

Chapter 6

Diagnosing convergence

6.1 Practical considerations

The theory of Markov chains we have seen in chapter 1 guarantees that a Markov chain that is irreducible and has invariant distribution f converges to the invariant distribution. The ergodic theorems 4.6 and 5.5 allow for approximating expectations $\mathbb{E}_f(h(\mathbf{X}))$ by their the corresponding means

$$\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \longrightarrow \mathbb{E}_f(h(\mathbf{X}))$$

using the *entire* chain. In practise, however, often only a subset of the chain $(\mathbf{X}^{(t)})_t$ is used:

Burn-in Depending on how $\mathbf{X}^{(0)}$ is chosen, the distribution of $(\mathbf{X}^{(t)})_t$ for small t might still be far from the stationary distribution f . Thus it might be beneficial to discard the first iterations $\mathbf{X}^{(t)}$, $t = 1, \dots, T_0$. This early stage of the sampling process is often referred to as *burn-in* period. How large T_0 has to be chosen depends on how fast mixing the Markov chain $(\mathbf{X}^{(t)})_t$ is. Figure 6.1 illustrates the idea of a burn-in period.

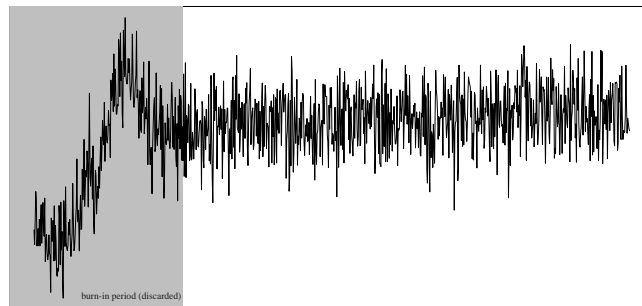


Figure 6.1. Illustration of the idea of a burn-in period.

Thinning Markov chain Monte Carlo methods typically yield a Markov chain with positive autocorrelation, i.e. $\rho(X_k^{(t)}, X_k^{(t+\tau)})$ is positive for small τ . This suggests building a subchain by only keeping every m -th value ($m > 1$), i.e. we consider a Markov chain $(\mathbf{Y}^{(t)})_t$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$ instead of $(\mathbf{X}^{(t)})_t$. If the correlation $\rho(\mathbf{X}^{(t)}, \mathbf{X}^{(t+\tau)})$ decreases monotonically in τ , then

$$\rho(Y_k^{(t)}, Y_k^{(t+\tau)}) = \rho(X_k^{(t)}, X_k^{(m \cdot t + \tau)}) < \rho(X_k^{(t)}, X_k^{(t+\tau)}),$$

i.e. the thinned chain $(\mathbf{Y}^{(t)})_t$ exhibits less autocorrelation than the original chain $(\mathbf{X}^{(t)})_t$. Thus thinning can be seen as a technique for reducing the autocorrelation, however at the price of yielding a chain $(\mathbf{Y}^{(t)})_{t=1, \dots, \lfloor T/m \rfloor}$,

whose length is reduced to $(1/m)$ -th of the length of the original chain $(\mathbf{X}^{(t)})_{t=1,\dots,T}$. Even though thinning is very popular, it cannot be justified when the objective is estimating $\mathbb{E}_f(h(\mathbf{X}))$, as the following lemma shows.

Lemma 6.1. *Let $(\mathbf{X}^{(t)})_{t=1,\dots,T}$ be a sequence of random variables (e.g. from a Markov chain) with $\mathbf{X}^{(t)} \sim f$ and $(\mathbf{Y}^{(t)})_{t=1,\dots,\lfloor T/m \rfloor}$ a second sequence defined by $\mathbf{Y}^{(t)} := \mathbf{X}^{(m \cdot t)}$. If $\text{Var}_f(h(\mathbf{X}^{(t)})) < +\infty$, then*

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

Proof. To simplify the proof we assume that T is divisible by m , i.e. $T/m \in \mathbb{N}$. Using

$$\sum_{t=1}^T h(\mathbf{X}^{(t)}) = \sum_{\tau=0}^{m-1} \sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \tau)})$$

and

$$\text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \tau_1)}) \right) = \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \tau_2)}) \right)$$

for $\tau_1, \tau_2 \in \{0, \dots, m-1\}$, we obtain that

$$\begin{aligned} \text{Var} \left(\sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) &= \text{Var} \left(\sum_{\tau=0}^{m-1} \sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \tau)}) \right) \\ &= m \cdot \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m)}) \right) + \underbrace{\sum_{\eta \neq \tau=0}^{m-1} \text{Cov} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \eta)}), \sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m + \tau)}) \right)}_{\leq \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m)}) \right)} \\ &\leq m^2 \cdot \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{X}^{(t \cdot m)}) \right) = m^2 \cdot \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{Y}^{(t)}) \right). \end{aligned}$$

