Markov Chain Monte Carlo

Bayes methods

2023-02-11

Target distribution P.

$X_0, X_1, X_2, \cdots$

Markov chain. (Finite)

(State space)

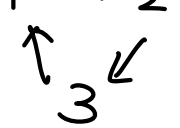
① irreducible. ($\forall i,j \in \Omega$, $i \leftrightarrow j$)

(不可约) (互通)

② aperiodic ($\forall i \in \Omega$, $d_i = 1$, $d = \gcd\{n, p_{ii}^{(n)} > 0\}$)

(非周期)

example,



$d_1 = \gcd(3, 6, 9) = 3$

$1 \leftrightarrow 2 \leftrightarrow 3$

$d_1 = d_2 = d_3 = 3$, not aperiodic.

$d_1 = \gcd(2, 3, \cdots) = 1$

$d_1 = d_2 = d_3 = 1$, aperiodic

stationary distribution.

$pP = p$. ($X_t \xrightarrow{D} p$)

p: Stationary distribution

P: transition probability matrix

Retailed balance (细致平衡条件)

time reversible (时间可逆) ☆☆☆☆

$$P_i P_{i,j} = P_j P_{j,i} \quad \forall\ i,j \in \Omega.$$

rate of $i \to j$ rate of $j \to i$.

(☆) P_{ij} is a TPR of a irreducible chain if we can find a sequence $\{\pi_i\}$ $\sum_i \pi_i = 1$ and an another TPR Q_{ij} satisfying:

$$\pi_i P_{ij} = \pi_j Q_{ji}$$

Q_{ji} is the TPR of the reversible chain,

$\{\pi_i\}$ is the stationary probability.

convergence and the Ergodic Theorem (遍历)

$$\frac{\sum_{t=1}^{T} f(X_t)}{T} \xrightarrow{a.s.} E_{x\sim p}(f(x))$$

M-H Algorithm. (target distribution p)

$X^{(t)} \to X^{(t+1)}$

(在每一步转移，若转移后能提高状态 X_t 在目标分布 P 中有密度值，就接受转移结果；否则，以一定的概率决定是转移还是不动)

Let $q(j|i) = Q_{ij}$, proposal distribution (simple).

$$q(j|i) > 0 \iff q(i|j) > 0$$

$X_t = i$ (discrete)

1. Draw $j \sim q(\cdot|i)$ and $u \sim U[0,1]$

2. If $u \le \alpha(j|i)$ where (If $i = j$, $\alpha(j|i) = 1$.)

$$\alpha(j|i) = \min\left\{1, \frac{p_j q(i|j)}{p_i q(j|i)}\right\}$$

(accept j with probability $\alpha(j|i)$)

then set $X_{t+1} = j$, otherwise set $X_{t+1} = i$.

(Start at $X_0 = i_0$)

Prove distribution of MCMC converges to p

$$P_{i,j} = q(j|i)\,\alpha(j|i) \qquad (j \neq i).$$

$$p_i P_{i,j} = p_i\, q(j|i)\,\alpha(j|i)$$

$$= p_i\, q(j|i) \min\left\{1, \frac{p_j\, q(i|j)}{p_i\, q(j|i)}\right\}$$

$$= \min\{p_i\, q(j|i),\ p_j\, q(i|j)\}$$

$$= p_j\, q(i|j) \min\left\{\frac{p_i\, q(j|i)}{p_j\, q(i|j)}, 1\right\}$$

$$= p_j\, q(i|j)\,\alpha(i|j)$$

$$= p_j\, P_{j,i}.$$

$\Rightarrow \{p_i\}$ is the stationary distribution.

Example. Hyper Geom $(k; K, N, n)$

$$\Omega = \{k \in \mathbb{Z} : \underbrace{\max\{0, n+K-N\}}_{B^-} \le k \le \underbrace{\min\{n, K\}}_{B^+}\}.$$

$$p_k = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad B^- \le k \le B^+$$

$$p_k = 0, \quad k \notin [B^-, B^+]$$

$$q(j|i) = \begin{cases} \frac{1}{2}, & \text{for } j = i \pm 1 \\ 0, & \text{other.} \end{cases}$$

① Draw $j \sim \underbrace{U\{i-1, i+1\}}_{\text{(discrete uniform)}}$ and $u \sim U(0,1)$

② If $B^- \le j \le B^+$,

$$u \le \min\left\{1, \frac{p_j \, q(i|j)}{p_i \, q(j|i)}_{\frac{1}{2}}^{=\frac{1}{2}}\right\}$$

$$= \min\left\{1, \frac{p_j}{p_i}\right\}$$

$$= \min\left\{1, \frac{\binom{K}{j}\binom{N-K}{n-j}}{\binom{K}{i}\binom{N-K}{n-i}}\right\}$$

then set $X_{t+1} = j$, otherwise $X_{t+1} = i$.

