frac{1}{M}\int_A f(y)\,dy \geq \frac{C}{M}\pi(A)$$

and the result follows from Lemma 8.

Random walk Metropolis (RWM)

Idea: perform only local moves with proposed increment distributions identical and symmetric, i.e. $q(x, y) = q(|y - x|)$.

Tipical case $q(x, \cdot) = N(x, \sigma^2 I_{d \times d})$.

This algorithm leads to geometric ergodicity under the following (sufficient) conditions (see [Jarner-Hansen, 2000])

- f has super-exponential tails, i.e. it is positive, continuous and satisfies

$$\lim_{|x| \to \infty} \frac{x}{|x|} \cdot \nabla \log f(x) = -\infty$$

- f satisfies

$$\limsup_{|x| \to \infty} \frac{x}{|x|} \cdot \frac{\nabla f(x)}{|\nabla f(x)|} < 0$$

- q is bounded away from zero in some region around zero:

$$\exists \delta_q, \epsilon_q > 0 \text{ s.t.} \quad q(x) \geq \epsilon_q, \quad \text{for } |x| \leq \delta_q$$

One variable at a time

- Suppose $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(d)}$ and $x \in \mathcal{X}$ has components $x = (x^{(1)}, \ldots, x^{(d)})$; Notation: $x^{(\sim i)} = (x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(d)})$.
- Consider a family of proposal transition densities $q^{(i)} : \mathcal{X} \times \mathcal{X}^{(i)} \to \mathbb{R}_+$

Idea: update one component at the time either chosen randomly or by performing a systematic sweep over the components.

Algorithm: One variable at a time MH with random selection.

Generate $X_0 \sim \lambda$
for $n = 0, 1, \ldots$ **do**
   Draw index $i_n \sim \beta$ (p.m.f on $\{1, \ldots, d\}$). Set $x = X_n^{(i_n)}$
   Draw $y \sim q_{i_n}(X_n, \cdot)$ and set $Y_{n+1} = (y, X_n^{(\sim i_n)})$
   Compute $\alpha_{i_n}(X_n, Y_{n+1}) = \min\left\{ \frac{f(Y_{n+1})}{f(X_n)} \frac{q_{i_n}(Y_{n+1}, x)}{q_{i_n}(X_n, y)}, 1 \right\}$
   Set $X_{n+1} = \begin{cases} Y_{n+1} & \text{with prob. } \alpha_{i_n}(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$
end

One variable at a time

Algorithm: One variable at a time MH with systematic sweep.

Generate $X_0 \sim \lambda$

for $n = 0, 1, \ldots$ **do**

  Set $Y_{n+1,0} = X_n$

  for $i = 1, \ldots, d$ **do**

    Draw $y \sim q_i(X_n, \cdot)$ and set $\tilde{Y} = (y, Y^{(\sim i)}_{n+1,i-1})$

    Set $Y_{n+1,i} = \begin{cases} \tilde{Y}, & \text{with prob. } \alpha_i(Y_{n+1,i-1}, \tilde{Y}) \\ Y_{n+1,i-1}, & \textit{otherwise} \end{cases}$

  end

  $X_{n+1} = Y_{n+1,d}$

end

Gibbs sampler

Consider a one variable at a time sampler using the conditional distributions as proposal densities $q_i(x,\cdot) = f_{X^{(i)} \mid X^{(\sim i)}}(\cdot \mid x^{(\sim i)})$

Given $x = (x^{(i)}, x^{(\sim i)})$ and $y = (y^{(i)}, x^{(\sim i)})$, the acceptance rate is

$$\alpha_i(x,y) = \min\left\{ \frac{f(y)}{f(x)} \frac{f_{X^{(i)} \mid X^{(\sim i)}}(x^{(i)} \mid x^{(\sim i)})}{f_{X^{(i)} \mid X^{(\sim i)}}(y^{(i)} \mid x^{(\sim i)})}, 1 \right\}$$

$$= \min\left\{ \frac{f(y)}{f(x)} \frac{f(x)/f_{X^{(\sim i)}}(x^{(\sim i)})}{f(y)/f_{X^{(\sim i)}}(x^{(\sim i)})}, 1 \right\} = 1$$

hence, in Gibbs sampler all the moves are accepted, provided one is able to generate exactly from the conditional distributions $f_{X^{(i)} \mid X^{(\sim i)}}(\cdot \mid x^{(\sim i)})$.