Thus

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) = \frac{1}{T^2} \text{Var} \left(\sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) \leq \frac{m^2}{T^2} \text{Var} \left(\sum_{t=1}^{T/m} h(\mathbf{Y}^{(t)}) \right) = \text{Var} \left(\frac{1}{T/m} \sum_{t=1}^{T/m} h(\mathbf{Y}^{(t)}) \right).$$

□

The concept of thinning can be useful for other reasons. If the computer's memory cannot hold the entire chain $(\mathbf{X}^{(t)})_t$, thinning is a good choice. Further, it can be easier to assess the convergence of the thinned chain $(\mathbf{Y}^{(t)})_t$ as opposed to entire chain $(\mathbf{X}^{(t)})_t$.

6.2 Tools for monitoring convergence

Although the theory presented in the preceding chapters guarantees the convergence of the Markov chains to the required distributions, this does not imply that a *finite* sample from such a chain yields a good approximation to the target distribution. As with all approximating methods this must be confirmed in practise.

This section tries to give a brief overview over various approaches to diagnosing convergence. A more detailed review with many practical examples can be diagnosed in (Guihenec-Jouyaux et al., 1998) or (Robert and Casella, 2004, chapter 12). There is an R package (CODA) that provides a vast selection of tools for diagnosing convergence. Diagnosing convergence is an art. The techniques presented in the following are nothing other than exploratory tools that help you judging whether the chain has reached its stationary regime. This section contains several cautionary examples where the different tools for diagnosing convergence fail.

Broadly speaking, convergence assessment can be split into the following three tasks of diagnosing different aspects of convergence:

Convergence to the target distribution. The first, and most important, question is whether $(\mathbf{X}^{(t)})_t$ yields a sample from the target distribution? In order to answer this question we need to assess . . .

- whether $(\mathbf{X}^{(t)})_t$ has reached a stationary regime, and
- whether $(\mathbf{X}^{(t)})_t$ covers the entire 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?

Comparison to i.i.d. sampling. How much information is contained in the sample from the Markov chain compared to i.i.d. sampling?

6.2.1 Basic plots

The most basic approach to diagnosing the output of a Markov Chain Monte Carlo algorithm is to plot the sample path $(\mathbf{X}^{(t)})_t$ as in figures 4.4 (b) (c), 4.5 (b) (c), 5.3 (a), and 5.4. Note that the convergence of $(\mathbf{X}^{(t)})_t$ is in distribution, i.e. the sample path is *not* supposed to converge to a single value. Ideally, the plot should be oscillating very fast and show very little structure or trend (like for example figure 4.4). The smoother the plot seems (like for example figure 4.5), the slower mixing the resulting chain is.

Note however that this plot suffers from the “you’ve only seen where you’ve been” problem. It is impossible to see from a plot of the sample path whether the chain has explored the entire support of the distribution.

Example 6.1 (A simple mixture of two Gaussians). In this example we sample from a mixture of two well-separated Gaussians

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

(see figure 6.2 (a) for a plot of the density) using a random walk Metropolis algorithm with proposed value $X = X^{(t-1)} + \varepsilon$ with $\varepsilon \sim N(0, \text{Var}(\varepsilon))$. If we choose the proposal variance $\text{Var}(\varepsilon)$ too small, we only sample from one population instead of both. Figure 6.2 shows the sample paths for two choices of $\text{Var}(\varepsilon)$: $\text{Var}(\varepsilon) = 0.4^2$ and $\text{Var}(\varepsilon) = 1.2^2$. The first choice of $\text{Var}(\varepsilon)$ is too small: the chain is very likely to remain in one of the two modes of the distribution. Note that it is impossible to tell from figure 6.2 (b) alone that the chain has not explored the entire support of the target. ◀