If $j < B^-$ or $j > B^+$, reject.

continuous

$$P(X_{t+1} \in A \mid X_t = i)$$

$$\underline{K(\theta, d\theta')} = P(X_{t+1} = d\theta' \mid X_t = \theta)$$

MH algorithm. (continuous) (target distribution p)

$q(\theta' \mid \theta)$. proposal probability density. (1-d)

$$q(\theta' \mid \theta) > 0 \iff q(\theta \mid \theta') > 0.$$

1. Draw $\theta' \sim q(\,\cdot \mid \theta)$ and $u \sim U[0,1]$

2. If $u \le \alpha(\theta' \mid \theta)$, where

$$\alpha(\theta' \mid \theta) = \min\left\{1, \frac{p(\theta')\, q(\theta \mid \theta')}{p(\theta)\, q(\theta' \mid \theta)}\right\}$$

Set $X_{t+1} = \theta'$, otherwise $X_{t+1} = \theta$.

$\theta = (\theta_1, \cdots \theta_p)$
↘ component

Updating different components using different proposal distribution. Suppose we have N different proposal densities $q_k(\theta'|\theta)$ and we select k'th one to make a proposal with probability ξ_k and then accept or reject:

1. Draw $k \sim \text{Multinom}(\xi_1, \cdots \xi_N)$

$\theta' \sim q_k(\,|\theta)$ and $u \sim U[0,1]$

2. If $u \in \alpha_k(\theta'|\theta)$,

$$\alpha_k(\theta'|\theta) = \min\left\{1, \frac{p(\theta')\, q_k(\theta|\theta')}{p(\theta)\, q_k(\theta'|\theta)}\right\}$$

then set $X_{t+1} = \theta'$, otherwise set $X_{t+1} = \theta$.

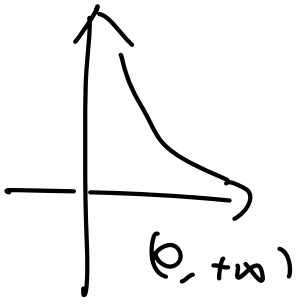
If we have p components, $\theta = (\theta_1, \cdots \theta_p)$.

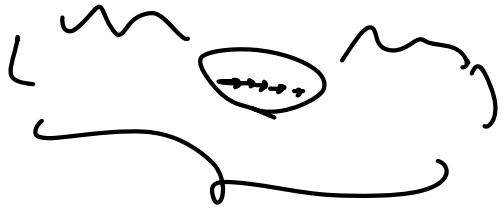
Take $N = p$, if we choose to update component k.

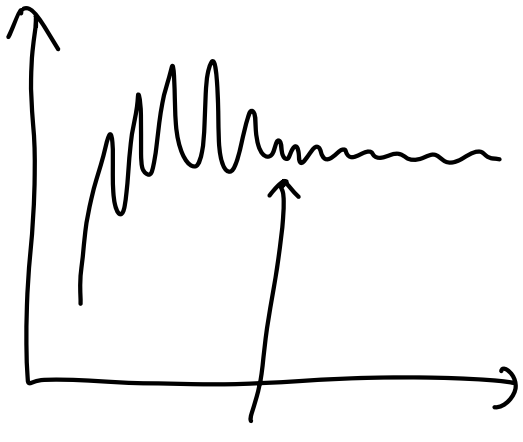
then propose $\theta'_{-k} = \theta_{-k}$ and update $\theta'_k \sim q_k(\cdot|\theta)$

$|\theta' - \theta|$ small, $p(\theta') \approx p(\theta)$.

Smaller changes will typically have a higher acceptance probability. The chain takes more steps to explore the support of the target density in Ω







convergence time.

Example, mixture of two bivariate normals.

$$\pi(\theta) = (2\pi)^{-1}\left[0.5\, e^{-(\theta-\mu_1)\Sigma_1^{-1}(\theta-\mu_1)/2} + 0.5\, e^{-(\theta-\mu_2)\Sigma_2^{-1}(\theta-\mu_2)/2}\right]$$

$\theta = (\theta_1, \theta_2)$. $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ known.

$\left[\begin{smallmatrix} 1 & 2 \\ 1 & 2 \end{smallmatrix}\right]$

proposal distribution.

$\theta_i' \sim U(\theta_i - a, \theta_i + a)$

$q(\theta_1' | \theta_1) = \frac{1}{2a}$

$\left[\begin{smallmatrix} \theta \\ \theta \end{smallmatrix}\right]$

$q(\theta' | \theta) = \left(\frac{1}{2a}\right)^2 = \frac{1}{4a^2}$

$\theta^{(n)} = (\theta_1, \theta_2)$

1. for $i = 1, 2$ simulate $\theta_i' \sim U(\theta_i - a, \theta_i + a)$ and $u \sim \text{Unif}(0,1)$

2. If $u \le \alpha(\theta' | \theta)$,

$$\alpha(\theta' | \theta) = \min\left\{1, \frac{\pi(\theta')\, q(\theta | \theta')}{\pi(\theta)\, q(\theta' | \theta)}\right\}$$

$$= \min\left\{1, \frac{\pi(\theta')}{\pi(\theta)}\right\}$$

比较大

$\theta^{(n+1)} = \theta'$, otherwise $\theta^{(n+1)} = \theta$.

