VII - Introduction to Markov chain Monte Carlo methods

We start this chapter with a brief history of Bayesian statistics:

The emergence of Bayesian statistics has a long and interesting history dating back to 1763 when Thomas Bayes laid down the basic ideas of his new probability theory (Bayes and Price, 1763, published posthumously by Richard Price). It was rediscovered independently by Laplace (Laplace, 1774) and used in a wide variety of contexts, e.g., celestial mechanics, population statistics, reliability, and jurisprudence. However, after that it was largely ignored. A few scientists, like Bruno de Finetti and Harold Jeffreys, kept the Bayesian theory alive in the first half of the 20th century. Harold Jeffreys published the book Theory of Probability (Jeffreys, 1939), which for a long time remained the main reference for using the Bayes theorem. The Bayes theorem was used in the Second World War at Bletchley Park, United Kingdom, for cracking the German Enigma code, but its use remained classified for many years afterwards. From 1950 onwards, the tide turned towards Bayesian methods. However, the lack of proper tools to do Bayesian inference remained a challenge. The frequentist methods in comparison were simpler to implement which made them more popular. (Sharma, S., 2017. Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy. arXiv preprint arXiv:1706.01629.)

MCMC methods

Markov chain Monte Carlo (MCMC) algorithms allow to generate a Markov chain having the distribution we want to sample from (called the "target distribution") as invariant distribution.

The existence of MCMC methods dates back to Metropolis et al. (1953) and Hasting (1970). However, it is only the (relatively) recent increases of the computational power that has made these sampling techniques (and hence Bayesian inference) widely applicable.

Nowadays, MCMC methods in general, and the Metropolis-Hastings algorithm in particular, are the most popular tools used in practice to approximate the posterior distributions arising in Bayesian statistics.

In this chapter we present the Metropolis-Hastings (M-H) algorithm and one of its variant, the Gibbs sampler.

We first present theses algorithms and develop the theory in the case where the target distribution has a finite support. Then, we give their general versions and state (without any proofs) the corresponding main theoretical results.

Notation

Let $\mathcal{Y} = \{1, \ldots, m\}$ for some $m \in \mathbb{N}$. Each $i \in \mathcal{Y}$ is called a state and $\mathcal{Y}$ is called the state space. Let $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions on $\mathcal{Y}$.

We say that a matrix $P = (p_{ij},\, i, j \in \mathcal{Y})$ is stochastic if every row is a distribution on $\mathcal{Y}$; that is

$$\sum_{j=1}^{m} p_{ij} = 1, \quad \min_{j \in \mathcal{Y}} p_{ij} \geq 0, \quad \forall i \in \mathcal{Y}.$$

Definition 7.1 *We say that $(Y_t)_{t \geq 0}$ is a Markov chain with initial distribution $\lambda_0 \in \mathcal{P}(\mathcal{Y})$ and transition matrix P if*

1. *Y_0 has distribution λ_0*
2. *For all $t \geq 0$ and $(i_0, \ldots, i_{t+1}) \in \mathcal{Y}^{t+2}$,*

$$\mathbb{P}(Y_{t+1} = i_{t+1} | Y_t = i_t, \ldots, Y_0 = i_0) = \mathbb{P}(Y_{t+1} = i_{t+1} | Y_t = i_t)$$

3. *For all $t \geq 0$ and $(i, j) \in \mathcal{Y}^2$,*

$$\mathbb{P}(Y_{t+1} = j | Y_t = i) = p_{ij}.$$

We say that $(Y_t)_{t \geq 0}$ is Markov(λ_0, P) in short.

For a Markov(λ_0, P) process $(Y_t)_{t \geq 0}$ we use below the shorthand

$$p_{ij}(t) = \mathbb{P}(Y_t = j | Y_0 = i), \quad t \geq 1, \quad (i, j) \in \mathcal{Y}^2$$

and, for $t \geq 0$, we denote by λ_t the marginal distribution of Y_t; that is

$$\lambda_t := \big(\mathbb{P}(Y_t = 1), \ldots, \mathbb{P}(Y_t = m)\big) \in \mathcal{P}(\mathcal{Y}). \tag{1}$$

Lastly, for $t \geq 0$ we let $p_{ij}^{(t)}$ be the element (i, j) of P^t, so that $P^t = (p_{ij}^{(t)})_{i,j=1}^{m}$.

Some key definitions and properties

The following proposition collects two simple results.

Proposition 7.1 *For any $t \geq 1$, $p_{ij}(t) = p_{ij}^{(t)}$ for all $(i,j) \in \mathcal{Y}^2$ and, for any $t \geq 0$, $\lambda_t^T = \lambda_0^T P^t$.*

Proof: Done in class.

Definition 7.2 *P is irreducible if for any $(i,j) \in \mathcal{Y}^2$ there exists a $t \geq 1$ such that $p_{ij}^{(t)} > 0$.*

In words, P is irreducible if it is possible to go to any state j from any state i.

Definition 7.3 *P is aperiodic if, for all $i \in \mathcal{Y}$, we have $p_{ii}^{(t)} > 0$ for all sufficiently large t.*

In words, P is aperiodic if for any i the event $\{Y_t = i\}$ can happen at irregular times.

Definition 7.4 *A probability measure $\mu \in \mathcal{P}(\mathcal{Y})$ is invariant for P if $\mu^T P = \mu^T$.*

Remark: If μ is invariant for P and $(Y_t)_{t \geq 0}$ is Markov(μ, P) then $Y_t \sim \mu$ for all $t \geq 0$.

An invariant distribution μ of P is often called stationary/equilibrium distribution of P because of the following result.

Theorem 7.1 *Assume that, for some $i \in \mathcal{Y}$, $\lim_{t \to +\infty} p_{ij}^{(t)}$ exists for all $j \in \mathcal{Y}$. Then,*

$$\mu := \left(\lim_{t \to +\infty} p_{ij}^{(t)},\ j \in \mathcal{Y} \right)$$

is an invariant distribution of P.

Proof: Done in class.

Convergence of Markov chains to equilibrium

Definitions 7.3 and 7.4 are important because of the following theorem.

