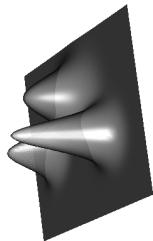
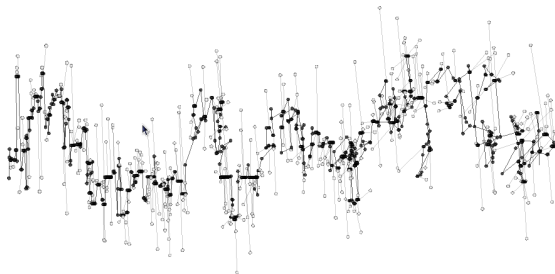


Markov Chains and Monte Carlo Methods

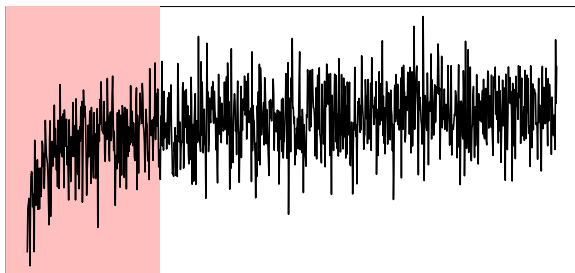
Chapter 6: Convergence diagnostics



Ioana Cosma
March 10, 2010

Practical considerations: Burn-in period

- Theory (ergodic theorems) allows for the use of the entire chain $(\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots)$ to approximate $\mathbb{E}_f(h(\mathbf{X}))$.
- However distribution of $(\mathbf{X}^{(t)})$ for small t might still be far from the stationary distribution f .
- Can be beneficial to discard the first iterations $\mathbf{X}^{(t)}, t = 1, \dots, T_0$ (**burn-in period**).
- Optimal T_0 depends on mixing properties of the chain.



Practical considerations: Thinning (1)

- MCMC methods typically yield positively correlated chain: $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ large for small τ .
- Idea: build a subchain by only keeping every m -th value: Consider a Markov chain $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$ instead of $(\mathbf{X}^{(t)})_{t=1, \dots, T}$ (**thinning**).
- $(\mathbf{Y}^{(t)})_t$ exhibits less autocorrelation than $(\mathbf{X}^{(t)})_t$, i.e.

$$\rho(\mathbf{Y}^{(t)}, \mathbf{Y}^{(t+\tau)}) = \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(m \cdot t + \tau)}) < \rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)}),$$

if the correlation $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ decreases monotonically in τ .

- Price we have to pay: length of $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ is only $(1/m)$ -th of the length of $(\mathbf{X}^{(t)})_{t=1, \dots, T}$.

Practical considerations: Thinning (2)

- If $\mathbf{X}^{(t)} \sim f$ and corresponding variances exist,

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) \leq \text{Var} \left(\frac{1}{\lfloor T/m \rfloor} \sum_{t=1}^{\lfloor T/m \rfloor} h(\mathbf{Y}^{(t)}) \right),$$

i.e. thinning cannot be justified when objective is estimating $\mathbb{E}_f(h(\mathbf{X}))$.

- Thinning can be a useful concept
 - if computer has insufficient memory.
 - for convergence diagnostics: $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$ is closer to an i.i.d. sample than $(\mathbf{X}^{(t)})_{t=1, \dots, T}$.

The need for convergence diagnostics

- Theory we have studied guarantees (under certain conditions) the convergence of the Markov chain $\mathbf{X}^{(t)}$ to the desired distribution.
- This does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution.
- Validity of the approximation must be confirmed in practise.
- Convergence diagnostics help answering this question (R package CODA).
- Convergence diagnostics are *not* perfect and should be treated with a good amount of scepticism; the techniques presented are **exploratory tools**.

The need for convergence diagnostics

- Theory we have studied guarantees (under certain conditions) the convergence of the Markov chain $\mathbf{X}^{(t)}$ to the desired distribution.
- This does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution.
- Validity of the approximation must be confirmed in practise.
- Convergence diagnostics help answering this question (R package CODA).
- Convergence diagnostics are *not* perfect and should be treated with a good amount of scepticism; the techniques presented are *exploratory tools*.