$\left(\theta^{(n)} = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\right)$

MH example: an equal mixture of bivariate normals

```
f<-function(x,mu1,mu2,S1i,S2i,p1=0.5) {
  c1<-exp(-t(x-mu1)%*%S1i%*%(x-mu1))
  c2<-exp(-t(x-mu2)%*%S2i%*%(x-mu2))
  return(p1*c1+(1-p1)*c2)
}
a=3; n=2000;mu1=c(1,1);mu2=c(4,4);
S=diag(2);
S1i=S2i=solve(S);
X=matrix(NA,2,n);
X[,1]=x=mu1 (initial value)
for (t in 1:(n-1)) {
  y<-x+(2*runif(2)-1)*a
  MHR<-f(y,mu1,mu2,S1i,S2i)/f(x,mu1,mu2,S1i,S2i)
  if (runif(1)<MHR) x<-y
  X[,t+1]<-x
}
```

$y = x_1 + [2\,U(0,1) - 1]\,a$.

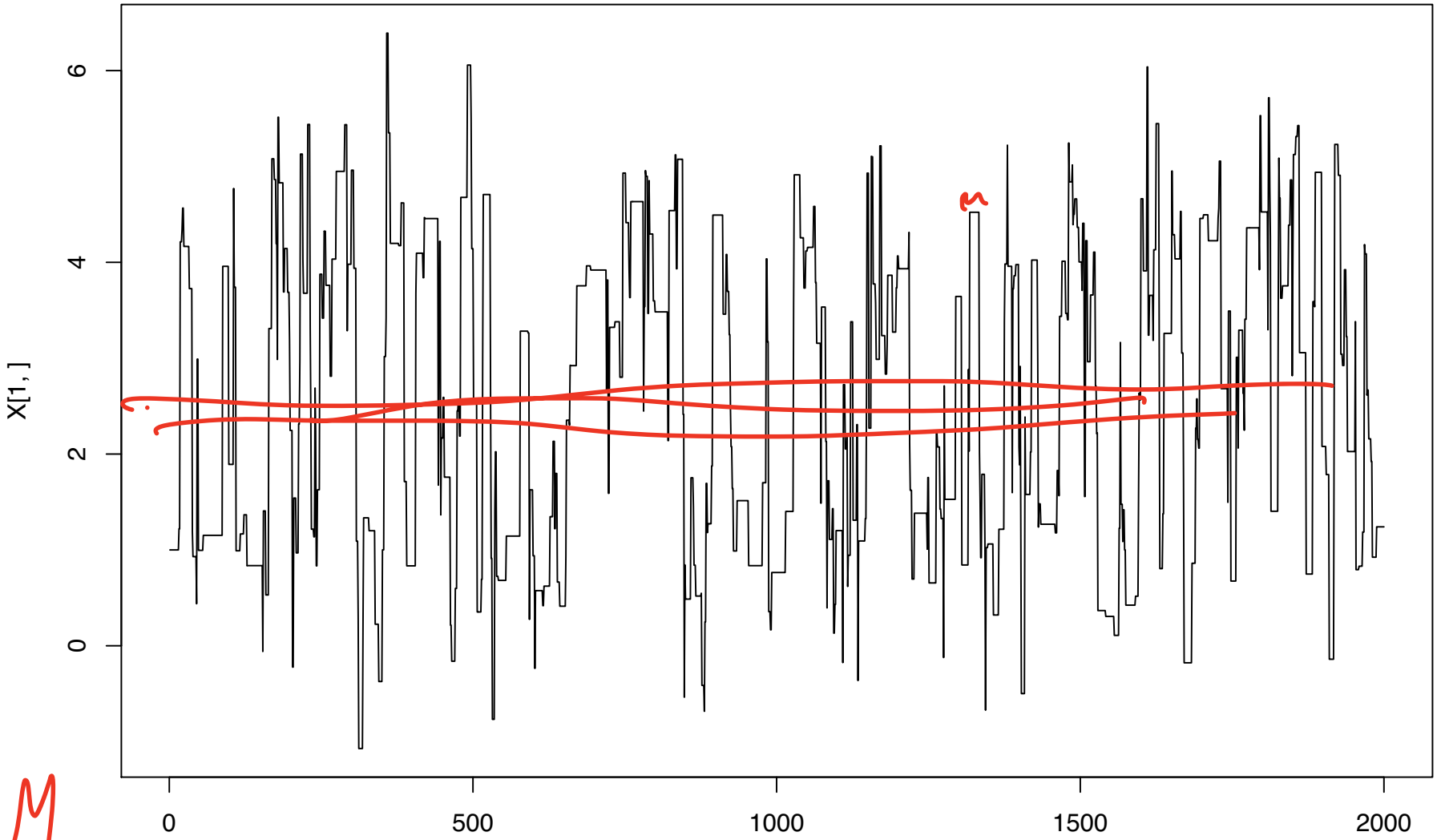
$= x_1 + [\,U(0,2) - 1\,]\,a$

$= x_1 + U(-1,1)\,a$

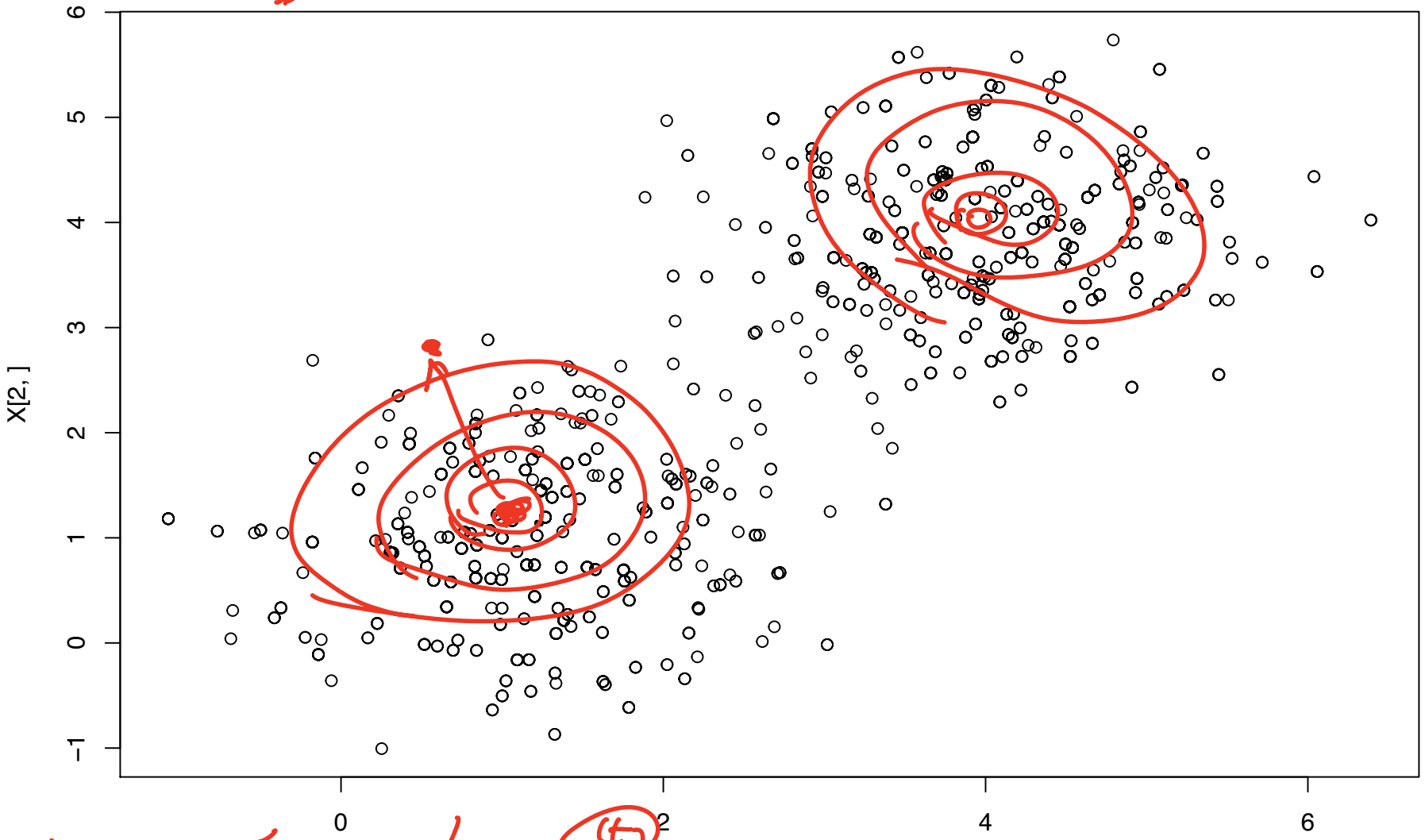
$= U(x_1 - a,\ x_1 + a)$

$X = \begin{pmatrix} & \\ & \end{pmatrix} \begin{matrix} \theta_1 \\ \theta_2 \end{matrix}$