Theorem 7.2 *Let P be an irreducible and aperiodic stochastic matrix with invariant distribution $\mu \in \mathcal{P}(\mathcal{Y})$. Let $\lambda_0 \in \mathcal{P}(\mathcal{Y})$ and $(Y_t)_{t\geq 0}$ be Markov(λ_0, P). Then, there exist constants $\rho \in (0,1)$ and $c \in (0, +\infty)$ such that, for all $(i,j) \in \mathcal{Y}^2$ and $t \geq 0$,*

$$|p_{ij}^{(t)} - \mu_j| \leq c\,\rho^t \quad \text{and} \quad \big|\mathbb{P}(Y_t = j) - \mu_j\big| \leq c\,\rho^t.$$

Proof: See Appendix 1.

Remark: Theorem 7.2 implies that if P is irreducible and aperiodic then P has at most one invariant distribution. In fact, it can be shown that an irreducible and aperiodic stochastic matrix has a unique invariant distribution (see Problem Sheet 5).

Remark: Theorem 7.2 implies that if $(Y_t)_{t\geq 0}$ is Markov(λ_0, P) for some irreducible and aperiodic stochastic matrix P then, as $t \to +\infty$, $Y_t \overset{\text{dist.}}{\Rightarrow} \mu$ where μ is the unique invariant distribution of P.

Corollary 7.1 *Consider the set-up of Theorem 7.2 and let $\varphi : \mathcal{Y} \to \mathbb{R}$. Then,*

$$\lim_{T\to+\infty} \mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t) - \sum_{i=1}^{m}\varphi(i)\mu_i\Big)^2\Big] = 0.$$

Proof: See Appendix 2.

Remark: It is also possible to show a strong law of large numbers, namely that

$$\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t) \to \sum_{i=1}^{m}\varphi(i)\mu_i, \quad \mathbb{P}\text{-almost surely.}$$

Building a Markov chain having the 'right' invariant distribution

So far we took P as given and assumed that P had an invariant distribution μ.

We now consider the converse problem: Given μ, the distribution we are interested in, how can we construct a transition matrix P such that P has μ as invariant distribution?

Before answering this question we need the following simple lemma.

Lemma 7.1 *Let $\mu \in \mathcal{P}(\mathcal{Y})$ and assume that there exists a transition matrix P such that*

$$\mu_i p_{ij} = \mu_j p_{ji}, \quad \forall (i,j) \in \mathcal{Y}^2. \tag{2}$$

Then, μ is an invariant distribution of P.

Proof: Done in class.

Condition (2) is known as the detailed balance condition. It implies that, at equilibrium (i.e. when $Y_t \sim \mu$), the joint probability $\mathbb{P}(Y_t = i, Y_{t+1} = j)$ is symmetric in t and $t+1$. When there exists a $\mu \in \mathcal{P}(\mathcal{Y})$ such that (2) holds we say that the Markov(λ_0, P) process $(Y_t)_{t \geq 0}$ is reversible.

Lemma 7.1 shows that we can construct a Markov chain having μ as invariant distribution provided that we can construct a transition matrix P such that (2) holds.

Surprisingly, not only it is always feasible to construct such a matrix P, but it is (very) easy using the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm

Theorem 7.3 *Let $Q = (q_{ij},\, i,j \in \mathcal{Y})$ be a transition matrix such that $q_{ij} > 0$ for all $(i,j) \in \mathcal{Y}^2$, $\mu \in \mathcal{P}(\mathcal{Y})$ be such that $\mu_i > 0$ for all $i \in \mathcal{Y}$ and $P^{\mathrm{MH}} = (p_{ij}^{\mathrm{MH}},\, i,j \in \mathcal{Y})$ be such that, for every $i \in \mathcal{Y}$,*

$$p_{ij}^{\mathrm{MH}} = \begin{cases} q_{ij}\, \min\left\{1, \frac{\mu_j\, q_{ji}}{\mu_i\, q_{ij}}\right\}, & j \neq i \\ 1 - \sum_{k \neq i} q_{ik}\, \min\left\{1, \frac{\mu_k\, q_{ki}}{\mu_i\, q_{ik}}\right\}, & j = i. \end{cases}$$

Then, P^{MH} is irreducible, aperiodic and has μ as unique invariant distribution.

Proof: Done in class.

We are now ready to write down the famous M-H algorithm.

Metropolis-Hastings algorithm (A1)

Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$ and a transition matrix Q on $\mathcal{Y}$

Set $Y_0 = y_0$

for $t \geq 1$ **do**

$\tilde{Y}_t \sim Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$

Set $Y_t = \tilde{Y}_t$ with probability $\alpha(Y_{t-1}, \tilde{Y}_t)$ and $Y_t = Y_{t-1}$ with probability $1 - \alpha(Y_{t-1}, \tilde{Y}_t)$.

end for

Notation: For $y \in \mathcal{Y}$ the notation $\tilde{Y} \sim Q(y, \mathrm{d}\tilde{y})$ means that the random variable $\tilde{Y}$ is such that $\mathbb{P}(\tilde{Y} = j) = q_{yj}$ while the mapping $\alpha : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ is defined by

$$\alpha(i,j) = \min\left\{1, \frac{\mu_j\, q_{ji}}{\mu_i\, q_{ij}}\right\}, \quad i,j \in \mathcal{Y}.$$

Remark: By Theorem 7.3, the M-H algorithm (A1) defines a Markov$(\delta_{y_0}, P_\mu^{\mathrm{MH}})$ process having μ as unique invariant distribution and such that the results of Theorem 7.2 and of Corollary 7.1 hold.

The Metropolis-Hastings algorithm on a general state space

The extension of the M-H algorithm (A1) to an arbitrary state space $\mathcal{Y}$ is straightforward:

Metropolis-Hastings algorithm (A2)

Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$ and a transition kernel Q on $\mathcal{Y}$.

Set $Y_0 = y_0$

for $t \geq 1$ **do**

$\tilde{Y}_t \sim Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$

Set $Y_t = \tilde{Y}_t$ with probability

$$\alpha(Y_{t-1}, \tilde{Y}_t) = \min\left\{1, \frac{\mu(\tilde{Y}_t) q(Y_{t-1}|\tilde{Y}_t)}{\mu(Y_{t-1}) q(\tilde{Y}_t|Y_{t-1})}\right\}$$

and $Y_t = Y_{t-1}$ otherwise.

end for

Remark: We abandon the notion of transition matrix to adopt the more general one of transition kernel.

Notation: $q(\tilde{y}|y)$ is the density of $Q(y, \mathrm{d}\tilde{y})$.