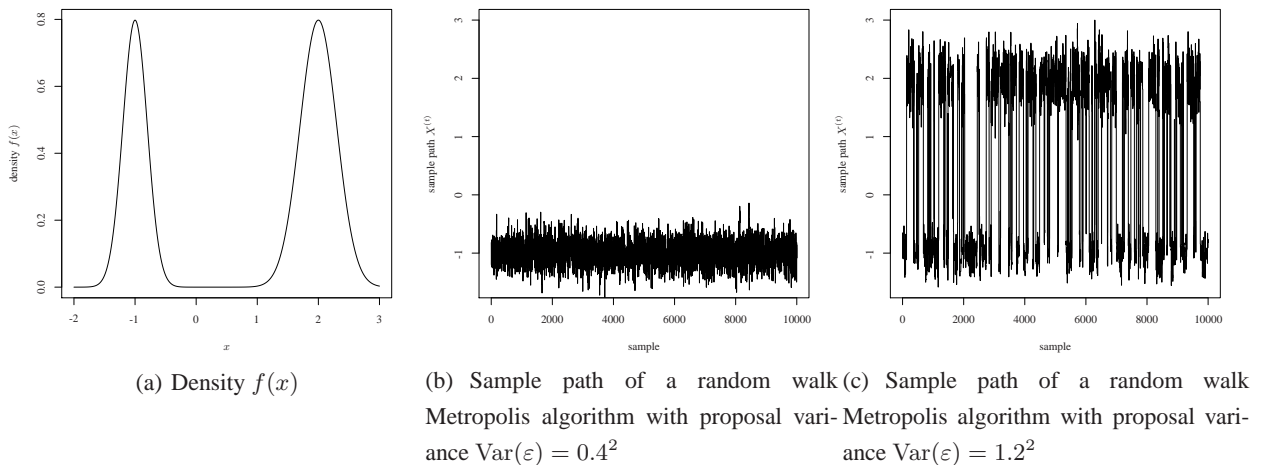


Figure 6.2. Density of the mixture distribution with two random walk Metropolis samples using two different variances $\text{Var}(\varepsilon)$ of the proposal.

In order to diagnose the convergence of the averages, one can look at a plot of the cumulative averages $(\sum_{\tau=1}^t h(X^{(\tau)})/t)_t$. Note that the convergence of the cumulative averages is — as the ergodic theorems suggest — to a value $(\mathbb{E}_f(h(\mathbf{X})))$. Figures 4.3, and 5.3 (b) show plots of the cumulative averages. An alternative

to plotting the cumulative means is using the so-called CUSUMs $\left(\bar{h}(X_j) - \sum_{\tau=1}^t h(X_j^{(\tau)})/t\right)_t$ with $\bar{h}(X_j) = \sum_{\tau=1}^T h(X_j^{(\tau)})/T$, which is nothing other than the difference between the cumulative averages and the estimate of the limit $\mathbb{E}_f(h(\mathbf{X}))$.

Example 6.2 (A pathological generator for the Beta distribution). The following MCMC algorithm (for details, see Robert and Casella, 2004, problem 7.5) yields a sample from the $\text{Beta}(\alpha, 1)$ 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 $X^{(t)} \sim \text{Beta}(\alpha + 1, 1)$.

This algorithm yields a very slowly converging Markov chain, to which no central limit theorem applies. This slow convergence can be seen in a plot of the cumulative means (figure 6.3 (b)). ◀

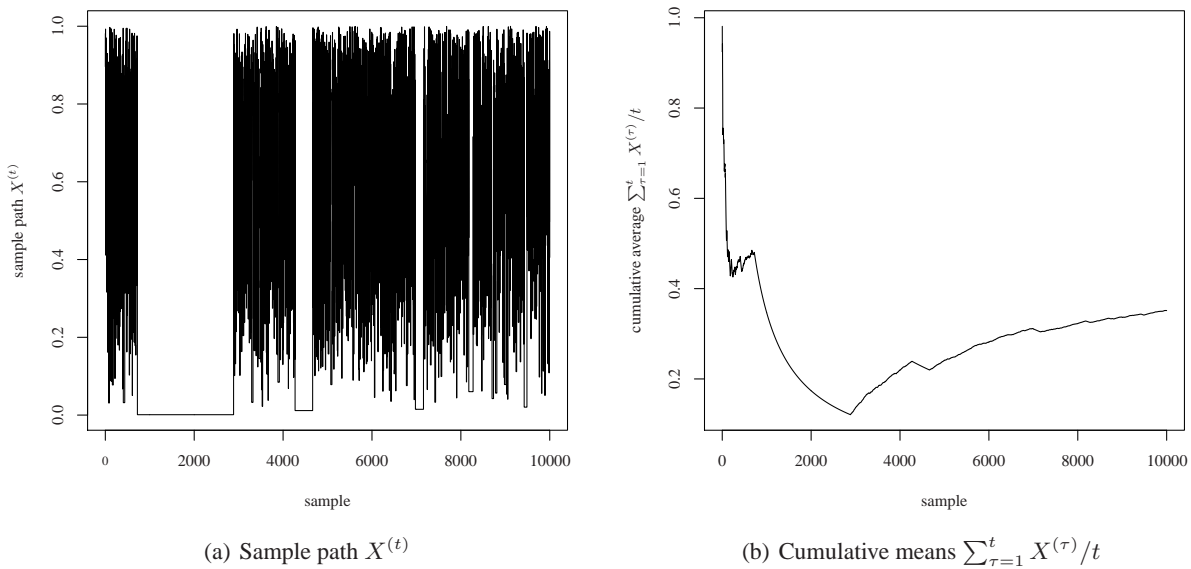


Figure 6.3. Sample paths and cumulative means obtained for the pathological Beta generator.