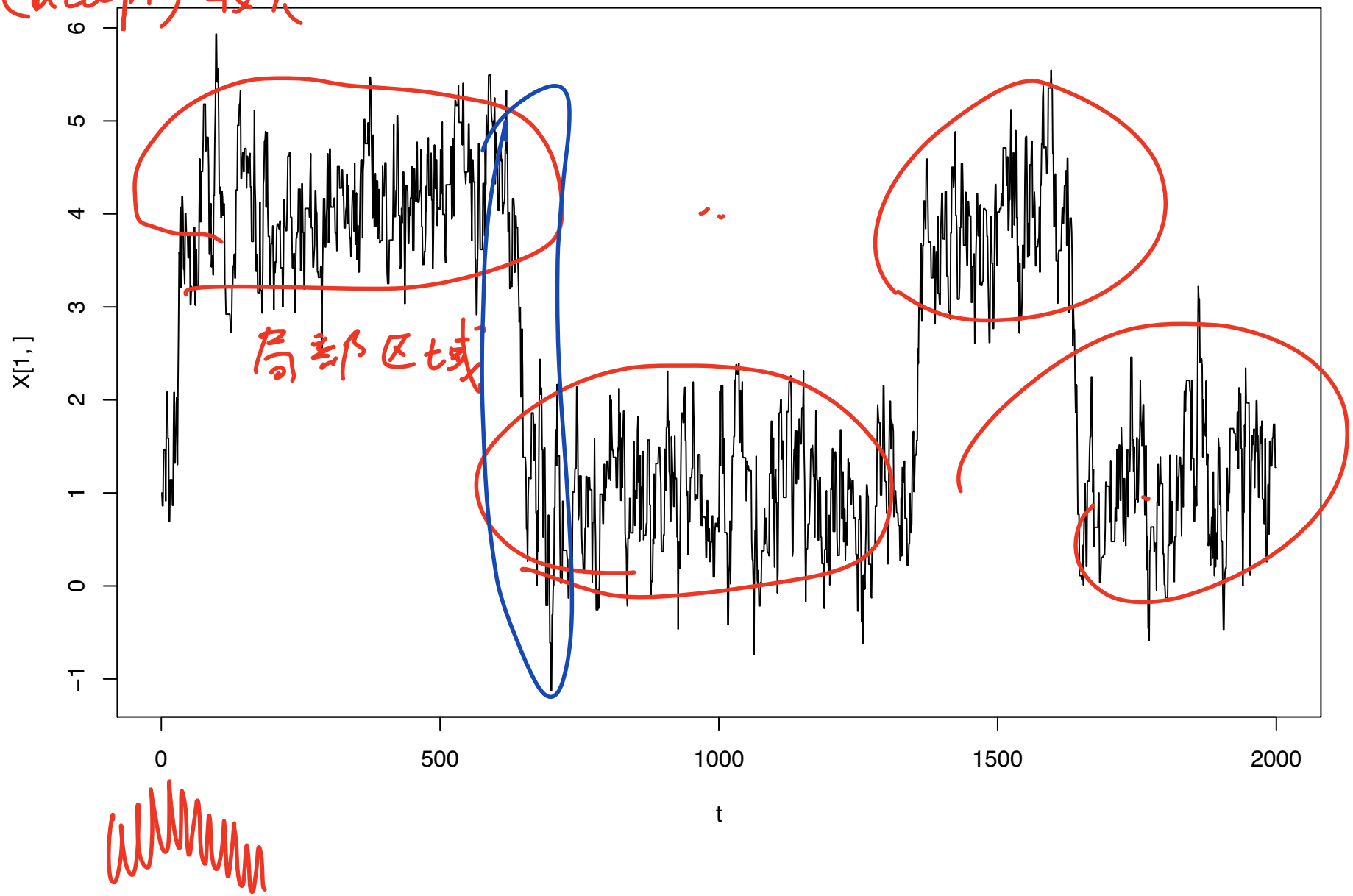
path of θ_1, $a = 3$, $n = 2000$



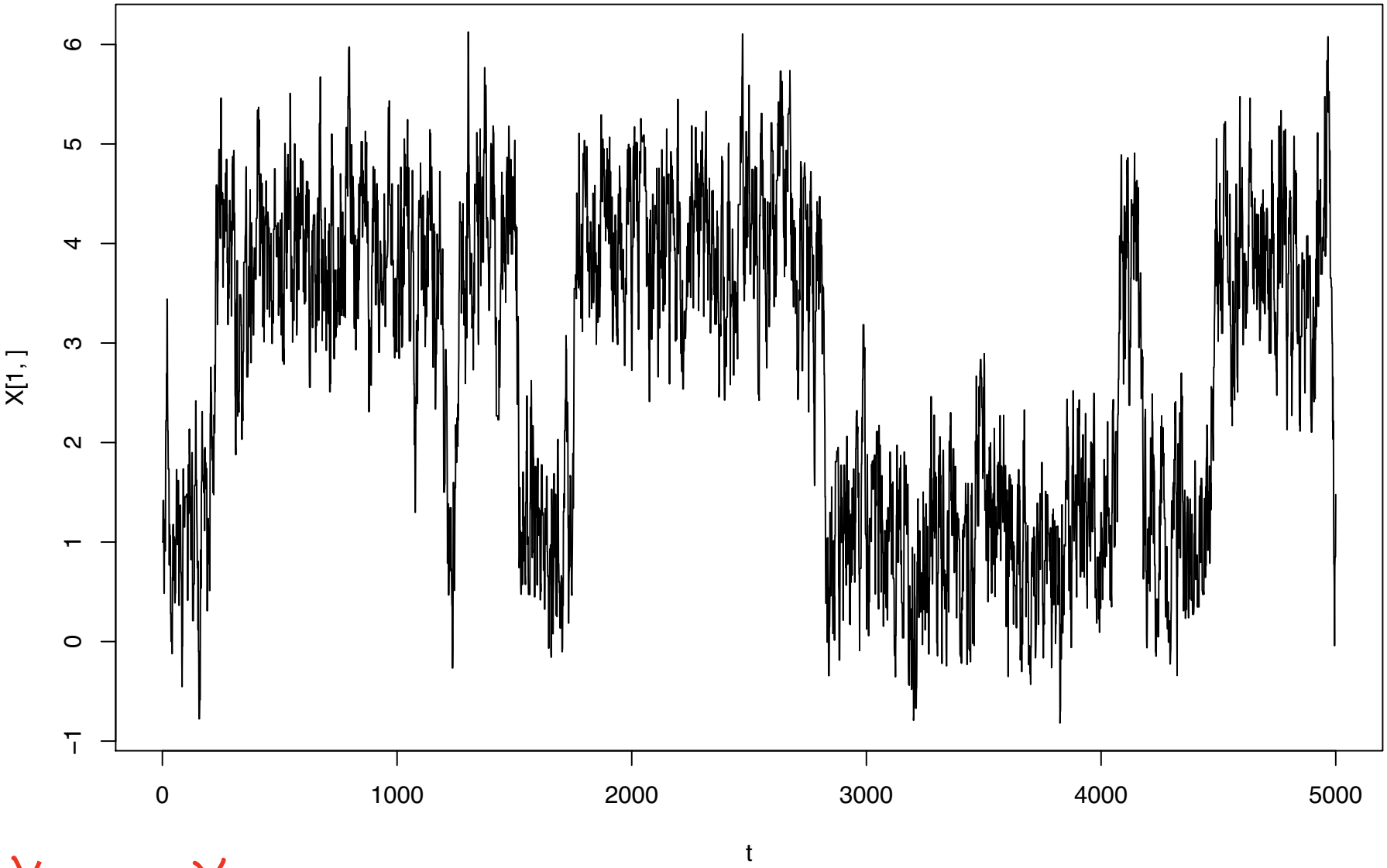
$\theta_2 \sim \theta_1, a = 3, n = 2000$