Jargon: $Q(y, \mathrm{d}\tilde{y})$ is often called the proposal distribution.

Technical remark: Now $\mathcal{P}(\mathcal{Y})$ is the set of probability distributions on $\mathcal{Y}$ that are absolutely continuous w.r.t. $\mathrm{d}y$.

Invariant distributions, irreducibility and aperiodicity for general state spaces

We start with the general definition of an invariant distribution.

Definition 7.5 *A probability measure $\mu \in \mathcal{P}(\mathcal{Y})$ is an invariant distribution for the transition kernel P if*

$$\int_{\mathcal{Y}} p(y|y')\mu(y')\mathrm{d}y' = \mu(y).$$

Recall that, when $\mathcal{Y}$ is finite, a transition matrix P is irreducible if it is possible to go to any state $j \in \mathcal{Y}$ from any state $i \in \mathcal{Y}$. If $\mathcal{Y}$ is a continuous state space such a requirement is impossible to full-fill and we therefore need to weaken the notion of irreducibility.

Definition 7.6 *Given a measure φ, the Markov chain $(Y_t)_{t\geq 0}$ with transition kernel P is φ-irreducible if, for every $y \in \mathcal{Y}$ and (measurable) set $A \subset \mathcal{Y}$ with $\varphi(A) > 0$, there exists a $t \geq 0$ such that $P^t(y, A) > 0$.*

In words, P is φ-irreducible if it is possible to go to any (measurable) set $A \subset \mathcal{Y}$ with $\varphi(A) > 0$ from any state $y \in \mathcal{Y}$.

Similarly, the notion of aperiodicity needs to be weakened.

Definition 7.7 *A Markov chain $(Y_t)_{t\geq 0}$ with transition kernel P and stationary distribution μ is aperiodic if there do not exist a $p \geq 2$ and disjoints subsets $\mathcal{Y}_1, \ldots, \mathcal{Y}_p \subset \mathcal{Y}$ with $P(y, \mathcal{Y}_{i+1}) = 1$ for all $y \in \mathcal{Y}_i$ ($i \in \{1, \ldots, p-1\}$), $P(y, \mathcal{Y}_1) = 1$ for all $y \in \mathcal{Y}_p$, and such that $\mu(\mathcal{Y}_1) > 0$ (and hence $\mu(\mathcal{Y}_i) > 0$ for all i).*

A general convergence result for M-H algorithms

We first show that the M-H algorithm (A2) indeed defines a Markov chain having μ as invariant distribution.

Lemma 7.2 *Let $\mu \in \mathcal{P}(\mathcal{Y})$ and assume that there exists a transition kernel P on $\mathcal{Y}$ such that*

$$\mu(y)p(\tilde{y}|y) = \mu(\tilde{y})p(y|\tilde{y}), \quad \forall (y, \tilde{y}) \in \mathcal{Y}^2. \tag{3}$$

Then, μ is an invariant distribution of P.

Proof: Obvious.

Corollary 7.2 *The Markov chain $(Y_t)_{t\geq 0}$ defined by the M-H algorithm (A2) admits μ as invariant distribution.*

Proof: Done in class.

The following result provides a simple way to check the validity of the M-H algorithm (A2)[a]

Theorem 7.4 *Consider the M-H algorithm (A2). Assume that $Q(y, A) > 0$ for all $y \in \mathcal{Y}$ and all (measurable) set $A \subset \mathcal{Y}$ such that $\mu(A) > 0$, and that*

$$\mathbb{P}\left(\frac{\mu(\tilde{Y}_t)q(Y_{t-1}|\tilde{Y}_t)}{\mu(Y_{t-1})q(\tilde{Y}_t|Y_{t-1})} < 1\right) > 0, \quad \forall t \geq 1.$$

Then, the resulting Markov chain $(Y_t)_{t\geq 0}$ is μ-irreducible and aperiodic, and consequently

$$\lim_{t\to+\infty} \mathbb{P}(Y_t \in A) = \mu(A)$$

for any measurable sets $A \subset \mathcal{Y}$.

[a] See e.g. Theorem 7.4, p.274, of Robert, C.P. and Casella, G. *Monte Carlo Statistical Methods. Springer-Verlag New York (2004).*

Central limit theorem for Markov chains

Let $(Y_t)_{t\geq 0}$ and $(Y'_t)_{t\geq 0}$ be two Markov chains defined by the M-H algorithm (A2) where the former is such that $Y_0 = y_0$ for some $y_0 \in \mathcal{Y}$ while the latter is such that $Y'_0 \sim \mu$ (the proposal distribution $Q(y, \mathrm{d}\tilde{y})$ being the same for the two processes).

Let $\mu(\varphi) := \int_{\mathcal{Y}} \varphi(y)\mu(y)\mathrm{d}y$ for some (measurable) function $\varphi : \mathcal{Y} \to \mathbb{R}$ verifying $\mu(\varphi^2) < +\infty$ and let

$$\hat{\mu}_T(\varphi) = \frac{1}{T}\sum_{t=1}^{T} \varphi(Y_t)$$

be an estimator of $\mu(\varphi)$.

Then, under some conditions, the following central limit theorem holds

$$\sqrt{T}\,\frac{\hat{\mu}_T(\varphi) - \mu(\varphi)}{\sigma} \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_1(0,1), \quad \text{as } T \to +\infty$$

with

$$\sigma^2 = \mathrm{Var}_\mu(\varphi(Y'_0)) + 2\sum_{t=1}^{\infty} \mathrm{Cov}\big(\varphi(Y'_0), \varphi(Y'_t)\big) = \mathrm{Var}(\varphi(Y'_0))\tau_\varphi$$

and where $\tau_\varphi = 1 + 2\sum_{t=1}^{\infty} \mathrm{Corr}\big(\varphi(Y'_0), \varphi(Y'_t)\big)$ is called the integrated auto-correlation time.

Remark: The asymptotic variance depends only on Q and not on y_0 (which is intuitive).

Choosing the proposal distribution $Q(y, \mathrm{d}\tilde{y})$

The choice of the proposal distribution $Q(y, \mathrm{d}\tilde{y})$ is important because

1. It influences the mixing time of the Markov chain, that is the speed at which it converges to its stationary distribution (the constant ρ in Theorem 7.2).
2. It influences the asymptotic variance of the estimator $\hat{\mu}_T(\varphi)$ of $\mu(\varphi)$ through the integrated auto-correlation time τ_φ.