Algorithm: Gibbs with random sweep.

Generate $X_0 \sim \lambda$
for $n = 0, 1, \dots$ **do**
    Draw i_n from a pmf β on $\{1, \dots, d\}$
    Generate $y^{(i_n)} \sim f(\cdot \mid X_n^{(\sim i_n)})$
    Set $X_{n+1} = (y^{(i_n)}, X_n^{(\sim i_n)})$
end

Metropolis Adjusted Langevin Algorithm (MALA)

Let $f : \mathbb{R}^d \to \mathbb{R}_+$ be our target probability density and consider the following Stochastic Differential Equation (Langevin dynamics)

$$dX_t = \nabla \log f(X_t) + \sqrt{2}\, dW_t, \quad t > 0, \quad X_0 \sim \lambda \tag{3}$$

with W_t a standard Wiener process and λ a probability measure on $\mathbb{R}^d$.

Let us denote by $\rho(x,t)$ the probability denisty function of X_t (provided it exists):

$$\int_A \rho(x,t)dx = \mathbb{P}_\lambda(X_t \in A)$$

Under quite general consitions on f, one has $\lim_{t\to\infty} \rho(x,t) = f(x)$, i.e. the distribution of X_t converges to f and f is an invariant probability density function for (3) (time continuous Markov chain)

Problem: usually, exact solutions of (3) are not available

Metropolis Adjusted Langevin Algorithm (MALA)

Remedy: use numerical discretization, e.g. Euler-Maruyama method

$$X_{n+1} = X_n + \Delta t \nabla \log f(X_n) + \sqrt{2\Delta t}\xi_n, \quad \xi_n \sim N(0, I) \tag{4}$$

However, the discrete time Markov chain $\{X_n\}_n$ will not have anymore f as invariant distribution due to the numerical discretization error

Idea: use (4) as a proposal distribution within a Metropolis-Hasting Algorithm

---

Algorithm: Metropolis Adjusted Langevin Algorithm (MALA).

Generate $X_0 \sim \lambda$
for $n = 0, 1, \dots$ **do**
  Generate $Y \sim N(X_n + \Delta t \nabla \log f(X_n), 2\Delta t I)$
  Compute $\alpha(X_n, Y) = \min\left\{1, \frac{f(Y)}{f(X_n)} \frac{\exp(-\|X_n - Y - \Delta t \nabla \log f(Y)\|^2/2\Delta t)}{\exp(-\|Y - X_n - \Delta t \nabla \log f(X_n)\|^2/2\Delta t)}\right\}$
  Set $X_{n+1} = \begin{cases} Y & \text{with prob. } \alpha(X_n, Y) \\ X_n & \text{otherwise} \end{cases}$
end

- Similar to a RWM; but proposal is not symmetric and uses gradients information

Ergodic estimator

- Let $\{X_n\}_n \sim \mathrm{Markov}(\lambda, P)$ on $\mathcal{X} \subset \mathbb{R}^d$, with unique invariant distribution π.
- We assume moreover that $\{X_n\}_n$ is geometrically ergodic, i.e. there exist $\gamma > 0$ and $h : \mathcal{X} \to \mathbb{R}_+$ s.t.

$$\|\pi^{n,\lambda} - \pi\|_{\mathrm{TV}} \leq \lambda(h) e^{-\gamma n}, \quad \lambda(h) = \int_{\mathcal{X}} h(x) d\lambda(x)$$

Recall that for $\mu, \nu \in \mathcal{M}_1(\mathcal{X})$

$$\|\mu - \nu\|_{\mathrm{TV}} = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)| = \sup_{\phi \in L^\infty(\mathcal{X})} \frac{\left|\int_{\mathcal{X}} \phi(x) d\mu(x) - \int_{\mathcal{X}} \phi(x) d\nu(x)\right|}{\|\phi\|_{L^\infty(\mathcal{X})}}$$

- Given a π-integrable function $\psi : \mathcal{X} \to \mathbb{R}$, we estimate $\mu = \mathbb{E}_\pi[\psi]$ by the ergodic estimator

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

- Question: how to monitor the approximation error $|\hat{\mu}_{N,b}^{MCMC} - \mu|$?

