

Advanced Statistical Computing

Fall 2018

Steve Qin

Markov Chain Monte Carlo

- The goal is to generate a sequence of random samples from an arbitrary probability density function (usually high dimensional),
- The sequence of random samples form a Markov chain,
- The purpose is simulation (Monte Carlo).

What is Monte Carlo



What is Monte Carlo?

- Rely on repeated sampling to study the results of a experiment or study the properties of certain procedure.
 - Often used in complex and uncertain scenarios
 - Difficult to formulate, high correlation.
 - Cheap
 - Take advantage of faster computers
- History
 - John von Neumann, Stanislaw Ulam, Nicholas Metropolis

Applications of Monte Carlo

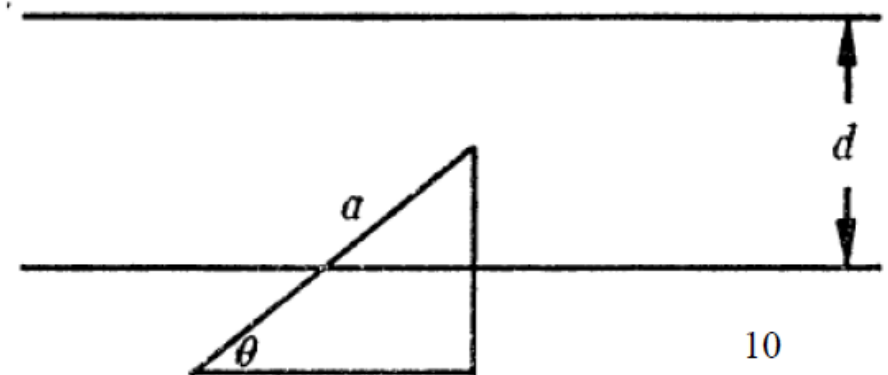
- Optimization
- Numerical integration
- Generate random samples

Buffon's needle

Georges-Louis Leclerc, Comte de Buffon
(1707-1788)



Given a needle of length a and an infinite grid of parallel lines with common distance d between them, what is the probability $P(E)$ that a needle, tossed at the grid randomly, will cross one of the parallel lines?



Buffon's needle

- Assume $a < d$

$$P(E) = \int_0^\pi \frac{a \sin \theta d\theta}{\pi d} = (a/\pi d) \int_0^\pi \sin \theta d\theta = 2a/\pi d.$$

<http://web.student.tuwien.ac.at/~e9527412/buffon.html>

Motivation

- Generate *iid* r.v. from high-dimensional arbitrary distributions is extremely difficult.
- Drop the “*independent*” requirement.
- How about also drop the “*identically distributed*” requirement?

Markov chain

- Assume a finite state, discrete Markov chain with N different states.

- Random process X_n , $n = 0, 1, 2, \dots$

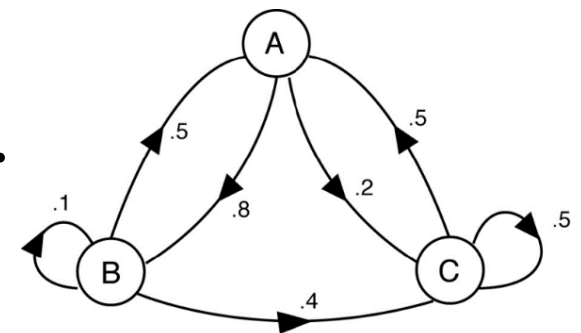
$$x_n \in S = \{1, 2, \dots, N\}$$

- Markov property,

$$P(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n)$$

- Time-homogeneous
- Order

– Future state depends on the past m states.



Markov graph of transition probabilities between states A, B and C

Key parameters

- Transition matrix

$$P(X_n = j \mid X_{n-1} = i) = p(i, j),$$

$$P = \{p(i, j)\}.$$

- Initial probability distribution $\pi^{(0)}$

$$\pi^{(n)}(i) = P(x_n = i).$$

- Stationary distribution (invariant/equilibrium)

$$\pi = \pi P.$$

Reducibility

- A state j is accessible from state i (written $i \rightarrow j$) if $P(X_n = j \mid X_0 = i) = p_{ij}^{(n)} > 0$.
- A Markov chain is *irreducible* if it is possible to get to any state from any state.

Recurrence

- A state i is *transient* if given that we start in state i , there is a non-zero probability that we will never return to i . State i is *recurrent* if it is not transient.

Ergodicity

- A state i is *ergodic* if it is aperiodic and positive recurrent. If all states in an irreducible Markov chain are ergodic, the chain is *ergodic*.

Reversible Markov chains

- Consider an ergodic Markov chain that converges to an invariant distribution π . A Markov chain is *reversible* if for all $x, y \in S$,

$$\pi(x)p(x, y) = \pi(y)p(y, x).$$

which is known as the detailed balance equation.

- An ergodic chain in equilibrium and satisfying the detailed balance condition has π as its unique stationary distribution.

Markov Chain Monte Carlo

- The goal is to generate a sequence of random samples from an arbitrary probability density function (usually high dimensional),
- The sequence of random samples form a Markov chain,

in Markov chain, $P \rightarrow \pi$

in MCMC, $\pi \rightarrow P$.