path of $\theta_1, a = 1, n = 2000$

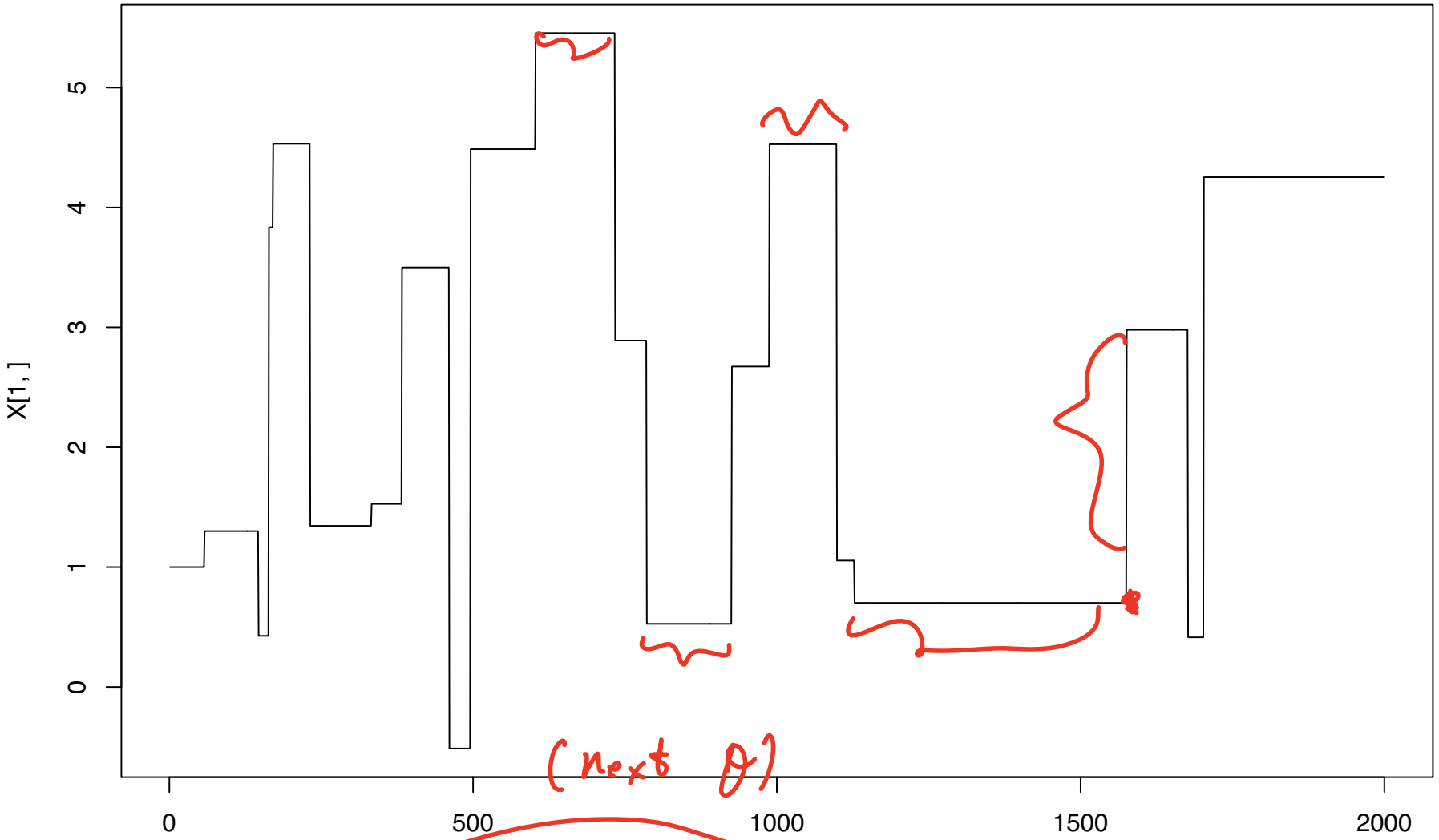


path of θ_1, $a = 1$, $n = 5000$



$X_t \neq X_{t+1}$

path of θ_1, $a = 20$, $n = 2000$ P(accept) 较小



It will make many proposals into very low density states in the tail of the density, which will

be rejected.

Initial state. (sample) from target distribution

$pP = p$.

(check 收敛的必要条件，充分条件不行)

Convergence and mixing.

MCMC samples $X_0 \cdots X_n \sim p(x)$

$\bar{f}_n = \frac{\sum_t f(x_t)}{n}$ $\xrightarrow{a.s.}$ $E_p[f(x)]$

bias, variance.

Initial state. (大概率不是 Stationary State

State in equilibrium)

$p^{(t)} \simeq p$ for $t \geq T$.

correlated

If $n \to \infty$, the retained samples are representative of P.

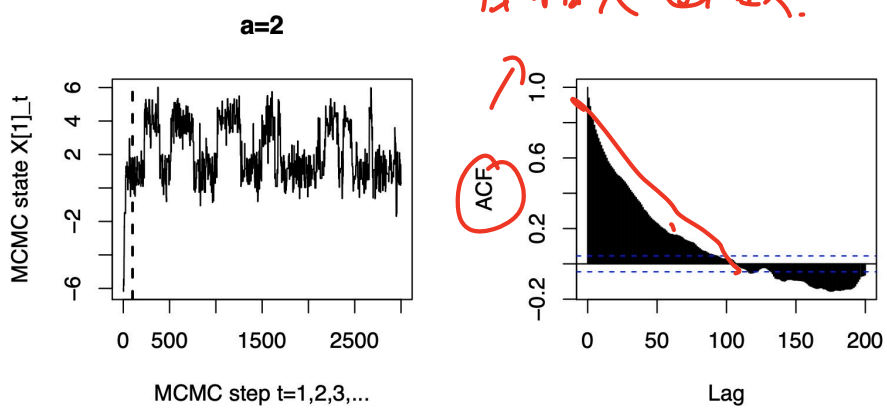
Second, assume start the chain in equilibrium,