The need for convergence diagnostics

- Theory we have studied guarantees (under certain conditions) the convergence of the Markov chain $\mathbf{X}^{(t)}$ to the desired distribution.
- This does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution.
- Validity of the approximation must be confirmed in practise.
- Convergence diagnostics help answering this question (R package CODA).
- Convergence diagnostics are *not* perfect and should be treated with a good amount of scepticism; the techniques presented are **exploratory tools**.

Different diagnostic tasks

Convergence to the target distribution Does $\mathbf{X}^{(t)}$ yields a sample from the target distribution? (**sample paths**)

- Has reached $(\mathbf{X}^{(t)})_t$ a stationary regime?
- Does $(\mathbf{X}^{(t)})_t$ cover the support of the target distribution?

Convergence of the averages Does $\sum_{t=1}^T h(\mathbf{X}^{(t)})/T$ provide a good approximation to the expectation $\mathbb{E}_f(h(\mathbf{X}))$ under the target distribution? (**plots of cumulative averages**)

Comparison to i.i.d. sampling How much information is contained in the sample from the Markov chain compared to i.i.d. sampling? (**effective sample size**)

Different diagnostic tasks

Convergence to the target distribution Does $\mathbf{X}^{(t)}$ yields a sample from the target distribution? (**sample paths**)

- Has reached $(\mathbf{X}^{(t)})_t$ a stationary regime?
- Does $(\mathbf{X}^{(t)})_t$ cover the support of the target distribution?

Convergence of the averages Does $\sum_{t=1}^T h(\mathbf{X}^{(t)})/T$ provide a good approximation to the expectation $\mathbb{E}_f(h(\mathbf{X}))$ under the target distribution? (**plots of cumulative averages**)

Comparison to i.i.d. sampling How much information is contained in the sample from the Markov chain compared to i.i.d. sampling? (**effective sample size**)

Different diagnostic tasks

Convergence to the target distribution Does $\mathbf{X}^{(t)}$ yields a sample from the target distribution? (**sample paths**)

- Has reached $(\mathbf{X}^{(t)})_t$ a stationary regime?
- Does $(\mathbf{X}^{(t)})_t$ cover the support of the target distribution?

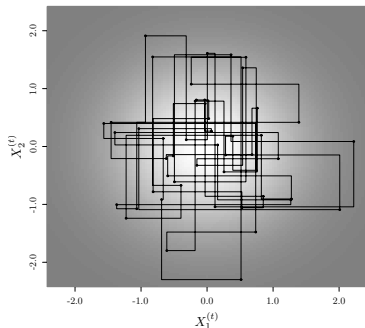
Convergence of the averages Does $\sum_{t=1}^T h(\mathbf{X}^{(t)})/T$ provide a good approximation to the expectation $\mathbb{E}_f(h(\mathbf{X}))$ under the target distribution? (**plots of cumulative averages**)

Comparison to i.i.d. sampling How much information is contained in the sample from the Markov chain compared to i.i.d. sampling? (**effective sample size**)

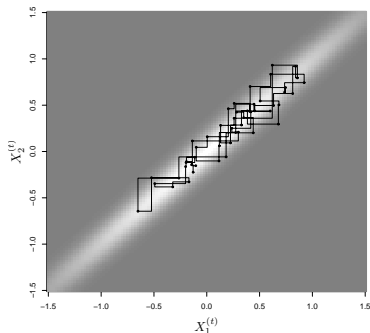
Pathological example 1: potentially slowly mixing

Gibbs sampler from a bivariate Gaussian with correlation $\rho(X_1, X_2)$

$$\rho(X_1, X_2) = 0.3$$



$$\rho(X_1, X_2) = 0.99$$



For correlations $\rho(X_1, X_2)$ close to ± 1 the chain can be poorly mixing.

Pathological example 2: no central limit theorem

The following MCMC algorithm has the $\text{Beta}(\alpha, 1)$ distribution as stationary distribution:

Starting with any $X^{(0)}$ iterate for $t = 1, 2, \dots$

1. With probability $1 - X^{(t-1)}$, set $X^{(t)} = X^{(t-1)}$.
2. Otherwise draw $X^{(t)} \sim \text{Beta}(\alpha + 1, 1)$.