Examples

$$P(X | E, \beta, \sigma^2) \propto \prod_{k=1}^K \prod_{E(i)=k} \prod_{j=1}^M \left((\sigma_{kj}^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_{kj}^2} (x_{ij} - \beta_{kj})^2} \right)$$

$$P(X | E) = \prod_{k=1}^K \prod_{j=1}^M \iint \prod_{E(i)=k} P(x_{ij} | \beta_{kj}, \sigma_{kj}^2) p(\beta_{kj} | \beta_0, \sigma_{kj}^2) p(\sigma_{kj}^2) d\beta_{kj} d\sigma_{kj}^2$$

$$= \prod_{k=1}^K \prod_{j=1}^M \left[\frac{b^a}{\Gamma(a)} \frac{(2\pi)^{-\frac{n_k}{2}}}{\sqrt{n_k+1}} \frac{\Gamma\left(\frac{n_k}{2} + a\right)}{\left(b + \frac{1}{2} \left(\sum_{E(i)=k} x_{ij}^2 + \beta_0^2 - \frac{\left(\sum_{E(i)=k} x_{ij} + \beta_0 \right)^2}{n_k + 1} \right) \right)^{\frac{n_k}{2} + a}} \right]$$

Bayesian Inference

Genotype $Y = (y_1, \dots, y_n)$

Haplotype $Z = (z_1, \dots, z_n)$

Frequency $\Theta = (\theta_1, \dots, \theta_m)$

Prior $\theta \sim \text{Dirichlet}(\beta)$

$$P(Y, Z, \Theta) = \prod_{i=1}^n \theta_{z_{i1}} \theta_{z_{i2}} \prod_{g=1}^m \theta_g^{\beta_g - 1}$$

History

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953).

Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Metropolis algorithm

- Direct sampling from the target distribution is difficult,
- Generating candidate draws from a proposal distribution,
- These draws then “corrected” so that asymptotically, they can be viewed as random samples from the desired target distribution.

Pseudo code

- Initialize X_0 ,
- Repeat
 - Sample $Y \sim q(x,.)$,
 - Sample $U \sim \text{Uniform}(0,1)$,
 - If $U \leq \alpha(X,Y)$, set $X_i = y$,
 - Otherwise $X_i = x$.

An informal derivation

- Find $\alpha(X, Y)$:
- Joint density of current Markov chain state and the proposal is $g(x, y) = q(x, y)\pi(x)$
- Suppose q satisfies detail balance
$$q(x, y) \pi(x) = q(y, x) \pi(y)$$
- If $q(x, y) \pi(x) > q(y, x) \pi(y)$, introduce
$$\alpha(x, y) < 1 \text{ and } \alpha(y, x) = 1$$
hence $\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}$.
- If $q(x, y) \pi(x) > q(y, x) \pi(y)$, ...
- The probability of acceptance is $\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$.

Metropolis-Hastings Algorithm

- Start with any $X^{(0)}=x_0$, and a “*proposal chain*” $T(x,y)$
- Suppose $X^{(t)}=x_t$. At time $t+1$,
 - **Draw** $y \sim T(x_t, y)$ (i.e., propose a move for the next step)
 - Compute “*goodness ratio*”

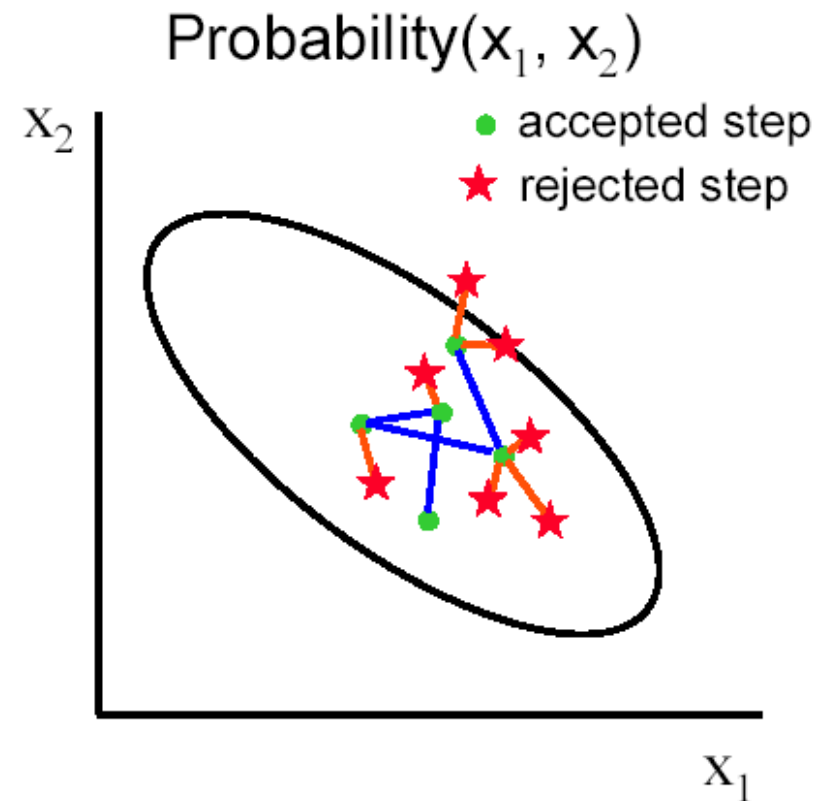
$$r = \frac{\pi(y)T(y, x_t)}{\pi(x_t)T(x_t, y)}$$

- **Acceptance/Rejection decision:** Let

$$X^{(t+1)} = \begin{cases} y, & \text{with } p = \min\{1, r\} \\ x_t, & \text{with } 1 - p \end{cases}$$

Remarks

- Relies only on calculation of target pdf up to a normalizing constant.



Remarks

- How to choose a good proposal function is crucial.
- Sometimes tuning is needed.
 - Rule of thumb: 30% acceptance rate
- Convergence is slow for high dimensional cases.