Note that it is impossible to tell from a plot of the cumulative means whether the Markov chain has explored the entire support of the target distribution.

6.2.2 Non-parametric tests of stationarity

This section presents the Kolmogorov-Smirnov test, which is an example of how nonparametric tests can be used as a tool for diagnosing whether a Markov chain has already converged.

In its simplest version, it is based on splitting the chain into three parts: $(\mathbf{X}^{(t)})_{t=1, \dots, \lfloor T/3 \rfloor}$, $(\mathbf{X}^{(t)})_{t=\lfloor T/3 \rfloor+1, \dots, 2\lfloor T/3 \rfloor}$, and $(\mathbf{X}^{(t)})_{t=2\lfloor T/3 \rfloor+1, \dots, T}$. The first block is considered to be the burn-in period. If the Markov chain has reached its stationary regime after $\lfloor T/3 \rfloor$ iterations, the second and third block should be from the same distribution. Thus we should be able to tell whether the chain has converged by comparing the distribution of $(\mathbf{X}^{(t)})_{t=\lfloor T/3 \rfloor+1, \dots, 2\lfloor T/3 \rfloor}$ to the one of $(\mathbf{X}^{(t)})_{t=2\lfloor T/3 \rfloor+1, \dots, T}$ using suitable nonparametric two-sample tests. One such test is the Kolmogorov-Smirnov test.

As the Kolmogorov-Smirnov test is designed for i.i.d. samples, we do not apply it to the $(\mathbf{X}^{(t)})_t$ directly, but to a thinned chain $(\mathbf{Y}^{(t)})_t$ with $\mathbf{Y}^{(t)} = \mathbf{X}^{(m \cdot t)}$: the thinned chain is less correlated and thus closer to being an i.i.d. sample. We can now compare the distribution of $(\mathbf{Y}^{(t)})_{t=\lfloor T/(3m) \rfloor+1, \dots, 2\lfloor T/(3m) \rfloor}$ to the one of

$(\mathbf{Y}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1, \dots, \lfloor T/m\rfloor}$ using the Kolmogorov-Smirnov statistic ¹

$$K = \sup_{x \in \mathbb{R}} \left| \hat{F}_{(\mathbf{Y}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1, \dots, 2\lfloor T/(3m)\rfloor}}(x) - \hat{F}_{(\mathbf{Y}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1, \dots, \lfloor T/m\rfloor}}(x) \right|.$$

As the thinned chain is not an i.i.d. sample, we cannot use the Kolmogorov-Smirnov test as a formal statistical test (besides we would run into problems of multiple testing). However, we can use it as an informal tool by monitoring the standardised statistic $\sqrt{t}K_t$ as a function of t .² As long as a significant proportion of the values of the standardised statistic are above the corresponding quantile of the asymptotic distribution, it is safe to assume that the chain has not yet reached its stationary regime.

Example 6.3 (Gibbs sampling from a bivariate Gaussian (continued)). In this example we consider sampling from a bivariate Gaussian distribution, once with $\rho(X_1, X_2) = 0.3$ (as in example 4.4) and once with $\rho(X_1, X_2) = 0.99$ (as in example 4.5). The former leads a fast mixing chain, the latter a very slowly mixing chain. Figure 6.4 shows the plots of the standardised Kolmogorov-Smirnov statistic. It suggests that the sample size of 10,000 is large enough for the low-correlation setting, but not large enough for the high-correlation setting. ◁