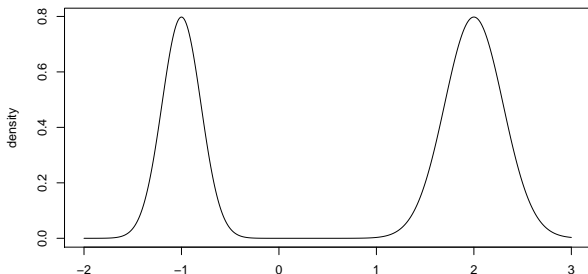
Markov chain converges very slowly (no central limit theorem applies).

Pathological example 3: nearly reducible chain

Metropolis-Hastings sample from a mixture of two well-separated Gaussians, i.e. the target is

$$f(x) = 0.4 \cdot \phi_{(-1,0.2^2)}(x) + 0.6 \cdot \phi_{(2,0.3^2)}(x)$$

If the variance of the proposal is too small, the chain cannot move from one population to the other.



Basic plots

- Plot the sample paths $(X_j^{(t)})_t$.
should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^t X_j^{(\tau)} / t)_t$.
should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^t X_j^{(\tau)} / t)_t$ with $\bar{X}_j = \sum_{\tau=1}^T X_j^{(\tau)} / T$.
should be converging to 0.
- Only very obvious problems visible in these plots.
- It is nearly impossible to tell that the chain has not explored the entire support of the target distribution: “you’ve only seen where you’ve been”!
- Difficult to assess multivariate distributions from univariate projections.

Basic plots

- Plot the sample paths $(X_j^{(t)})_t$.
should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^t X_j^{(\tau)} / t)_t$.
should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^t X_j^{(\tau)} / t)_t$ with $\bar{X}_j = \sum_{\tau=1}^T X_j^{(\tau)} / T$.
should be converging to 0.
- Only very obvious problems visible in these plots.
- It is nearly impossible to tell that the chain has not explored the entire support of the target distribution: “you’ve only seen where you’ve been”!
- Difficult to assess multivariate distributions from univariate projections.

Basic plots

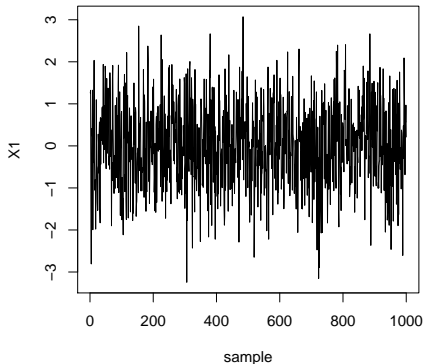
- Plot the sample paths $(X_j^{(t)})_t$.
should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^t X_j^{(\tau)} / t)_t$.
should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^t X_j^{(\tau)} / t)_t$ with $\bar{X}_j = \sum_{\tau=1}^T X_j^{(\tau)} / T$.
should be converging to 0.
- Only very obvious problems visible in these plots.
- It is nearly impossible to tell that the chain has not explored the entire support of the target distribution: **“you’ve only seen where you’ve been”!**
- Difficult to assess multivariate distributions from univariate projections.

Basic plots

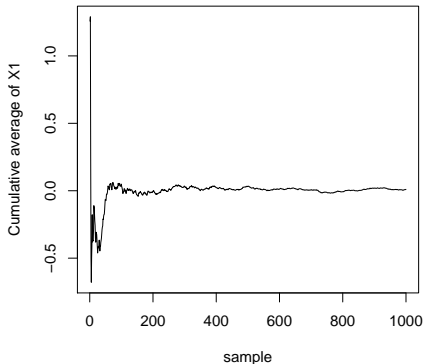
- Plot the sample paths $(X_j^{(t)})_t$.
should be oscillating very fast and show very little structure.
- Plot the cumulative averages $(\sum_{\tau=1}^t X_j^{(\tau)} / t)_t$.
should be converging to a value.
- Alternatively plot CUSUM $(\bar{X}_j - \sum_{\tau=1}^t X_j^{(\tau)} / t)_t$ with $\bar{X}_j = \sum_{\tau=1}^T X_j^{(\tau)} / T$.
should be converging to 0.
- Only very obvious problems visible in these plots.
- It is nearly impossible to tell that the chain has not explored the entire support of the target distribution: “you’ve only seen where you’ve been”!
- Difficult to assess multivariate distributions from univariate projections.