Illustration of Metropolis-Hastings

- Suppose we try to sample from a bi-variate normal distributions.

$$N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

- Start from $(0, 0)$

- Proposed move at each step is a two dimensional random walk

$$x_{t+1} = x_t + s \cos \theta$$

$$y_{t+1} = y_t + s \sin \theta$$

with

$$s \sim U(0,1)$$

$$\theta \sim U(0,2\pi)$$

Illustration of Metropolis-Hastings

- At each step, calculate $r = \frac{\pi(x_{t+1}, y_{t+1})}{\pi(x_t, y_t)}$

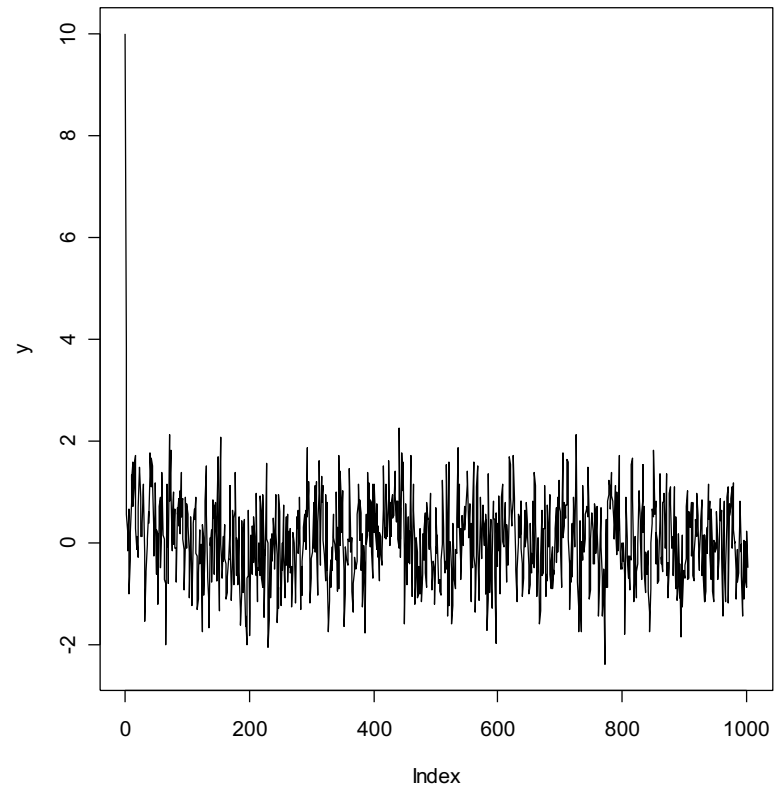
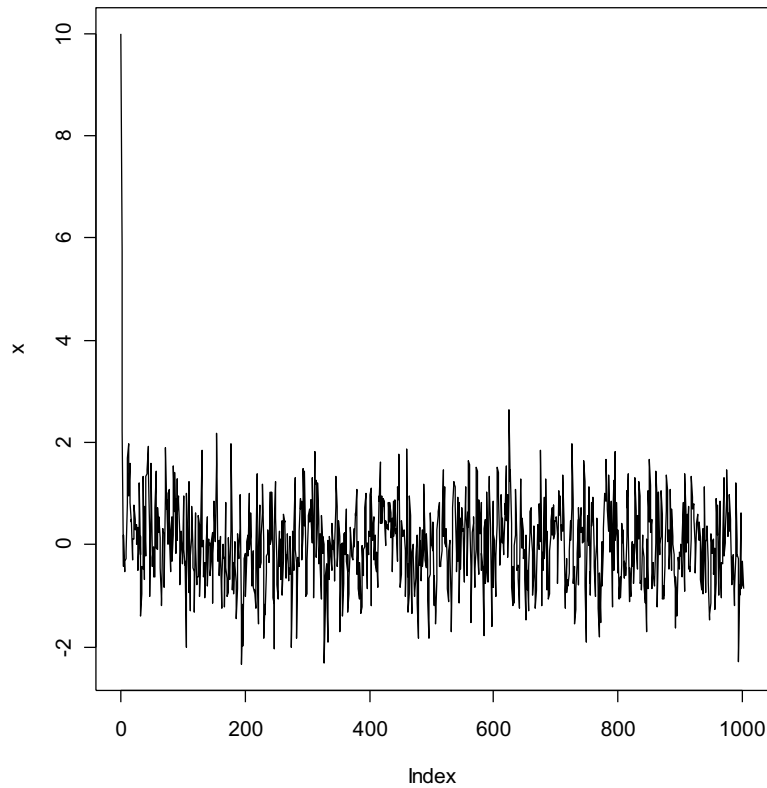
$$T((x_t, y_t), (x_{t+1}, y_{t+1})) = T((x_{t+1}, y_{t+1}), (x_t, y_t)) = 1/\pi^2$$

since

$$\begin{aligned} r &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x_{t+1}^2 - 2\rho x_{t+1}y_{t+1} + y_{t+1}^2)\right)}{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x_t^2 - 2\rho x_t y_t + y_t^2)\right)} \\ &= \exp\left(-\frac{1}{2(1-\rho^2)}\left((x_{t+1}^2 - 2\rho x_{t+1}y_{t+1} + y_{t+1}^2) - (x_t^2 - 2\rho x_t y_t + y_t^2)\right)\right) \end{aligned}$$