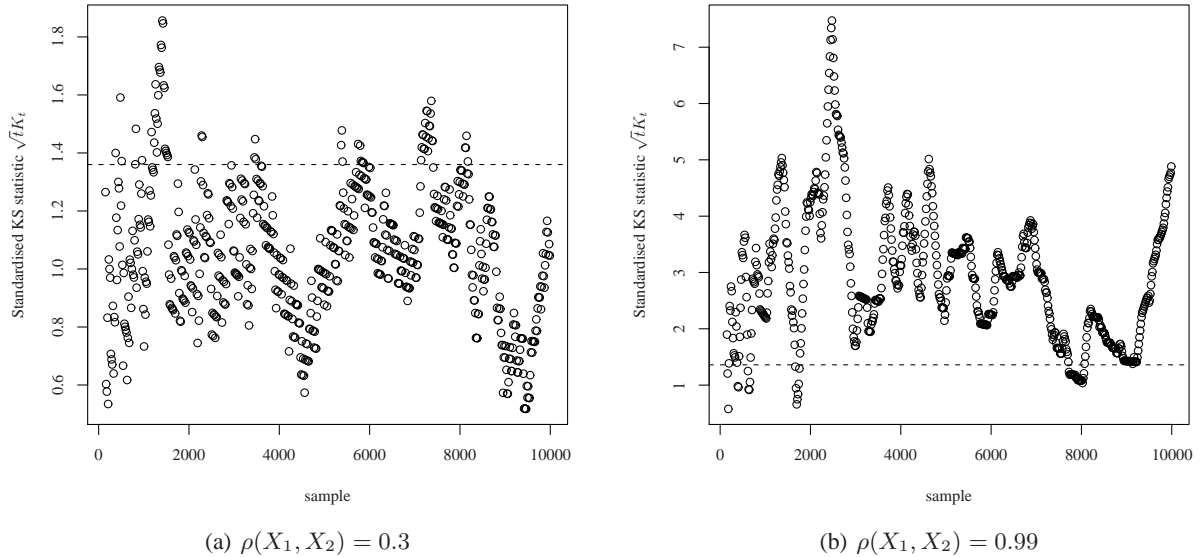


Figure 6.4. Standardised Kolmogorov-Smirnov statistic for $X_1^{(5-t)}$ from the Gibbs sampler from the bivariate Gaussian for two different correlations.

Note that the Kolmogorov-Smirnov test suffers from the “you’ve only seen where you’ve been” problem, as it is based on comparing $(\mathbf{Y}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1, \dots, 2\lfloor T/(3m)\rfloor}$ and $(\mathbf{Y}^{(t)})_{t=2\lfloor T/(3m)\rfloor+1, \dots, \lfloor T/m\rfloor}$. A plot of the Kolmogorov-Smirnov statistic for the chain with $\text{Var}(\varepsilon) = 0.4$ from example 6.1 would not reveal anything unusual.

¹ The two-sample Kolmogorov-Smirnov test for comparing two i.i.d. samples $Z_{1,1}, \dots, Z_{1,n}$ and $Z_{2,1}, \dots, Z_{2,n}$ is based on comparing their empirical CDFs

$$\hat{F}_k(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, z]}(Z_{k,i}).$$

The Kolmogorov-Smirnov test statistic is the maximum difference between the two empirical CDFs:

$$K = \sup_{z \in \mathbb{R}} |\hat{F}_1(z) - \hat{F}_2(z)|.$$

For $n \rightarrow \infty$ the CDF of $\sqrt{n} \cdot K$ converges to the CDF

$$R(k) = 1 - \sum_{i=1}^{+\infty} (-1)^{i-1} \exp(-2i^2 k^2).$$

² K_t is hereby the Kolmogorov-Smirnov statistic obtained from the sample consisting of the first t observations only.

6.2.3 Riemann sums and control variates

A simple tool for diagnosing convergence of a one-dimensional Markov chain can be based on the fact that

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

We can estimate this integral by the Riemann sum

$$\sum_{t=2}^T (X^{[t]} - X^{[t-1]}) f(X^{[t]}), \quad (6.1)$$

where $X^{[1]} \leq \dots \leq X^{[T]}$ is the ordered sample from the Markov chain. If the Markov chain has explored all the support of f , then (6.1) should be around 1. Note that this method, often referred to as Riemann sums (Philippe and Robert, 2001), requires that the density f is known inclusive of normalisation constants.

Example 6.4 (A simple mixture of two Gaussians (continued)). In example 6.1 we considered two random-walk Metropolis algorithms: one ($\text{Var}(\varepsilon) = 0.4^2$) failed to explore the entire support of the target distribution, whereas the other one ($\text{Var}(\varepsilon) = 1.2^2$) managed to. The corresponding Riemann sums are 0.598 and 1.001, clearly indicating that the first algorithm does not explore the entire support. \triangleleft

Riemann sums can be seen as a special case of a technique called *control variates*. The idea of control variates is comparing several ways of estimating the same quantity. As long as the different estimates disagree, the chain has not yet converged. Note that the technique of control variates is only useful if the different estimators converge about as fast as the quantity of interest — otherwise we would obtain an overly optimistic, or an overly conservative estimate of whether the chain has converged. In the special case of the Riemann sum we compare two quantities: the constant 1 and the Riemann sum (6.1).