The first point is particularly important because in practice we can only run algorithm (A2) for a finite number T of iterations. If Q is poorly chosen then the distribution of Y_T will be 'far away' from the equilibrium distribution μ and the output of the algorithm will do a poor job at approximating μ.

Finding a good proposal distribution Q is both difficult and problem dependent: a given Q may perform well for some target distributions and very poorly for others.

In practice, the only solution is often to tune Q manually; that is, to try different proposal distributions until the output of the algorithm (A2) suggests that the algorithm has converged (in the sense that Y_T is approximatively distributed according to μ).

Of course, we can never be sure that the M-H algorithm has converged but there are some ways to detect bad choices for Q, and notably the inspection of

1. the acceptance rate
2. the trace plots
3. the autocorrelation functions.

We explain these three complementary approaches in what follows.

Assessing the convergence using the acceptance rate

The first approach that can be used to assess the convergence of the M-H algorithm is to look at the acceptance rate:

$$r_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(y_t = \tilde{y}_t).$$

Indeed,

- A low value for this quantity indicates that the simulated trajectory $\{y_t\}_{t=1}^T$ remains for a long time at a given location before moving to a new state.
- A high acceptance rate usually (but not always) arises when Q is such that, with high probability, $\tilde{Y}_t$ is very 'close' to Y_{t-1}.

Hence, both a low and a large value of the acceptance rate is a sign that the mixing time of the Markov chain (i.e. the time needed to be close to equilibrium) is large and that the correlation between Y_t and Y_{t-k} is important even for large values of k.

There exist theoretical results suggesting that the "optimal" acceptance rate is 0.234. In practice, choosing a proposal distribution Q such that $r_T \approx 0.234$ works well.

The main advantage of this approach is its simplicity. However, by summarizing the behaviour of the Markov chain using a single number, the acceptance rate may hide important differences in the mixing times of the chain along its different coordinates.

Assessing the convergence using the trace plots and the autocorrelation functions

Let $Y_{i,t}$ be the i-th coordinate of Y_t.

Then,

- The trace plot represents the simulated trajectory $\{y_{i,t}\}_{t=1}^{T}$ as a function of t.
- The auto-correlation function (ACF) returns, for integer $k \geq 0$, an estimate $\hat{\gamma}_T(k)$ of $\text{Corr}(Y'_{i,0}, Y'_{i,k})$ where, as per above, the Markov chain $(Y'_t)_{t\geq 0}$ has the same transition as $(Y_t)_{t\geq 0}$ but is such that $Y'_0 \sim \mu$.

  For instance,

$$\hat{\gamma}_T(k) = \frac{\frac{1}{T-k}\sum_{s=k+1}^{T}\left(y_{i,t} - \bar{y}_{i,T}\right)\left(y_{i,t-k} - \bar{y}_{i,T}\right)}{\sqrt{\frac{1}{T}\sum_{s=1}^{T}\left(y_{i,t} - \bar{y}_{i,T}\right)^2 \frac{1}{T-k}\sum_{s=k+1}^{T}\left(y_{i,t-k} - \bar{y}_{i,T}\right)^2}}$$

  with

$$\bar{y}_{i,T} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}.$$

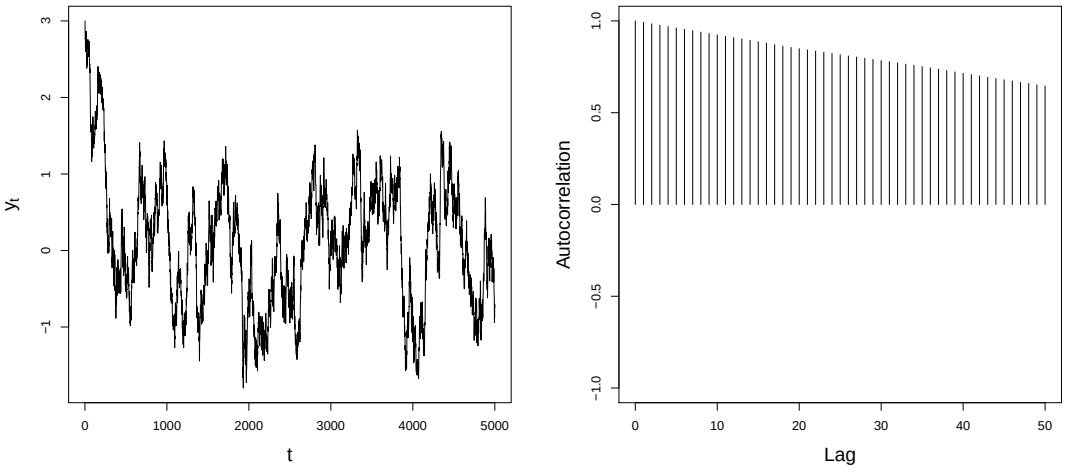
Looking at the trace plots and at the autocorrelation function therefore allows to assess the convergence of the Markov chain coordinate by coordinate.

The M-H algorithm: A simple example

Let $\mathcal{Y} = \mathbb{R}$, $\mu(y)$ be the p.d.f. of the $\mathcal{N}_1(0,1)$ distribution and $Q_\sigma(y, \mathrm{d}\tilde{y}) = q_\sigma(\tilde{y}|y)\mathrm{d}\tilde{y}$ with $q_\sigma(\tilde{y}|y)$ the p.d.f. of the $\mathcal{N}_1(0,\sigma^2)$ distribution.

Below we use M-H algorithm (A2) based on the transition kernel Q_σ and starting at $y_0 = 3$ to generate a sample that approximates μ.