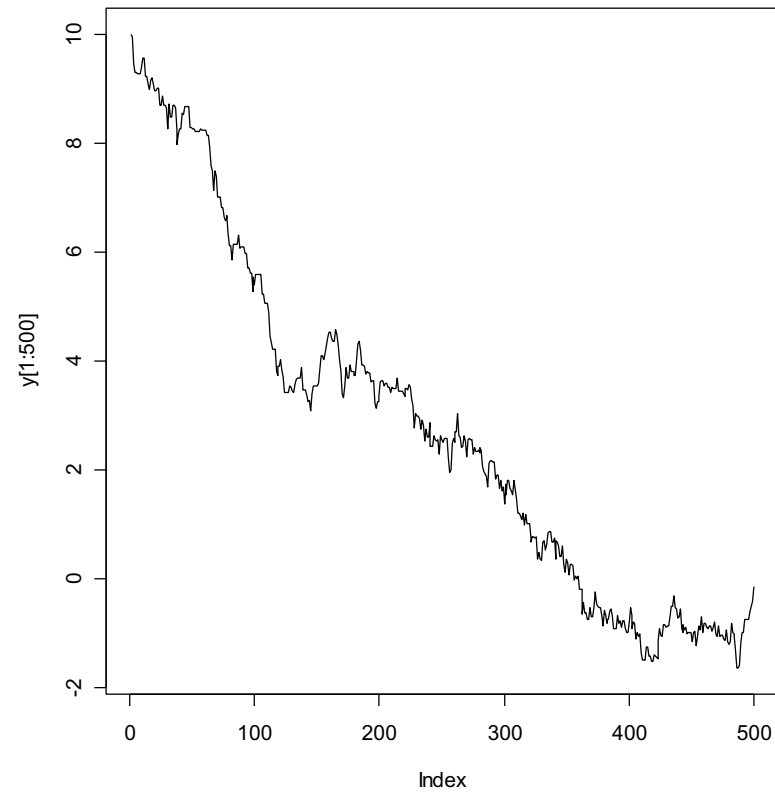
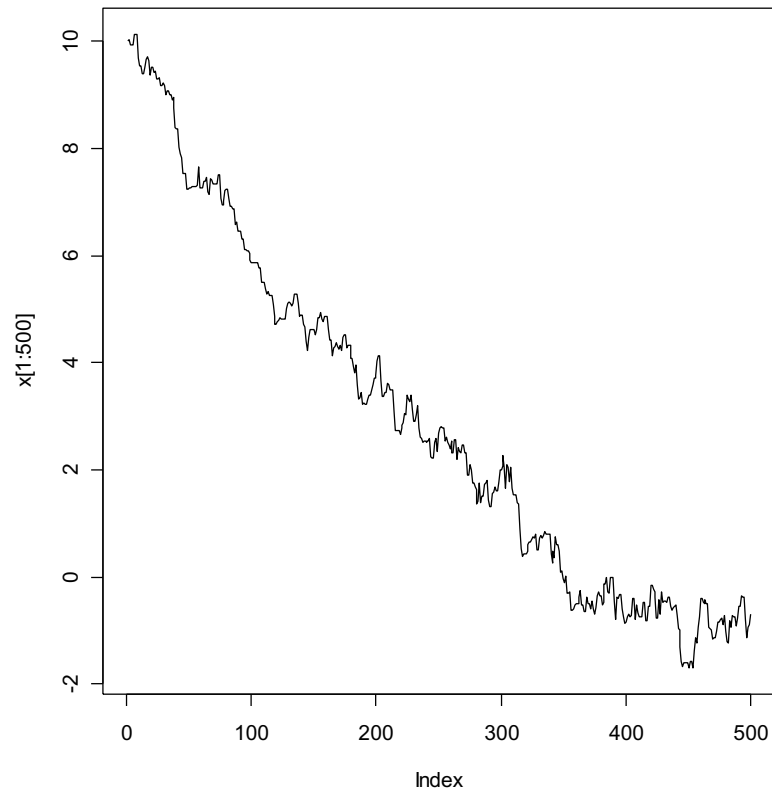
Convergence