Bias

If the chain is at stationarity ($\lambda = \pi$), then $\hat{\mu}_{N,b}^{MCMC}$ is unbiased. Indeed, $X_n \sim f$, $\forall n$ and

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

If, instead, the chain is not at stationarity ($\lambda \neq \pi$), the estimator $\hat{\mu}_{N,b}^{MCMC}$ is biased ! However

$$\begin{aligned}
|\mathbb{E}_\lambda[\hat{\mu}_{N,b}^{MCMC} - \mu]| &= \left| \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_\lambda[\psi(X_{i+b}) - \mu] \right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \left| \int_{\mathcal{X}} \psi(y)\pi^{i+b,\lambda}(dy) - \int_{\mathcal{X}} \psi(y)d\pi(y) \right| \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \|\psi\|_{L^\infty(\mathcal{X})} \|\pi^{i+b,\lambda} - \pi\|_{\mathrm{TV}} \\
&\leq \frac{1}{N}\|\psi\|_{L^\infty(\mathcal{X})}\lambda(h) \sum_{i=b+1}^{N} e^{-\gamma i} \leq \frac{e^{-\gamma b}}{N} \frac{\|\psi\|_{L^\infty(\mathcal{X})}\lambda(h)}{1 - e^{-\gamma}}
\end{aligned}$$

Bias

- The Bias decays as $O(\frac{1}{N})$, faster than the standard deviation (which is $O(\frac{1}{\sqrt{N}})$)
- Moreover, it decays as $O(e^{-\gamma b})$ and can be dramatically reduced by increasing the *burn-in* b.
- Reasonable values of b can be guessed from a trace-plot of the chain $\{\psi(X_n)\}_n$ (smallest time after which the chain looks at stationarity)

Asymptotic variance

Assume that a sufficient burn-in period has been removed and the chain is essentially at stationarity. Then, the following result on the *asymptotic variance* holds

Lemma

Let $\{X_n\} \sim \mathrm{Markov}\,(\pi, P)$ *with* π *invariant for* P*, and denote*

$$c(k) = Cov_\pi(\psi(X_0), \psi(X_k)) = Cov_\pi(\psi(X_j), \psi(X_{j+k})).$$

Then

$$\mathbb{V}\mathrm{ar}_\pi[\hat{\mu}_{N,b}^{MCMC}] = \frac{\sigma^2_{MCMC,N}}{N}, \quad \textit{with } \sigma^2_{MCMC,N} = c(0) + 2\sum_{\ell=1}^{N-1}\left(1 - \frac{\ell}{N}\right)c(\ell).$$

Moreover, if $\sum_{k=0}^{\infty}|c(k)| < +\infty$*, then*

$$\lim_{N\to\infty} N\mathbb{V}\mathrm{ar}\left[\hat{\mu}_{N,b}^{MCMC}\right] = \sigma^2_{MCMC}$$

with $\sigma^2_{MCMC} = c(0) + 2\sum_{k=1}^{\infty} c(k)$.

Proof

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}_\pi[\hat{\mu}^{MCMC}_{N,b}] &= \mathbb{E}_\pi\left[\left(\frac{1}{N}\sum_{j=1}^{N}\psi(X_{j+b})-\mu\right)^2\right] \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathbb{E}_\pi[(\psi(X_{j+b})-\mu)(\psi(X_{k+b})-\mu)] \\
&= \frac{1}{N^2}\left[\sum_{j=1}^{N}\underbrace{\mathbb{V}\mathrm{ar}_\pi[\psi(X_{j+b})]}_{c(0)}+2\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\underbrace{\mathrm{Cov}_\pi(\psi(X_{j+b}),\psi(X_{k+b}))}_{c(k-j)}\right] \\
&= \frac{c(0)}{N}+\frac{2}{N^2}\sum_{j=1}^{N-1}\sum_{\ell=1}^{N-j}c(\ell) \\
&= \frac{c(0)}{N}+\frac{2}{N}\sum_{\ell=1}^{N-1}\frac{N-\ell}{N}c(\ell) \\
&= \frac{1}{N}\left(c(0)+2\sum_{\ell=1}^{N-1}\left(1-\frac{\ell}{N}\right)c(\ell)\right).
\end{aligned}
$$

Under the assumption $\sum_{\ell=0}^{\infty}|c(\ell)|<+\infty$, it follows that $\lim_{N\to\infty} N\mathbb{V}\mathrm{ar}_\pi[\hat{\mu}^{MCMC}_{N,b}]=\sigma^2_{MCMC}$.

