MATH-414 – Stochastic simulation

Lecture 12: Metropolis-Hastings algorithms

Prof. Fabio Nobile

EPFL

Outline

Metropolis-Hastings algorithms

Convergence diagnostics

Independence sampler

Idea: take proposal density $q(x, y) = g(y)$ independent of the 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: $X_0 \sim \lambda$, $\text{supp}(\lambda) \subset \text{supp}(f)$

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

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

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

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

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

5 **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 $\{X_n\} \sim \text{Markov}(\lambda, P)$ and $\{Y_n\} \sim \text{Markov}(\pi, P)$ 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 $Z_n \sim \text{Be}(\epsilon)$, $\mathbb{P}(Z_n = 1) = \epsilon$, $\mathbb{P}(Z_n = 0) = 1 - \epsilon$
  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 $\{X_n\} \sim \text{Markov}(\lambda, P)$ and $\{Y_n\} \sim \text{Markov}(\pi, P)$.
- 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 $X_n \sim Y_n$, $n > T$.

$$\|\pi^{n,\lambda} - \pi\|_{\text{TV}} = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\pi^{n,\lambda}(A) - \pi(A)| = 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\mathbb{P}(X_n \in A) - \mathbb{P}(Y_n \in A)|$$

$$= 2 \sup_A |\mathbb{P}(X_n \in A, T < n) + \mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T < n) - \mathbb{P}(Y_n \in A, T \geq n)|$$

$$= 2 \sup_A |\mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T \geq n)|$$

$$= 2 \sup_A |\mathbb{P}(X_n \in A, Y_n \notin A, T \geq n) - \mathbb{P}(X_n \notin A, Y_n \in A, T \geq n)| \leq 2\mathbb{P}(T \geq n).$$

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

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

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 1.

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: $x^{(-i)} = (x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(d)})$.
- 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.

---

1 Generate $X_0 \sim \lambda$

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

3 Draw index $i_n \sim \beta$ (p.m.f on $\{1, \ldots, d\}$). Set $x = X_n^{(i_n)}$

4 Draw $y \sim q_{i_n}(X_n, \cdot)$ and set $Y_{n+1} = (y, X_n^{(-i_n)})$

5 Compute $\alpha_{i_n}(X_n, Y_{n+1}) = \min\left\{\frac{f(Y_{n+1})}{f(X_n)} \frac{q_{i_n}(Y_{n+1}, x)}{q_{i_n}(X_n, y)}, 1\right\}$

6 Set $X_{n+1} = \begin{cases} Y_{n+1} & \text{with prob. } \alpha_{i_n}(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$

7 **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 $\tilde{Y} = (y, Y^{(-i)}_{n+1,i-1})$
    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

Detailed balance for one variable at a time

Let p_i be the transition density of the algorithm when the index i is selected:

$$\begin{aligned} p_i(x,y) &= p_i((x^{(i)},x^{(-i)}),(y^{(i)},y^{(-i)})) \\ &= \left(\alpha_i(x,y^{(i)})q_i(x,y^{(i)}) + (1-\alpha_i^*(x))\delta_{x^{(i)}}(y^{(i)})\right)\delta_{x^{(-i)}}(y^{(-i)}) \\ &= \underbrace{\alpha_i(x,y^{(i)})q_i(x,y^{(i)})\delta_{x^{(-i)}}(y^{(-i)})}_{p_i^1(x,y)} + \underbrace{(1-\alpha_i^*(x))\delta_{x^{(i)}}(y^{(i)})\delta_{x^{(-i)}}(y^{(-i)})}_{p_i^2(x,y)} \end{aligned}$$

with $\alpha_i^*(x) = \int_{\mathcal{X}} \alpha_i(x,y^{(i)})q_i(x,y^{(i)})dy^{(i)}$

Denote $P_i(x,A) = \int_A p_i(x,y)dy$ the corresponding Markov transition kernel and $\mathcal{P}_i$ the associated Markov transition operator.

Lemma

Let $P_i : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ is in detailed balance with f.

Proof of the Lemma

$$\int_A P_i(x,B)f(x)dx = \underbrace{\int_A\int_B p_i^1(x,y)f(x)dydx}_{T_1} + \underbrace{\int_A\int_B p_i^2(x,y)dydx}_{T_2}$$