- Trace plot Good



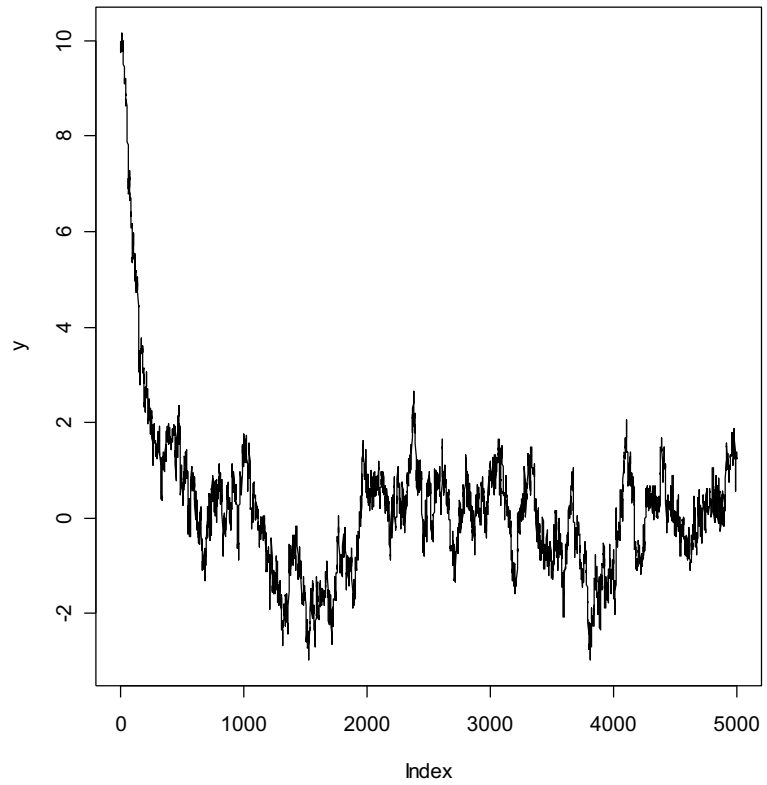
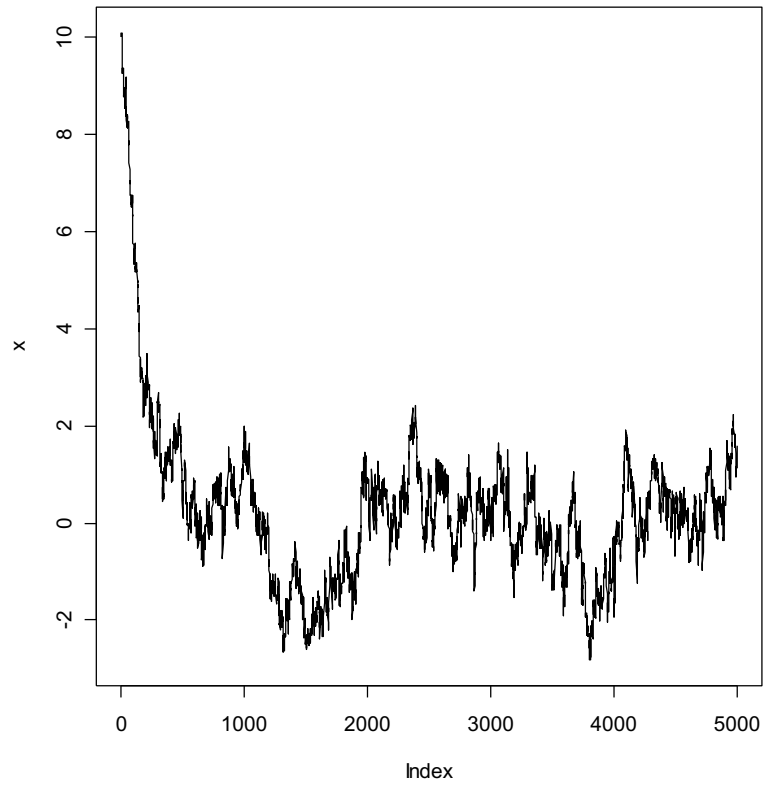
Convergence

- Trace plot Bad



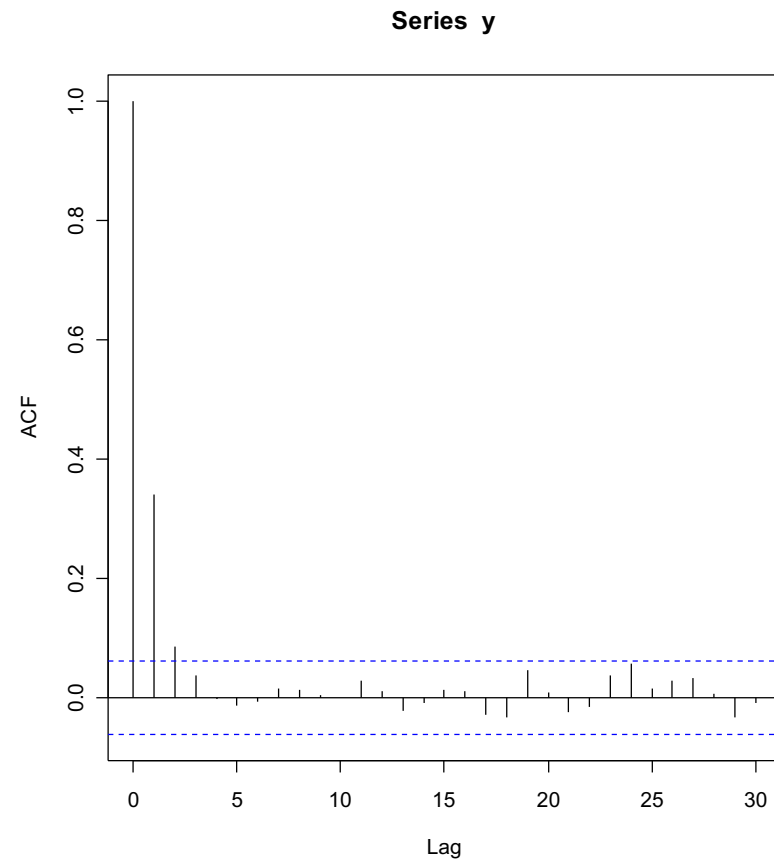
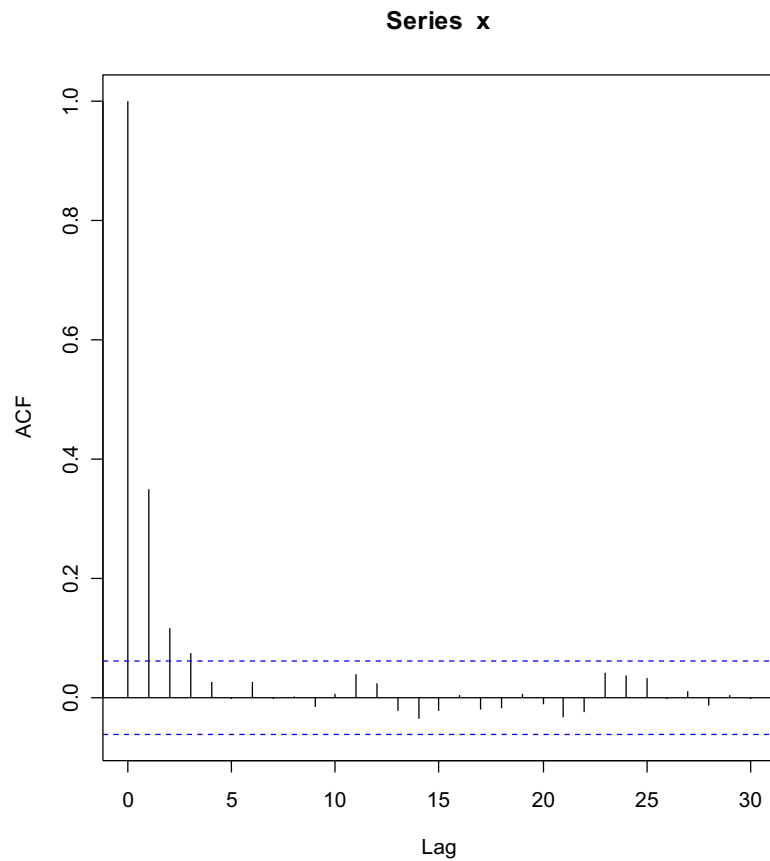
Convergence