6.2.4 Comparing multiple chains

A family of convergence diagnostics (see e.g. Gelman and Rubin, 1992; Brooks and Gelman, 1998) is based on running $L > 1$ chains — which we will denote by $(\mathbf{X}^{(1,t)})_t, \dots, (\mathbf{X}^{(L,t)})_t$ — with overdispersed³ starting values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$, covering at least the support of the target distribution.

All L chains should converge to the same distribution, so comparing the plots from section 6.2.1 for the L different chains should not reveal any difference. A more formal approach to diagnosing whether the L chains are all from the same distribution can be based on comparing the inter-quantile distances.

We can estimate the inter-quantile distances in two ways. The first consists of estimating the inter-quantile distance for each of the L chain and averaging over these results, i.e. our estimate is $\sum_{l=1}^L \delta_\alpha^{(l)} / L$, where $\delta_\alpha^{(l)}$ is the distance between the α and $(1 - \alpha)$ quantile of the l -th chain $(X^{(l,t)})_t$. Alternatively, we can pool the data first, and then compute the distance between the α and $(1 - \alpha)$ quantile of the pooled data. If all chains are a sample from the same distribution, both estimates should be roughly the same, so their ratio

$$\hat{S}_\alpha^{\text{interval}} = \frac{\sum_{l=1}^L \delta_\alpha^{(l)} / L}{\delta_\alpha^{(\cdot)}}$$

can be used as a tool to diagnose whether all chains sampled from the same distribution, in which case the ratio should be around 1.

Alternatively, one could compare the variances within the L chains to the pooled estimate of the variance (see Brooks and Gelman, 1998, for more details).

³ i.e. the variance of the starting values should be larger than the variance of the target distribution.

Example 6.5 (A simple mixture of two Gaussians (continued)). In the example of the mixture of two Gaussians we will consider $L = 8$ chains initialised from a $N(0, 10^2)$ distribution. Figure 6.5 shows the sample paths of the 8 chains for both choices of $\text{Var}(\varepsilon)$. The corresponding values of $\hat{S}_{0.05}^{\text{interval}}$ are:

$$\begin{aligned} \text{Var}(\varepsilon) = 0.4^2 : \hat{S}_{0.05}^{\text{interval}} &= \frac{0.9789992}{3.630008} = 0.2696962 \\ \text{Var}(\varepsilon) = 1.2^2 : \hat{S}_{0.05}^{\text{interval}} &= \frac{3.634382}{3.646463} = 0.996687. \end{aligned} \quad \triangleleft$$

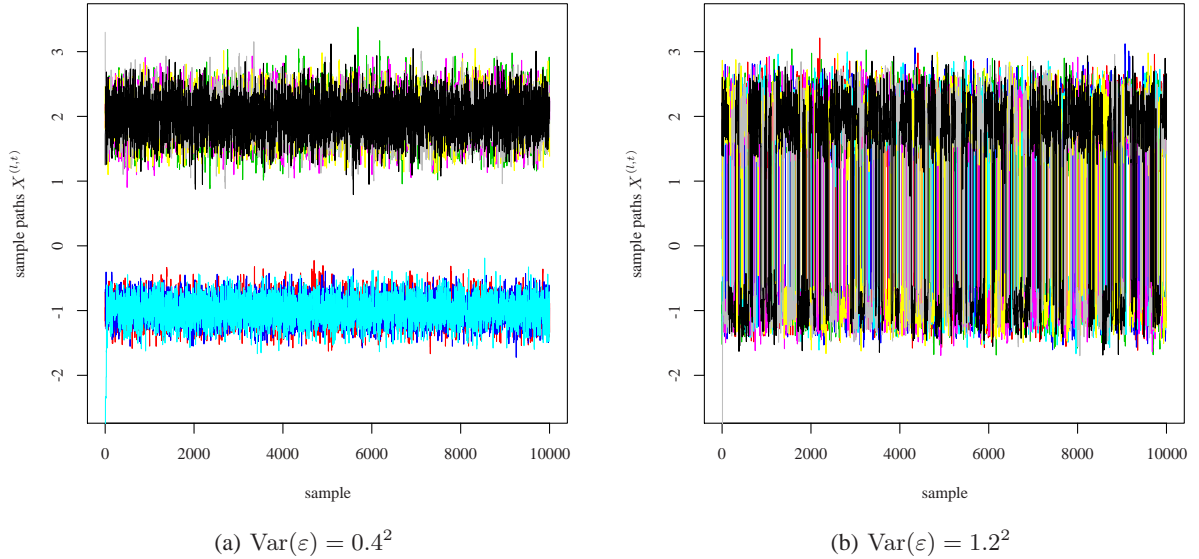


Figure 6.5. Comparison of the sample paths for $L = 8$ chains for the mixture of two Gaussians.