Basic plots for pathological ex. 1 ($\rho(X_1, X_2) = 0.3$)

Sample paths



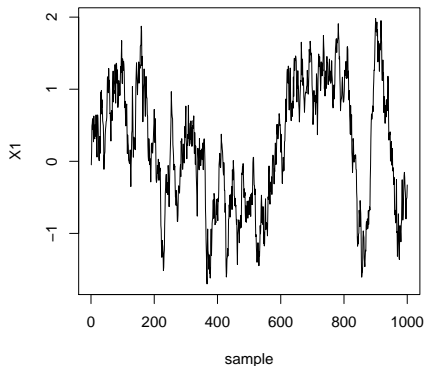
Cumulative averages



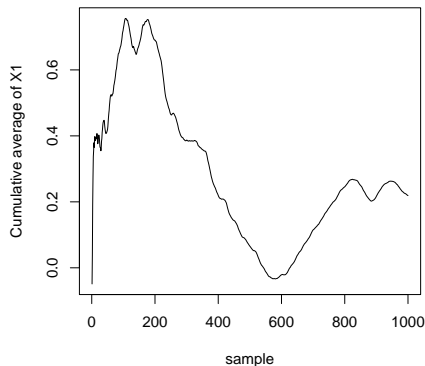
Looks OK.

Basic plots for pathological ex. 1 ($\rho(X_1, X_2) = 0.99$)

Sample paths



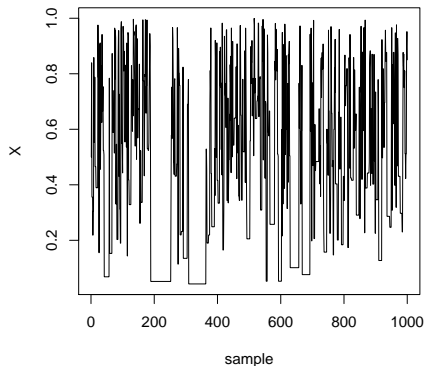
Cumulative averages



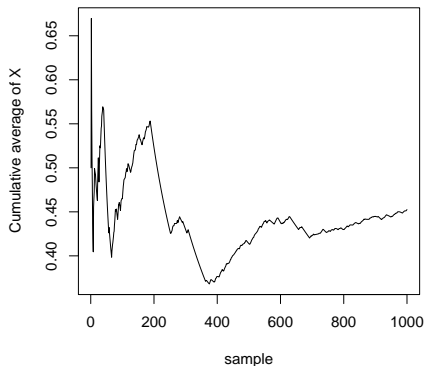
Slow mixing speed can be detected.

Basic plots for pathological example 2

Sample paths



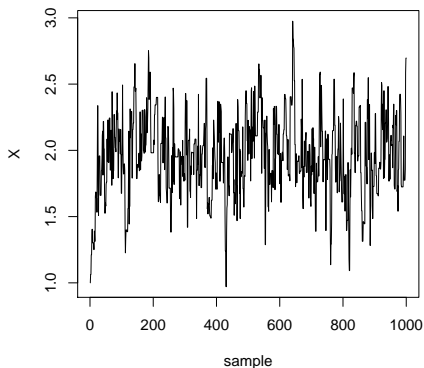
Cumulative averages



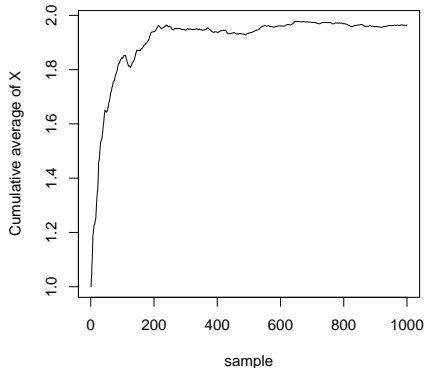
Slow convergence of the mean can be detected.