- Trace plot



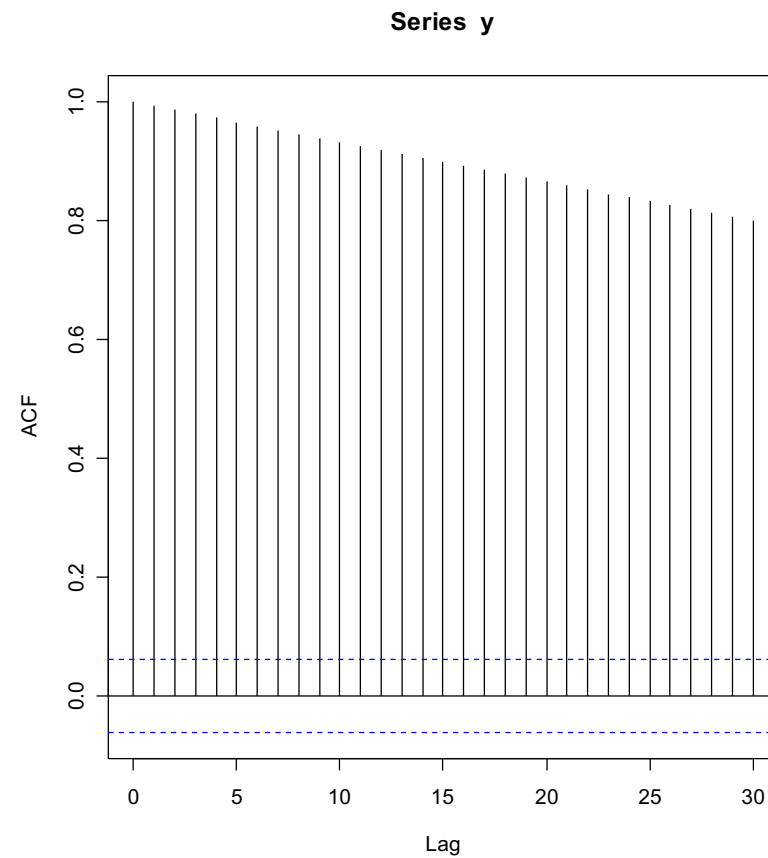
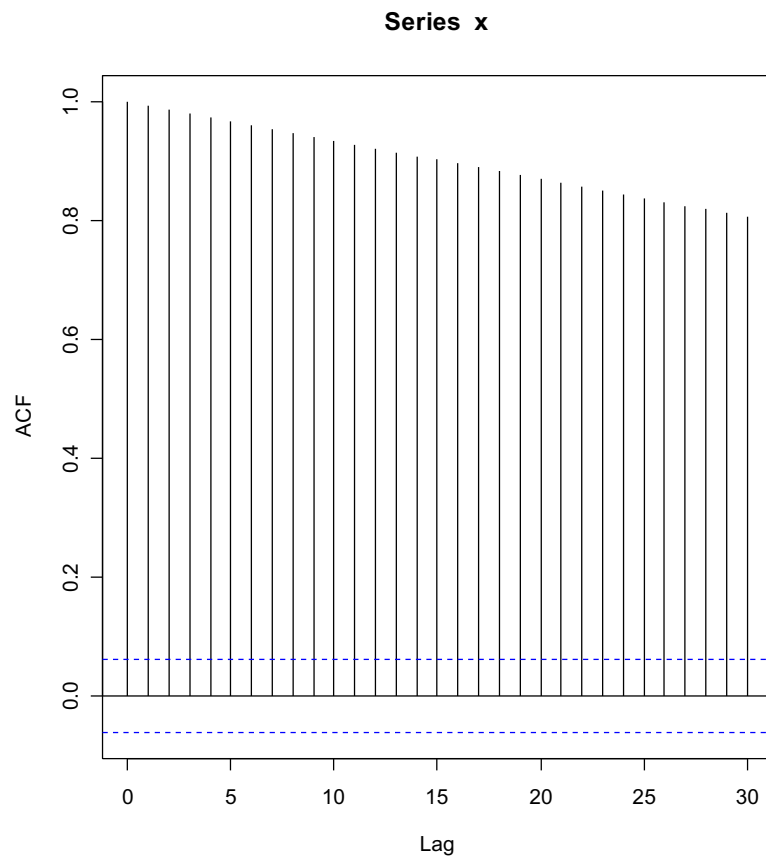
Convergence