Note that this method depends crucially on the choice of initial values $\mathbf{X}^{(1,0)}, \dots, \mathbf{X}^{(L,0)}$, and thus can easily fail, as the following example shows.

Example 6.6 (Witch's hat distribution). Consider a distribution with the following density:

$$f(x_1, x_2) \propto \begin{cases} (1 - \delta)\phi_{(\boldsymbol{\mu}, \sigma^2, \mathbb{I})}(x_1, x_2) + \delta & \text{if } x_1, x_2 \in (0, 1) \\ 0 & \text{else,} \end{cases}$$

which is a mixture of a Gaussian and a uniform distribution, both truncated to $[0, 1] \times [0, 1]$. Figure 6.6 illustrates the density. For very small σ^2 , the Gaussian component is concentrated in a very small area around $\boldsymbol{\mu}$.

The conditional distribution of $X_1|X_2$ is

$$f(x_1|x_2) = \begin{cases} (1 - \delta_{x_2})\phi_{(\boldsymbol{\mu}, \sigma^2, \mathbb{I})}(x_1, x_2) + \delta_{x_2} & \text{for } x_1 \in (0, 1) \\ 0 & \text{else.} \end{cases}$$

with $\delta_{x_2} = \frac{\delta}{\delta + (1 - \delta)\phi_{(\boldsymbol{\mu}_2, \sigma^2)}(x_2)}$.

Assume we want to estimate $\mathbb{P}(0.49 < X_1, X_2 \leq 0.51)$ for $\delta = 10^{-3}$, $\boldsymbol{\mu} = (0.5, 0.5)'$, and $\sigma = 10^{-5}$ using a Gibbs sampler. Note that 99.9% of the mass of the distribution is concentrated in a very small area around $(0.5, 0.5)$, i.e. $\mathbb{P}(0.49 < X_1, X_2 \leq 0.51) = 0.999$.

Nonetheless, it is very unlikely that the Gibbs sampler visits this part of the distribution. This is due to the fact that unless x_2 (or x_1) is very close to μ_2 (or μ_1), δ_{x_2} (or δ_{x_1}) is almost 1, i.e. the Gibbs sampler only samples from the uniform component of the distribution. Figure 6.6 shows the samples obtained from 15 runs of the Gibbs

sampler (first 100 iterations only) all using different initialisations. On average only 0.04% of the sampled values lie in $(0.49, 0.51) \times (0.49, 0.51)$ yielding an estimate of $\hat{\mathbb{P}}(0.49 < X_1, X_2 \leq 0.51) = 0.0004$ (as opposed to $\mathbb{P}(0.49 < X_1, X_2 \leq 0.51) = 0.999$).

It is however close to impossible to detect this problem with any technique based on multiple initialisations. The Gibbs sampler shows this behaviour for practically all starting values. In figure 6.6 all 15 starting values yield a Gibbs sampler that is stuck in the “brim” of the witch’s hat and thus misses 99.9% of the probability mass of the target distribution. \triangleleft

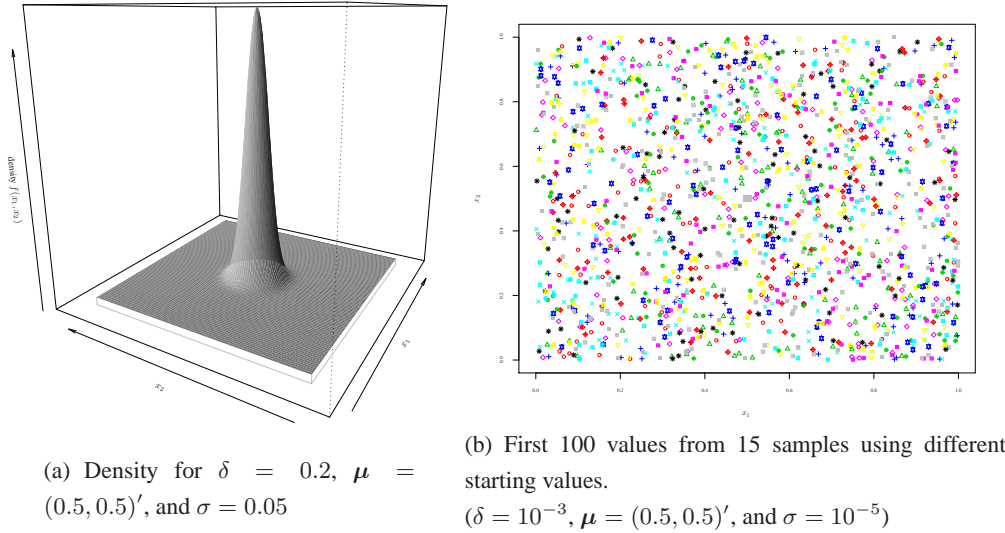


Figure 6.6. Density and sample from the witch’s hat distribution.