Basic plots for pathological example 3

Sample paths



Cumulative averages



Techniques fail to detect lack of convergence: we **cannot** detect that the sample only covers one part of the distribution (“you’ve only seen where you’ve been”).

Other tests for diagnosing convergence:

- Non-parametric test of stationarity: the Kolmogorov-Smirnov test.
- Comparing multiple chains: estimate inter-quantile distances in two different ways.
- Riemann sums and control variates: estimate $\int_E f(x)dx$.
- Compute the *effective sample size*: the size an i.i.d. sample would have to have in order to obtain the same variance as the estimate from the Markov chain.

Comparing multiple chains

- Compare $L > 1$ chains $(\mathbf{X}^{(1,t)})_t, \dots, (\mathbf{X}^{(L,t)})_t$.
- Initialised using overdispersed starting values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$.
- **Idea:** Variance and range of each chain $(\mathbf{X}^{(l,t)})_t$ should equal the range and variance of all chains pooled together.
- Compare basic plots for the different chains.
- Quantitative measure:
 - Compute distance $\delta_\alpha^{(l)}$ between α and $(1 - \alpha)$ quantile of $(X_k^{(l,t)})_t$.
 - Compute distance $\delta_\alpha^{(\cdot)}$ between α and $(1 - \alpha)$ quantile of the pooled data.
 - The ratio $\hat{S}_\alpha^{\text{interval}} = \frac{\sum_{l=1}^L \delta_\alpha^{(l)}}{\delta_\alpha^{(\cdot)}}$ should be around 1.
- Alternative: compare variance within each chain with the pooled variance estimate.
- Choosing suitable initial values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$ is difficult in high dimensions.

Comparing multiple chains

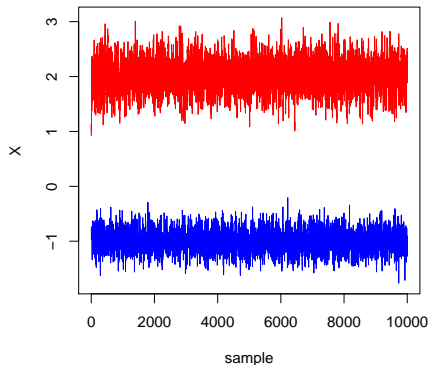
- Compare $L > 1$ chains $(\mathbf{X}^{(1,t)})_t, \dots, (\mathbf{X}^{(L,t)})_t$.
- Initialised using overdispersed starting values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$.
- **Idea:** Variance and range of each chain $(\mathbf{X}^{(l,t)})_t$ should equal the range and variance of all chains pooled together.
- Compare basic plots for the different chains.
- Quantitative measure:
 - Compute distance $\delta_\alpha^{(l)}$ between α and $(1 - \alpha)$ quantile of $(X_k^{(l,t)})_t$.
 - Compute distance $\delta_\alpha^{(\cdot)}$ between α and $(1 - \alpha)$ quantile of the pooled data.
 - The ratio $\hat{S}_\alpha^{\text{interval}} = \frac{\sum_{l=1}^L \delta_\alpha^{(l)} / L}{\delta_\alpha^{(\cdot)}}$ should be around 1.
- Alternative: compare variance within each chain with the pooled variance estimate.
- Choosing suitable initial values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$ is difficult in high dimensions.

Comparing multiple chains

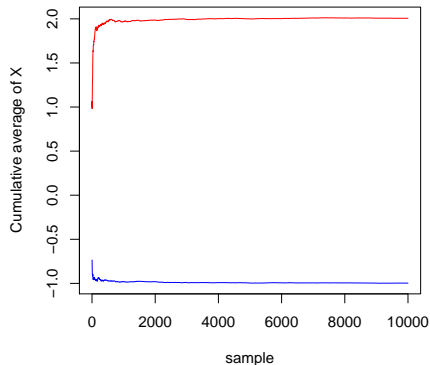
- Compare $L > 1$ chains $(\mathbf{X}^{(1,t)})_t, \dots, (\mathbf{X}^{(L,t)})_t$.
- Initialised using overdispersed starting values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$.
- **Idea:** Variance and range of each chain $(\mathbf{X}^{(l,t)})_t$ should equal the range and variance of all chains pooled together.
- Compare basic plots for the different chains.
- Quantitative measure:
 - Compute distance $\delta_\alpha^{(l)}$ between α and $(1 - \alpha)$ quantile of $(X_k^{(l,t)})_t$.
 - Compute distance $\delta_\alpha^{(\cdot)}$ between α and $(1 - \alpha)$ quantile of the pooled data.
 - The ratio $\hat{S}_\alpha^{\text{interval}} = \frac{\sum_{l=1}^L \delta_\alpha^{(l)} / L}{\delta_\alpha^{(\cdot)}}$ should be around 1.
- Alternative: compare variance within each chain with the pooled variance estimate.
- Choosing suitable initial values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$ is difficult in high dimensions.