$$\begin{aligned}
T_1 &= \int_A\int_B \min\left\{1, \frac{f(y^{(i)},x^{(-i)})q_i((y^{(i)},x^{(-i)}),x^{(i)})}{f(x)q_i(x,y^{(i)})}\right\} q_i(x,y^{(i)})\delta_{x^{(-i)}}(y^{(-i)})f(x)dxdy \\
&= \int_A\int_B \min\left\{1, \frac{f(y)q_i(y,x^{(i)})}{f(x)q_i(x,y^{(i)})}\right\} q_i(x,y^{(i)})\delta_{x^{(-i)}}(y^{(-i)})f(x)dydx \\
&= \int_A\int_B \min\left\{\frac{f(x)q_i(x,y^{(i)})}{f(y)q_i(y,x^{(i)})}, 1\right\} q_i(y,x^{(i)})\delta_{x^{(-i)}}(y^{(-i)})f(y)dydx \\
&= \int_A\int_B \min\left\{\frac{f(x^{(i)},y^{(-i)})q_i((x^{(i)},y^{(-i)}),y^{(i)})}{f(y)q_i(y,x^{(i)})}, 1\right\} q_i(y,x^{(i)})\delta_{y^{(-i)}}(x^{(-i)})f(y)dydx \\
&= \int_A\int_B \alpha_i(y,x^{(i)})q_i(y,x^{(i)})\delta_{y^{(-i)}}(x^{(-i)})f(y)dxdy = \int_B\int_A p_i^1(y,x)f(y)dxdy
\end{aligned}$$

$$T_2 = \int_A\int_B (1-\alpha_i^*(x))\delta_x(y)f(x)dydx = \int_B\int_A p_i^2(y,x)f(y)dxdy$$

EPFL

Detailed balance for one variable at a time

Random selection algorithm. Markov Transition kernel:

$$P^{\text{rand}}(x, A) = \mathbb{P}\left(X_{n+1} \in A \mid X_n = x\right)$$
$$= \sum_{i=1}^{d} \mathbb{P}\left(X_{n+1} \in A \mid X_n = x, i_n = i\right) \mathbb{P}\left(i_n = i\right) = \frac{1}{d} \sum_{i=1}^{d} P_i(x, A)$$

Associated operator: $\mathcal{P}^{\text{rand}} = \frac{1}{d} \sum_{i=1}^{d} \mathcal{P}_i$

(P_i, f) in detailed balance $\implies$ (P^{rand}, f) in detailed balance (hence $\mathcal{P}^{\text{rand}}$ reversible and with invariant distribution f)

Detailed balance for one variable at a time

Systematic sweep algorithm. Markov Transition kernel:

$$P^{\text{sweep}}(x, A) = \mathbb{P}(X_{n+1} \in A \mid X_n = x)$$
$$= \int_A \int_{\mathcal{X}^{d-1}} P_d(y_{d-1}, dy_d) \cdots P_2(y_1, dy_2) P_1(x, dy_1)$$

Associated operator: $\mathcal{P}^{\text{sweep}} = \mathcal{P}_1 \cdots \mathcal{P}_d$

f invariant for $P_i \quad \Longrightarrow \quad f$ invariant for $\mathcal{P}^{\text{sweep}}$. However $\mathcal{P}^{\text{sweep}}$ is not reversible!

Markov transition operator of the reversed chain $\widehat{\mathcal{P}}^{\text{sweep}} = \mathcal{P}_d \cdots \mathcal{P}_1$

Possible fix to recover reversibility (if important)

$$\mathcal{P}^{\text{sweep}} = \mathcal{P}_1 \cdots \mathcal{P}_d \cdots \mathcal{P}_1$$

forward sweep $i = 1, \ldots, d$ followed by backward sweep $i = d-1, \ldots, 1$.

Gibbs sampler

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

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

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

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

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

Algorithm: Gibbs with random sweep.

1 Generate $X_0 \sim \lambda$
2 **for** $n = 0, 1, \ldots$ **do**
3 | Draw i_n from a pmf β on $\{1, \ldots, d\}$
4 | Generate $y^{(i_n)} \sim f_{X^{(i_n)} \mid X^{(-i_n)}}(\cdot \mid X_n^{(-i_n)})$
5 | Set $X_{n+1} = (y^{(i_n)}, X_n^{(-i_n)})$
6 **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-Hastings Algorithm

---

Algorithm: Metropolis Adjusted Langevin Algorithm (MALA).

---

1 Generate $X_0 \sim \lambda$

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

3 Generate $Y \sim N(X_n + \Delta t \nabla \log f(X_n), 2\Delta t I)$

4 Compute $\alpha(X_n, Y) = \min\left\{1, \frac{f(Y)}{f(X_n)} \frac{\exp(-\|X_n - Y - \Delta t \nabla \log f(Y)\|^2 / 2\Delta t)}{\exp(-\|Y - X_n - \Delta t \nabla \log f(X_n)\|^2 / 2\Delta t)}\right\}$

5 Set $X_{n+1} = \begin{cases} Y & \text{with prob. } \alpha(X_n, Y) \\ X_n & \text{otherwise} \end{cases}$

6 **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

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

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

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

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

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

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

Proof

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

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

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 $\mathbb{V}\mathrm{ar}\left[\hat{\mu}^{MC}_N\right] = \frac{c(0)}{N}$.
- 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:

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

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

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

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

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