Asymptotic variance

- The quantity

$$\sigma^2_{MCMC} = c(0) + 2\sum_{k=1}^{\infty} c(k)$$

  is called time-average variance constant (TAVC) or asymptotic variance
- If $\{X_n\}_n$ were iid and distributed as f (pure Monte Carlo sampling) then the variance of the Monte Carlo estimator would be $\mathbb{V}\mathrm{ar}\left[\hat{\mu}_N^{MC}\right] = \frac{c(0)}{N}$.
- Given N, we call effective sample size (ESS) the sample size that a Monte Carlo estimator would use to achieve the same variance as the MCMC one:

$$\mathbb{V}\mathrm{ar}\left[\hat{\mu}_{N,b}^{MCMC}\right] \to \frac{\sigma^2_{MCMC}}{N} = \frac{c(0)}{ESS} \qquad \Longrightarrow \qquad ESS = N\frac{c(0)}{\sigma^2_{MCMC}}$$

- For reversible, geometrically ergodic, Markov chains, a CLT holds

$$\sqrt{N}(\hat{\mu}_{N,b}^{MCMC} - \mu) \xrightarrow{\mathrm{d}} N(0, \sigma^2_{MCMC})$$

Estimating the asymptotic variance – covariance method

Given a path $\{X_n\}_n$ and a burn-in time, we can estimate the covariances

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

and

$$\hat{\sigma}^2_{MCMC} = \hat{c}(0) + 2\sum_{k=1}^{N-2}\hat{c}(k).$$

However, the last terms in the sum are very unstable. Better estimator

$$\hat{\sigma}^2_M = \hat{c}(0) + 2\sum_{k=1}^{M}\hat{c}(k), \quad \text{with } M = 2\min\{k : \hat{c}(2k) + \hat{c}(2k+1) < 0\}.$$

(valid for reversible Markov Chains)

Estimating the asymptotic variance - batch means

An alternative idea to estimate σ^2_{MCMC} is to split the sequence $\{X_n\}_{n=b+1}^{N+b}$ into M blocks of size $T = N/M$

Then we can build M different sample averages

$$\hat{\mu}^{(i)} = \frac{1}{T} \sum_{j=(i-1)T+b+1}^{iT+b} \psi(X_j), \quad \text{and} \quad \hat{\mu}^{MCMC}_{N,b} = \frac{1}{M} \sum_{i=1}^{M} \hat{\mu}^{(i)}.$$

If T is sufficiently large (larger than the relaxation time), the M blocks are nearly independent so that

$$\mathbb{V}\mathrm{ar}\left[\hat{\mu}^{MCMC}_{N,b}\right] \approx \frac{\sigma^2_{MCMC}}{N} \approx \frac{\mathbb{V}\mathrm{ar}\left[\hat{\mu}^{(1)}\right]}{M}$$

and $\mathbb{V}\mathrm{ar}\left[\hat{\mu}^{(1)}\right]$ can be estimated by a sample variance estimator

$$\mathbb{V}\mathrm{ar}\left[\hat{\mu}^{(1)}\right] \approx \hat{\sigma}^2_{\hat{\mu}^{(1)}} = \frac{1}{M-1} \sum_{i=1}^{M} \left(\hat{\mu}^{(i)} - \hat{\mu}^{MCMC}_{N,b}\right)^2.$$

Finally, an estimator for σ^2_{MCMC} is

$$\hat{\sigma}^2_{MCMC} = \frac{N}{M} \hat{\sigma}^2_{\hat{\mu}^{(1)}} = \frac{T}{M-1} \sum_{i=1}^{M} \left(\hat{\mu}^{(i)} - \hat{\mu}^{MCMC}_{N,b}\right)^2.$$