Comparing multiple chains plots for pathological example 3

Sample paths



Cumulative averages



$$\hat{S}_{\alpha}^{\text{interval}} = 0.2703 \ll 1$$

We can detect that the sample only covers one part of the distribution (provided the chains are initialised appropriately).

Riemann sums and control variates

- Consider order statistic $X^{[1]} \leq \dots \leq X^{[T]}$.
- Provided $(X^{[t]})_{t=1\dots,T}$ covers the support of the target, the Riemann sum

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]})f(X^{[t]})$$

converges to

$$\int f(x)dx = 1.$$

- Thus if $\sum_{t=2}^T (X^{[t]} - X^{[t-1]})f(X^{[t]}) \ll 1$, the Markov chain has failed to explore all the support of the target.
- Requires that the target density f be available inclusive of normalisation constants.
- Only effective in 1D.
- Riemann sums can be seen as a special case of *control variates*.

Riemann sums and control variates

- Consider order statistic $X^{[1]} \leq \dots \leq X^{[T]}$.
- Provided $(X^{[t]})_{t=1,\dots,T}$ covers the support of the target, the Riemann sum

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]})f(X^{[t]})$$

converges to

$$\int f(x)dx = 1.$$

- Thus if $\sum_{t=2}^T (X^{[t]} - X^{[t-1]})f(X^{[t]}) \ll 1$, the Markov chain has failed to explore all the support of the target.
- Requires that the target density f be available inclusive of normalisation constants.
- Only effective in 1D.
- Riemann sums can be seen as a special case of *control variates*.

Riemann sums for pathological example 3

For the chain stuck in the population with mean 2 we obtain

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]})f(X^{[t]}) = 0.598 \ll 1,$$

so we can detect that we have not explored the whole distribution.

Effective sample size

- MCMC algorithms yield a positively correlated sample $(\mathbf{X}^{(t)})_{t=1,\dots,T}$.
- MCMC sample of size T thus contains less information than an i.i.d. sample of size T .
- Question: how much less information?
- Approximate $(h(\mathbf{X}^{(t)}))_{t=1,\dots,T}$ by an $AR(1)$ process, i.e. we assume that

$$\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho^{|\tau|}.$$

- Variance of the estimator is

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) \approx \frac{1 + \rho}{1 - \rho} \cdot \frac{1}{T} \text{Var} \left(h(\mathbf{X}^{(t)}) \right)$$

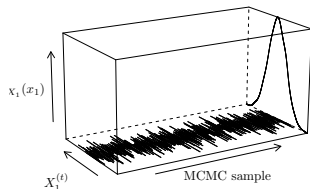
- Same variance as an i.i.d. sample of the size $T \cdot \frac{1 - \rho}{1 + \rho}$.
- Thus define $T \cdot \frac{1 - \rho}{1 + \rho}$ as **effective sample size**.

Effective sample for pathological example 1

Rapidly mixing chain

$$(\rho(X_1, X_2) = 0.3)$$

10,000 samples



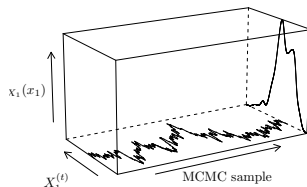
$$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.078$$

ESS for estimating $\mathbb{E}_f(X_1)$ is 8,547.

Slowly mixing chain

$$(\rho(X_1, X_2) = 0.99)$$

10,000 samples



$$\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.979$$

ESS for estimating $\mathbb{E}_f(X_1)$ is 105.