6.2.5 Comparison to i.i.d. sampling and the effective sample size

MCMC algorithms typically yield a positively correlated sample $(\mathbf{X}^{(t)})_{t=1, \dots, T}$, which contains less information than an i.i.d. sample of size T . If the $(\mathbf{X}^{(t)})_{t=1, \dots, T}$ are positively correlated, then the variance of the average

$$\text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) \quad (6.2)$$

is larger than the variance we would obtain from an i.i.d. sample, which is $\text{Var}(h(\mathbf{X}^{(t)}))/T$.

The effective sample size (ESS) allows to quantify this loss of information caused by the positive correlation. The effective sample size is the size an i.i.d. sample would have to have in order to obtain the same variance (6.2) as the estimate from the Markov chain $(\mathbf{X}^{(t)})_{t=1, \dots, T}$.

In order to compute the variance (6.2) we make the simplifying assumption that $(h(\mathbf{X}^{(t)}))_{t=1, \dots, T}$ is from a second-order stationary time series, i.e. $\text{Var}(h(\mathbf{X}^{(t)})) = \sigma^2$, and $\rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho(\tau)$. Then

$$\begin{aligned} \text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) &= \frac{1}{T^2} \left(\sum_{t=1}^T \underbrace{\text{Var}(h(\mathbf{X}^{(t)}))}_{=\sigma^2} + 2 \sum_{1 \leq s < t \leq T} \underbrace{\text{Cov}(h(\mathbf{X}^{(s)}), h(\mathbf{X}^{(t)}))}_{=\sigma^2 \cdot \rho(t-s)} \right) \\ &= \frac{\sigma^2}{T^2} \left(T + 2 \sum_{\tau=1}^{T-1} (T - \tau) \rho(\tau) \right) = \frac{\sigma^2}{T} \left(1 + 2 \sum_{\tau=1}^{T-1} \left(1 - \frac{\tau}{T}\right) \rho(\tau) \right). \end{aligned}$$

If $\sum_{\tau=1}^{+\infty} |\rho(\tau)| < +\infty$, then we can obtain from the dominated convergence theorem⁴ that

⁴ see e.g. Brockwell and Davis (1991, theorem 7.1.1) for details.

$$T \cdot \text{Var} \left(\frac{1}{T} \sum_{t=1}^T h(\mathbf{X}^{(t)}) \right) \longrightarrow \sigma^2 \left(1 + 2 \sum_{\tau=1}^{+\infty} \rho(\tau) \right)$$

as $T \rightarrow \infty$. Note that the variance would be σ^2/T_{ESS} if we were to use an i.i.d. sample of size T_{ESS} . We can now obtain the effective sample size T_{ESS} by equating these two variances and solving for T_{ESS} , yielding

$$T_{\text{ESS}} = \frac{1}{1 + 2 \sum_{\tau=1}^{+\infty} \rho(\tau)} \cdot T.$$

If we assume that $(h(\mathbf{X}^{(t)}))_{t=1, \dots, T}$ is a first-order autoregressive time series (AR(1)), i.e. $\rho(\tau) = \rho(h(\mathbf{X}^{(t)}), h(\mathbf{X}^{(t+\tau)})) = \rho^{|\tau|}$, then we obtain using $1 + 2 \sum_{\tau=1}^{+\infty} \rho^\tau = (1 + \rho)/(1 - \rho)$ that

$$T_{\text{ESS}} = \frac{1 - \rho}{1 + \rho} \cdot T.$$

Example 6.7 (Gibbs sampling from a bivariate Gaussian (continued)). In examples 4.4 and 4.5 we obtained for the low-correlation setting that $\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.078$, thus the effective sample size is

$$T_{\text{ESS}} = \frac{1 - 0.078}{1 + 0.078} \cdot 10000 = 8547.$$

For the high-correlation setting we obtained $\rho(X_1^{(t-1)}, X_1^{(t)}) = 0.979$, thus the effective sample size is considerably smaller:

$$T_{\text{ESS}} = \frac{1 - 0.979}{1 + 0.979} \cdot 10000 = 105.$$

◁