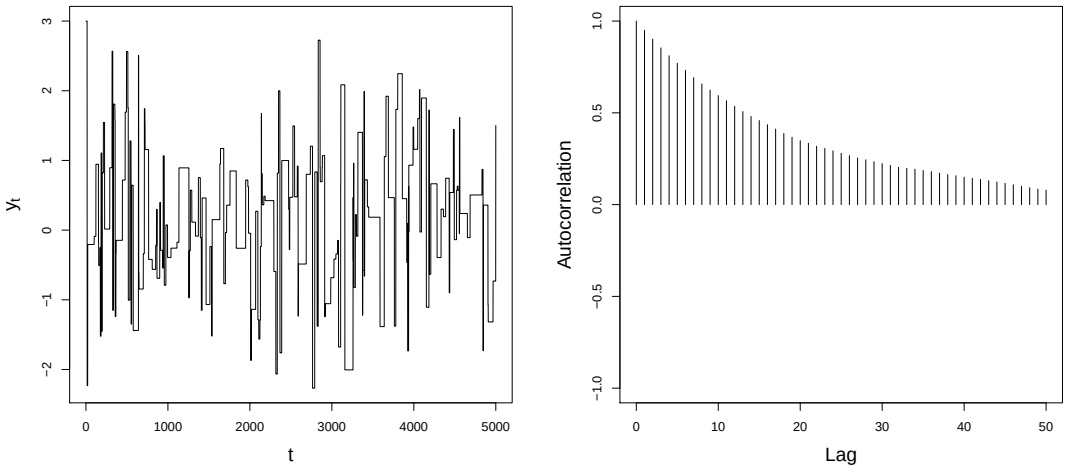
Results for $\sigma = 0.1$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.9736$.

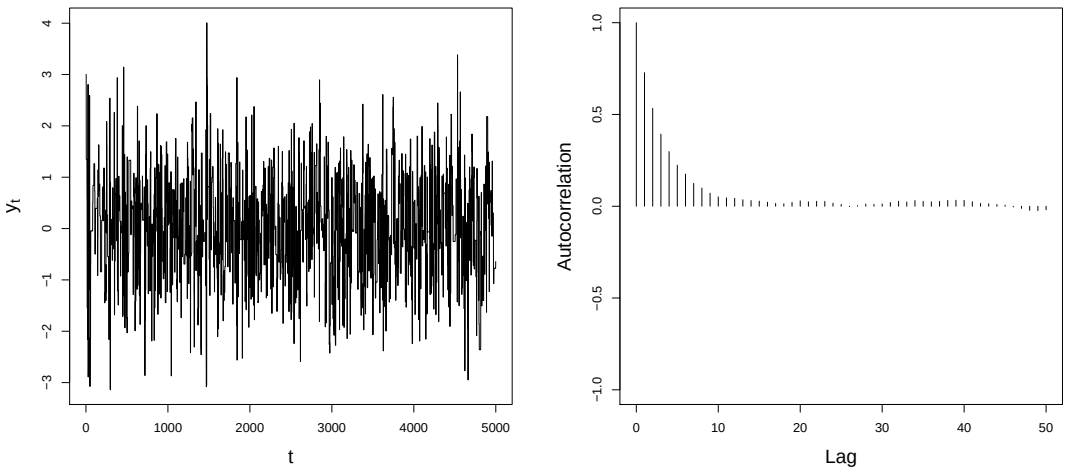
The M-H algorithm: A simple example (end)

Results for $\sigma = 40$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.0364$.

Results for $\sigma = 5$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.2402$.

The Gibbs sampler

The Gibbs sampler, which can be used only when the distribution μ we want to sample from is multivariate, is a particular case of the M-H algorithm (A2) where $Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$ is such that we have $\mathbb{P}(\alpha(Y_{t-1}, \tilde{Y}_t) = 1) = 1$ for all $t \geq 1$ (see Problem Sheet 5).

Let $\mathcal{Y} = \times_{i=1}^{d} \mathcal{Y}_i$ for some (measurable) sets $\mathcal{Y}_i \subset \mathbb{R}^{d_i}$, $y^{(i)} \in \mathcal{Y}_i$ be the i-th 'block' of $y \in \mathcal{Y}$, so that

$$y = (y^{(1)}, \ldots, y^{(d)}).$$

We also define $y^{(k:p)} = (y^{(k)}, \ldots, y^{(p)})$ for integers $1 \leq k < p \leq d$ and denote by $y^{(i)}$ the vector y without its i-th block, that is (with obvious conventions when $i \in \{1, d\}$).

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

For $\mu \in \mathcal{P}(\mathcal{Y})$, $i \in \{1, \ldots, d\}$ and $y \in \mathcal{Y}$, we let $\mu^{(i)}(\cdot | y^{(-i)})$ be the p.d.f. on $\mathcal{Y}_i$ defined by

$$\mu^{(i)}(\tilde{y} | y^{(-i)}) = \frac{\mu\big(y^{(1:i-1)}, \tilde{y}, y^{(i+1):d}\big)}{\int_{\mathcal{Y}_i} \mu\big(y^{(1:i-1)}, z, y^{(i+1):d}\big) \mathrm{d}z}, \quad \forall \tilde{y} \in \mathcal{Y}_i$$

In words, $\mu^{(i)}(\cdot | y^{(-i)})$ is the p.d.f. of the distribution of $Y^{(i)}$ under μ, conditional to $Y^{(-i)} = y^{(-i)}$.

The Gibbs sampler on a general state space: The algorithm

Using the above notation the Gibbs sampler on a general state space $\mathcal{Y}$ works as follows.

Gibbs sampler (A3)

Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$

Set $Y_0 = y_0$

for $t \geq 1$ **do**

  for $i = 1, \ldots, d$ **do**

    $Y_t^{(i)} \sim \mu^{(i)}(y_t^{(i)} | Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)}) \mathrm{d}y_t^{(i)}$

  end for

end for

The main advantage of the Gibbs sampler is that it does not require to choose a proposal distribution Q; that is, implementing the Gibbs sampler only requires to specify the distribution of interest $\mu \in \mathcal{P}(\mathcal{Y})$ and a starting value $y_0 \in \mathcal{Y}$.

However:

1. The resulting Markov chain may have a large mixing time if the correlation among the different coordinates is important (see the example below).
2. To implement algorithm (A3) we must be able to sample from the full conditional distribution $\mu^{(i)}(\cdot | y^{(-i)})$ for all i, which is rarely the case in practice (see below for a solution to this problem).