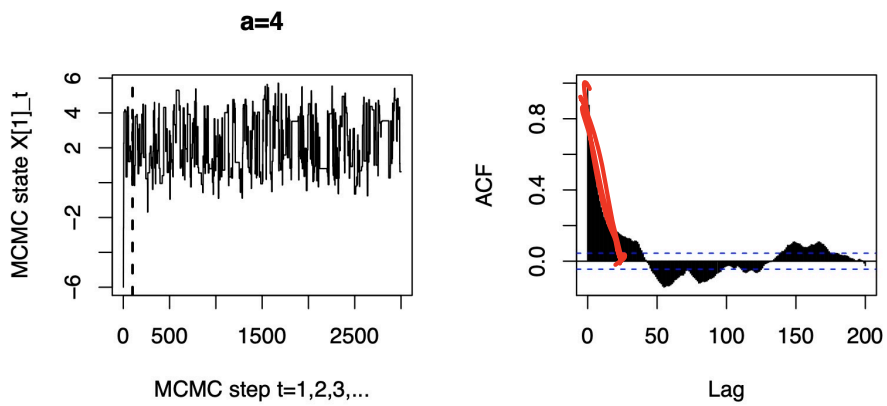
$Var(\bar{f}_n) \downarrow$ as $n \uparrow$, $\bar{f}_n$ has useful precision.

$Var(\bar{f}_n)$ calculation is complex as the MCMC samples are correlated.

自相关函数.

At X_{t+r}.





——— (no serial correlation)

$$\mathrm{Var}(\bar{f}_n) = \frac{\mathrm{Var}(f(X))}{ESS} \qquad X \sim p.$$

effective sample size.

(The number of independent samples which would give the same variance reduction as our n correlated samples!)

$ESS \ll n$.

$ESS = n$ (MCMC samples were independent.)

$\hat{\sigma}_f^2$ is the estimate of $\mathrm{Var}(f(x)) = (\sigma_f^2)$

Let $\sigma_{f,n}^2 = \mathrm{Var}(\bar{f}_n)$, $ESS = \dfrac{\sigma_f^2}{\hat{\sigma}_{f,n}^2}$.

Simulate $K \uparrow$ Markov chain, length $= n$.

$\bar{f}_{1,n}, \cdots \bar{f}_{k,n}$ ##

Sample $\theta^{(k,t)}$, $k = 1 \cdots K$.