- Autocorrelation plot Good



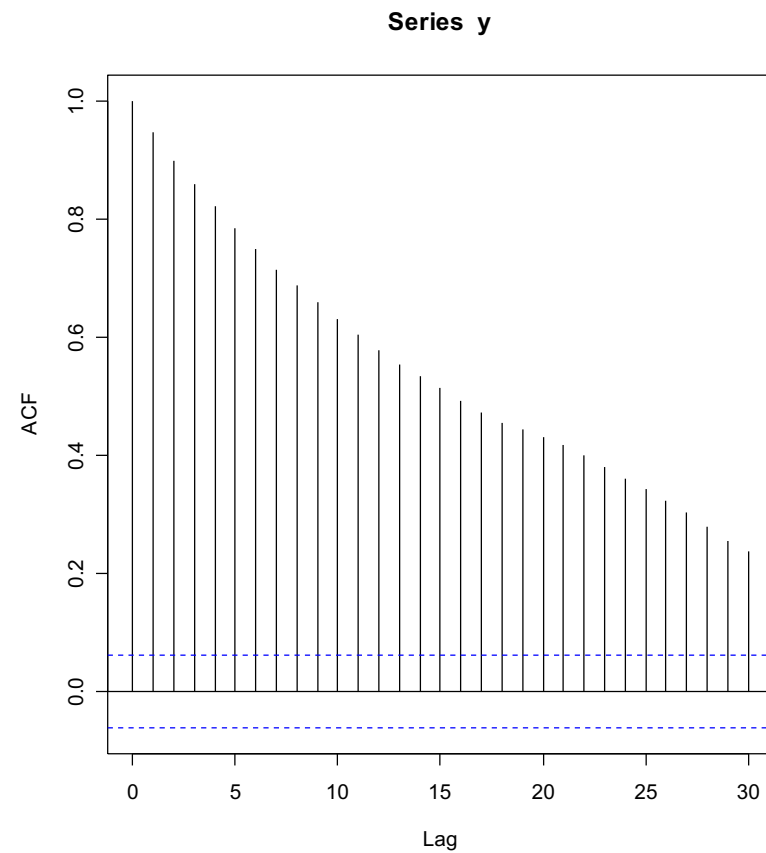
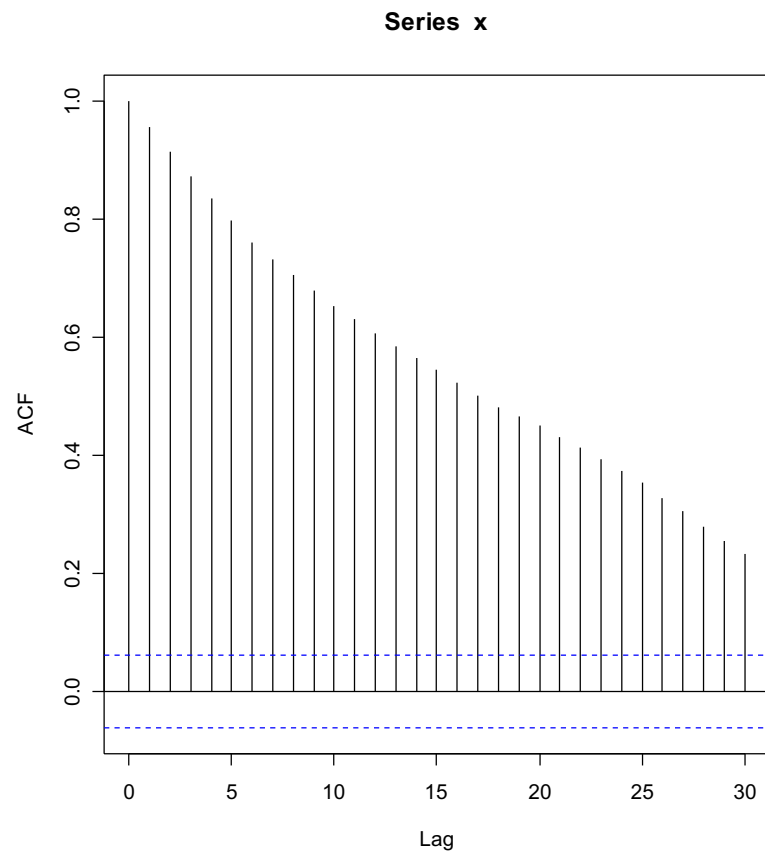
Convergence

- Autocorrelation plot Bad $s = 0.5$



Convergence

- Autocorrelation plot Okay $s = 3.0$



References

- Metropolis et al. 1953,
- Hastings 1973,
- Tutorial paper:
Chib and Greenberg (1995). Understanding the Metropolis--Hastings Algorithm. *The American Statistician* **49**, 327-335.

Gibbs Sampler

- **Purpose:** Draw random samples from a joint distribution (high dimensional)

$$x = (x_1, x_2, \dots, x_n) \text{ Target } \pi(x)$$

- **Method:** Iterative conditional sampling

$$\forall i, \text{ draw } x_i \sim \pi(x_i \mid x_{[-i]})$$

Illustration of Gibbs Sampler

- Suppose the target distribution is:

$$(X, Y) \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

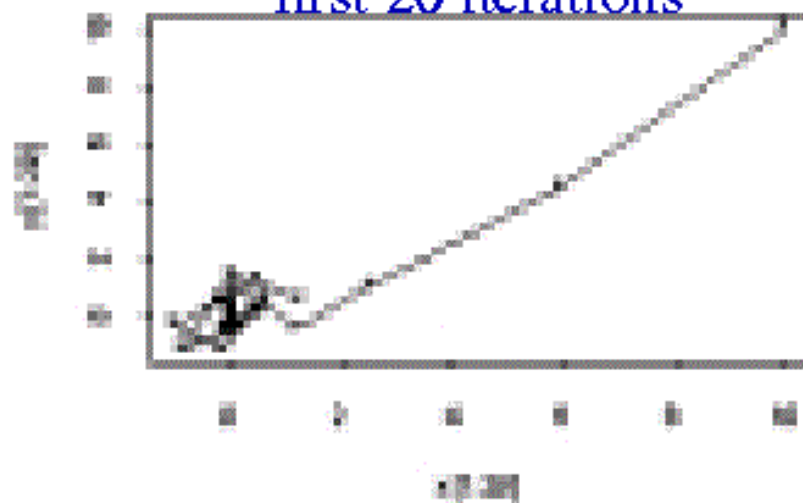
- Gibbs sampler:

$$[X|Y = y] \sim N(\rho y, 1 - \rho^2)$$

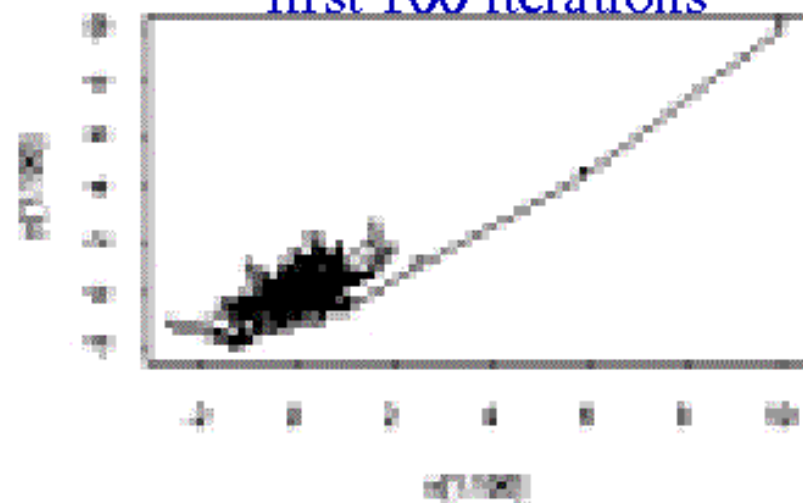
$$[Y|X = x] \sim N(\rho x, 1 - \rho^2)$$

Start from, say, $(X, Y) = (10, 10)$, we can take a look at the trajectories. We took $\rho = 0.6$.

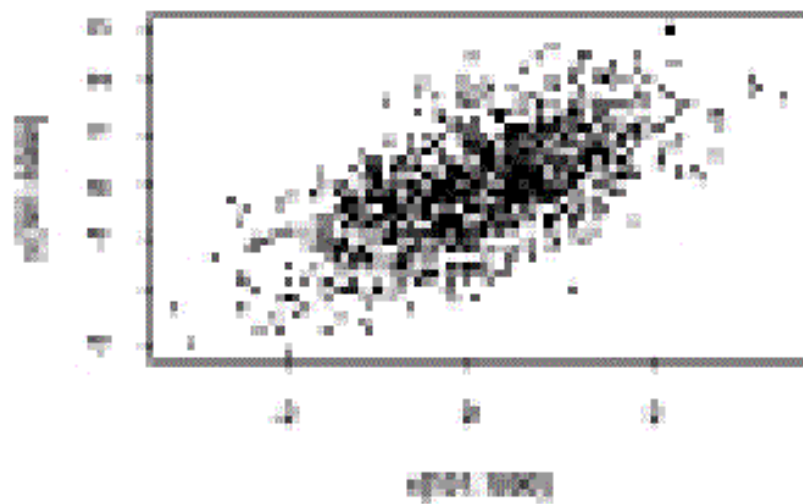
first 20 iterations



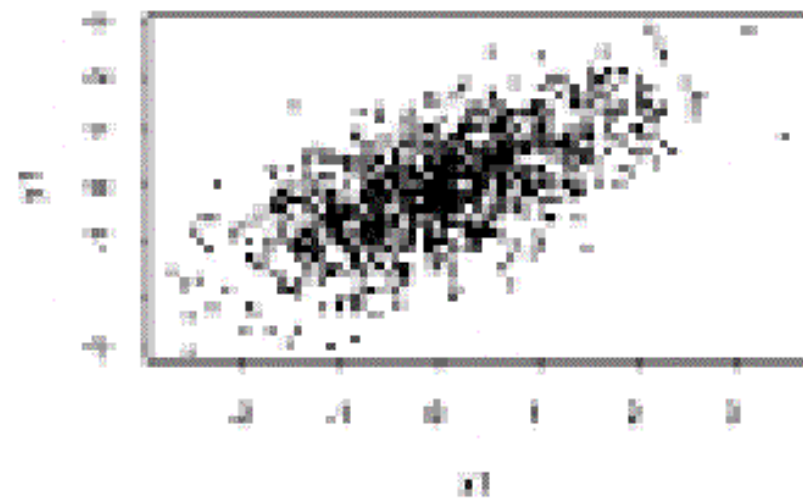
first 100 iterations



101-1000 iterations



900 iid samples



References

- Geman and Geman 1984,
- Gelfand and Smith 1990,
- Tutorial paper:
Casella and George (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

Remarks

- Gibbs Sampler is a special case of Metropolis-Hastings
- Compare to EM algorithm, Gibbs sampler and Metropolis-Hastings are stochastic procedures
- Verify convergence of the sequence
- Require Burn in
- Use multiple chains

References

- Geman and Geman 1984,
- Gelfand and Smith 1990,
- Tutorial paper:
Casella and George (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.