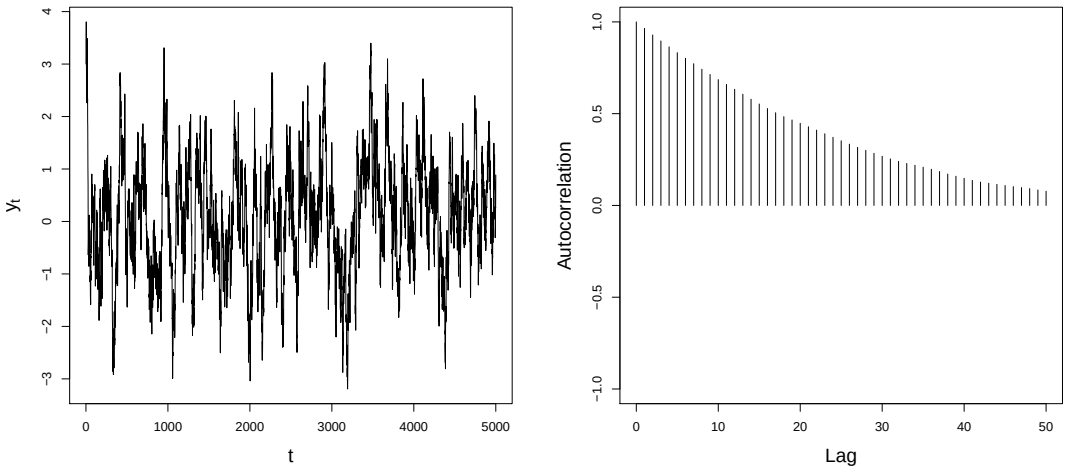
Gibbs sampler: A simple example

Let $\mathcal{Y} = \mathbb{R}^2$ and $\mu(y)$ be p.d.f. of the $\mathcal{N}_2\big(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\big)$ distribution.

Then, $\mu^{(i)}(\cdot|y^{(-i)})$ is the p.d.f. of the $\mathcal{N}_1(\rho\, y^{(j)}, 1-\rho^2)$ distribution, $i \in \{1, 2\}$.

Below we use the Gibbs sampler (A3) starting at $y_0 = (3, 3)$ to generate a sample that approximates μ for different values of ρ.

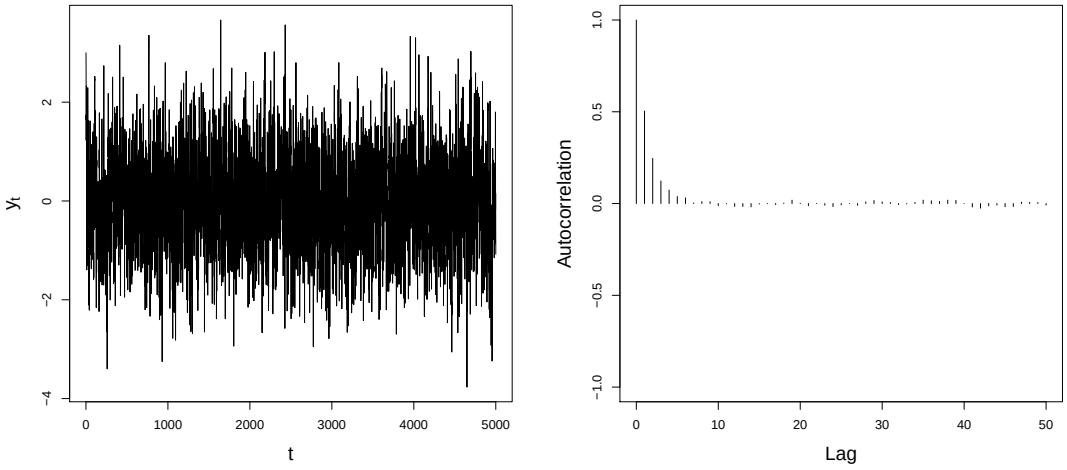
Results for $\rho = 0.98$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

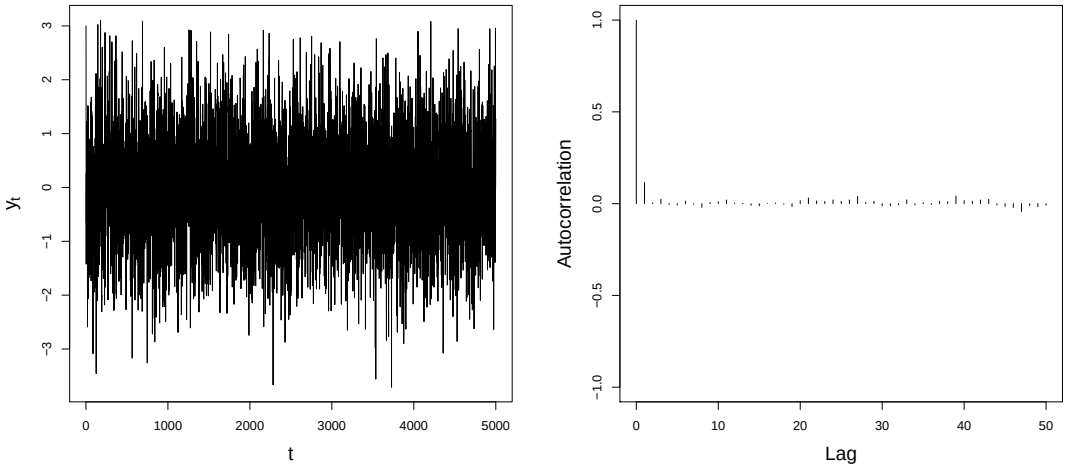
Gibbs sampler: A simple example

Results for $\rho = 0.7$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

Results for $\rho = 0.3$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

"Partial" Gibbs sampling

As mentioned above, the Gibbs sampler (A3) requires to be able to sample from $\mu^{(i)}(\cdot|y^{(-i)})$ for all i, which is rarely the case in practice.

We now assume that we can simulate from $\mu^{(i)}(\cdot|y^{(-i)})$ only for $i = 1, \ldots, d_1$, for some $d_1 < d$.

In this context, the following modified Gibbs sampler can be used.

"Partial" Gibbs sampler (A4)

Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$

Set $Y_0 = y_0$

for $t \geq 1$ **do**

  for $i = 1, \ldots, d_1$ **do**

    $Y_t^{(i)} \sim \mu^{(i)}(y_t^{(i)}|Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)})\mathrm{d}y_t$

  end for

  for $i = d_1 + 1, \ldots, d_2$ **do**

    $Y_t^{(i)} \sim P_{\mu^{(i)}(y_t^{(i)}|Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)})}(Y_{t-1}^{(i)}, \mathrm{d}y_t^{(i)})$

  end for

end for

Notation: We denote by P_η a transition kernel having η as invariant distribution.

Typically, $P_{\mu^{(i)}(\tilde{y}^{(i)}|y^{(1:i-1)}, y_{t-1}^{(i+1:d)})}$ is taken to be a M-H kernel and the resulting algorithm is known as the Metropolis-within-Gibbs algorithm.