$t = 1, \cdots n$.

$$\bar{f}_{k,n} = \frac{\sum_t f(\theta^{(k,t)})}{n}$$

$$\text{Var}(\bar{f}_n) = \hat{\sigma}^2_{f,n}$$

$$= \frac{1}{k-1} \sum_{k=1}^{K} \left(\bar{f}_{k,n} - \frac{\sum_{j=1}^{K} \bar{f}_{j,n}}{K} \right)^2.$$

$$ESS = \frac{\hat{\sigma}_f^2}{\hat{\sigma}^2_{f,n}}$$ (precision gain)

fix n the same, max ESS $\Rightarrow$ algorithm.

equilibrium

舍弃. $n = 1000$

前面都 drop burn-in

$$\rho_s = \frac{Cov\left(f(x_t), f(x_{t+s})\right)}{\sqrt{Var\, f(x_t)}\sqrt{Var\, f(x_{t+s})}}$$

$$= \frac{Cov\left(f(x_t), f(x_{t+s})\right)}{Var\, f(x_t)}$$

$$Var(\bar{f}_n) = Var\left(\frac{\sum_{t=1}^{n} f(x_t)}{n}\right)$$

$$= \frac{Var\left(f(x_1) + \cdots + f(x_n)\right)}{n^2}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(f(x_i), f(x_j)\right)}{n^2}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{|i-j|}\cdot\sigma^2$$

$$\sum_{i=1}^{n}\sum_{j=i}^{n} + \sum_{i=1}^{n}\sum_{j>i}^{n} + \left(\sum_{i=1}^{n}\sum_{j=i}^{n}\right)$$

n.

$$= \frac{\sigma^2}{n^2}\left(n + 2\sum_{i=1}^{n}\sum_{i>j}\rho_{i-j}\right)$$

$$= \frac{\sigma^2}{n^2}\left(n + 2\sum_{s=1}^{n-1}(n-s)\rho_s\right)$$

$$= \frac{\sigma^2}{n}\left(1 + 2\sum_{s=1}^{n-1}\left(1-\frac{s}{n}\right)\rho_s\right)$$

$$= \frac{\sigma^2}{n}\left(1 + 2\sum_{s=1}^{n-1}\rho_s - \sum_{s=1}^{n-1}s\rho_s\right)$$

$$= \frac{\sigma^2}{n}\left(1 + 2\sum_{s=1}^{n-1}\rho_s\right) - \frac{\sigma^2\sum_{s=1}^{n-1}s\rho_s}{n^2}$$

$$\simeq \frac{\sigma^2}{n}\left(1 + 2\sum_{s=1}^{n-1}\rho_s\right)$$

$$\frac{\sigma^2}{n}\cdot J_f$$

If $\rho_s \to 0$ rapidly at large s

$$\sum_{i=1}^{n}\sum_{i>j}\rho_{i-j}$$

$$\rho_1 + \rho_2 + \rho_1 + \rho_3 + \rho_2 + \rho_1 + \cdots$$

$$\sum_{s=1}^{n-1}(n-s)\rho_s$$

$$(n-1)\rho_1 + (n-2)\rho_2 + \cdots + \rho_{n-1}$$

$$ESS = \frac{\mathrm{Var}(f(x))}{\mathrm{Var}(\bar{f}_n)} = \frac{\sigma^2}{\frac{\sigma^2}{n}\cdot J_f} = \frac{n}{J_f}$$

Sample correlation estimate $\hat{\rho}_s = \frac{\hat{r}_s}{\hat{r}_0}$

$$\hat{\tau}_f = 1 + 2 \sum_{s=1}^{M} \hat{\rho}_s$$

$$1 + 2 \sum_{s=1}^{n-1} \hat{\rho}_s$$

① 在 s 后, $\hat{\rho}_s$ 本来就很小

② M 后 noise

Geyer, for a markov chain:

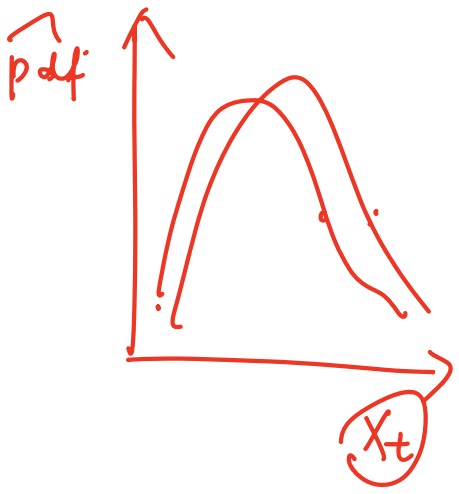
$\Gamma_t + \Gamma_{t+1}$ is positive, monotone, convex

choose M equal to the least s satisfying

$\hat{\Gamma}_s + \hat{\Gamma}_{s+1} > 0$ and $\hat{\Gamma}_s + \hat{\Gamma}_{s+1} > \hat{\Gamma}_{s-1} + \hat{\Gamma}_s$

① start at different initial states.

check marginal distributions agree?



② ACF plot. 看 correlation falls rapidly?

Calculate ESS.

③ plot MCMC trace.