Gibbs sampler v.s. Metropolis-within-Gibbs: An example

Let $\mathcal{Y} = \mathbb{R}^2$ and $\mu(y)$ be the p.d.f. of the $\mathcal{N}_2\big(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\big)$ distribution.

Below we use the Gibbs sampler and a Metropolis-within-Gibbs algorithm, both starting at $y_0 \in \mathcal{Y}$, to generate a sample that approximates μ.

The Metropolis-within-Gibbs algorithm we consider is as follows.

Metropolis-within-Gibb for the Bivariate Gaussian example

Input: $y_0 \in \mathcal{Y}$

Set $Y_0 = y_0$

for $t \geq 1$ **do**

$Y_t^{(1)} \sim \mathcal{N}_1(\rho Y_{t-1}^{(2)}, 1-\rho^2)$

$\tilde{Y}_t^{(2)} \sim \mathcal{N}_1(Y_{t-1}^{(2)}, \sigma^2)$

Set $Y_t^{(2)} = \tilde{Y}_t^{(2)}$ with probability

$$\min\left\{1, \frac{\varphi\big(\tilde{Y}_t^{(2)}; \rho Y_t^{(1)}, 1-\rho^2\big)}{\varphi\big(Y_{t-1}^{(2)}; \rho Y_t^{(1)}, 1-\rho^2\big)}\right\}$$

and $Y_t^{(2)} = Y_{t-1}^{(2)}$ otherwise.

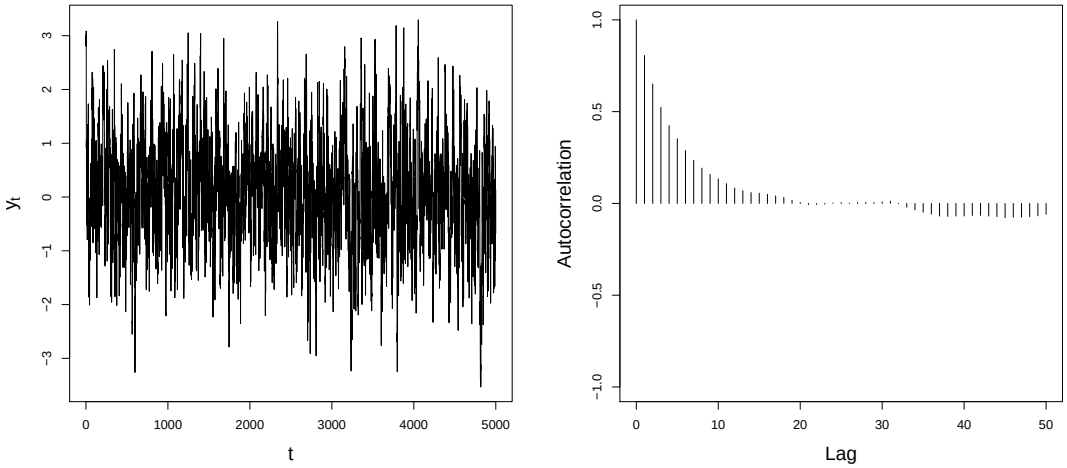
end for

Notation: $\varphi(\cdot; \mu, \sigma^2)$ denotes the p.d.f. of the $\mathcal{N}_1(\mu, \sigma^2)$ distribution.

Gibbs sampler v.s. Metropolis-within-Gibbs: An example (end)

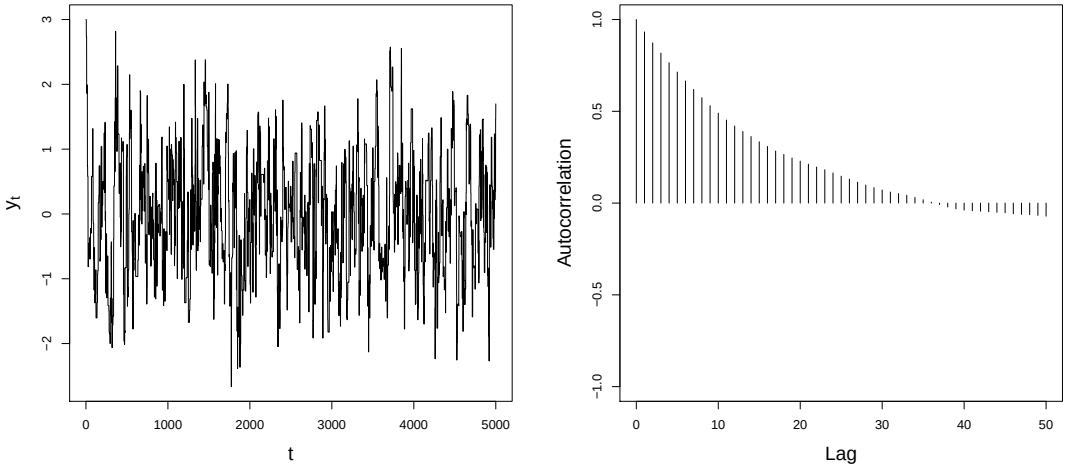
Let $y_0 = (3, 3)$, $\sigma = 2$ and $\rho = 0.9$.

Results for the Gibbs sampler:



Trace plot (left) and ACP (right) for $(y_t^{(2)})_{t\geq 0}$.

Results for the Metropolis-within-Gibbs algorithm:



Trace plot (left) and ACP (right) for $(y_t^{(2)})_{t\geq 1}$. The acceptance rate for $\{y_t^{(2)}\}_{t=1}^{T}$ is $r_T = 0.2586$.

Appendix 1: Proof of Theorem 7.2

We start with a preliminary result.

Lemma 7.3 *Let P be irreducible and aperiodic. Then, there exists a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$.*

Proof: Let $(i, j, k) \in \mathcal{Y}^3$ and remark first that, as P is irreducible, there exist a $t_1 \geq 1$ and a $t_2 \geq 1$ such that $p_{ij}^{(t_1)} > 0$ and $p_{ki}^{(t_2)} > 0$. In addition, as P is aperiodic, there exists a $t_3 \geq 1$ such that $p_{ii}^{(t)} > 0$ for all $t \geq t_3$.

Let $(Y_t)_{t\geq 0}$ be Markov(λ_0, P). Then, using the above observations, the Markov property of $(Y_t)_{t\geq 0}$ and Proposition 7.1, for any $t \geq t_3$ we have

$$\begin{aligned}
p_{kj}^{(t_1+t_2+t)} &= \mathbb{P}(Y_{t_1+t_2+t} = j | Y_0 = k) \\
&\geq \mathbb{P}(Y_{t_1+t_2+t} = j, Y_{t_2+t} = i, Y_{t_2} = i | Y_0 = k) \\
&= \mathbb{P}(Y_{t_1+t_2+t} = j | Y_{t_2+t} = i)\mathbb{P}(Y_{t_2+t} = i | Y_{t_2} = i)\mathbb{P}(Y_{t_2} = i | Y_0 = k) \\
&= p_{ij}^{(t_1)}\, p_{ii}^{(t)}\, p_{ki}^{(t_2)} \\
&> 0
\end{aligned}$$

showing that $p_{kj}^{(t)} > 0$ for all $t \geq t_1 + t_2 + t_3$. The result follows.

Appendix 1: Proof of Theorem 7.2 (continued)

We now prove[a] Theorem 7.2 and assume first that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$. Let $\mathbf{1} = (1, \ldots, 1)$ and $\Pi = \mathbf{1}\mu^T$. Note that

$$P\Pi = \Pi P = \Pi^2 = \Pi. \tag{4}$$

Note also that, because $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$, there exists an $\alpha \in (0, 1)$ such that all the elements of the matrix $P - \alpha\Pi$ are non-negative.

Let

$$\tilde{P} = \frac{1}{1-\alpha}(P - \alpha\Pi)$$

so that, since all the elements of $\tilde{P}$ are non-negative and $\tilde{P}\mathbf{1} = \mathbf{1}$, $\tilde{P}$ is a stochastic matrix. Note also that, by (4),

$$\tilde{P}\Pi = \Pi\tilde{P} = \Pi. \tag{5}$$

Next, let $t \geq 1$. Then,

$$\begin{aligned} P^t &= \Big((1-\alpha)\tilde{P} + \alpha\Pi\Big)^t \\ &= (1-\alpha)^t\tilde{P}^t + \sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\tilde{P}^{t-s}\alpha^s\Pi^s \\ &= (1-\alpha)^t\tilde{P}^t + \Pi\sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\alpha^s \\ &= (1-\alpha)^t\tilde{P}^t + \Pi\big(1-(1-\alpha)^t\big) \end{aligned} \tag{6}$$

where the second equality uses the Binomial expansion (that holds because the matrices $\tilde{P}$ and Π commute, by (5)), the third equality uses (5) and the last equality uses the fact that the sum is equal to $1 - \mathbb{P}(Z = t)$ where $Z \sim \text{Binomial}(t, 1-\alpha)$.

[a] This proof is due to Prof. Balint Toth

Appendix 1: Proof of Theorem 7.2 (end)

To proceed further for a matrix $A = (a_{ij})$ we let $|A| = (|a_{ij}|)$.

Then, using (6),

$$|P^t - \Pi| = (1-\alpha)^t |\tilde{P}^t - \Pi| \leq (1-\alpha)^t \mathbf{1}\mathbf{1}^T, \quad \forall t \geq 1 \tag{7}$$

where the inequality holds because $\tilde{P}$ and Π are both stochastic matrices.

Since $P^t - \Pi = (p_{ij}^{(t)} - \mu_j)_{i,j=1}^m$ inequality (7) yields

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha)^t, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t \geq 1$$

showing that the conclusion of Theorem 7.2 holds with $c = 1$ and $\rho = (1-\alpha)$ in the special case where $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$.

Assume now that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} = 0$. Then, as P is irreducible and aperiodic, there exists by Lemma 7.3 a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$. Let $P_0 = P^{t_0}$ so that, as per above, there exists an $\alpha_0 \in (0,1)$ such that all the elements of the matrix $P_0 - \alpha_0\Pi$ are non-negative.

Then, repeating the above computations with P replaced by P_0 and α replaced by α_0, we have, noting that P_0^t has elements $(p_{ij}^{(t_0+t)})_{i,j=1}^M$,

$$|p_{ij}^{(t_0+t)} - \mu_j| \leq (1-\alpha_0)^t, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t \geq 1.$$

or, equivalently,

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha_0)^{t-t_0}, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t > t_0.$$

Then, as $|p_{ij}^{(t)} - \mu_j| \leq 1$ for all $(i,j) \in \mathcal{Y}^2$ and all $t \geq 0$, the conclusion of Theorem 7.2 holds with $c = (1-\alpha_0)^{-t_0}$ and $\rho = (1-\alpha_0)$.

Appendix 2: Proof of Corollary 7.1

We first show the following result

Theorem 7.5 *Consider the set-up of Theorem 7.2. Then,*

$$\lim_{T\to+\infty} \mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{i\}}(Y_t) - \mu_i\Big)^2\Big] = 0, \quad \forall i \in \mathcal{Y}.$$

Proof: Fix i. Then,

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{i\}}(Y_t) - \mu_i\Big)^2\Big] = \frac{1}{T}\big(v_1(T) + v_2(T)\big)$$

where

$$v_1(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)^2\big]$$

and

$$v_2(T) = \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)\big].$$

Then, to prove the result it suffices to show that

$$\limsup_{T\to+\infty} v_1(T) < +\infty, \quad \limsup_{T\to+\infty} |v_2(T)| < +\infty.$$

For $v_1(T)$ we trivially have $v_1(T) \leq 1$ and thus, as required, $\limsup_{T\to+\infty} v_1(T) < +\infty$.

Appendix 2: Proof of Corollary 7.1 (continued)

We now study $v_2(T)$.

To this aim remark first that, by Proposition 7.1 and using the law of iterated expectations, for $1 \leq t < s$,

$$\begin{aligned}
&\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)\big] \\
&\qquad = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)\mathbb{E}[(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)|Y_t]\big] \\
&\qquad = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(p_{Y_t i}^{(s-t)} - \mu_i)\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
|v_2(T)| &\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)\Big]\right| \\
&= \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(p_{Y_t i}^{(s-t)} - \mu_i)\Big]\right| \\
&\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\Big[|\mathbf{1}_{\{i\}}(Y_t) - \mu_i|\,|p_{Y_t i}^{(s-t)} - \mu_i|\Big] \\
&\leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\rho^{s-t} \\
&\leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=1}^{\infty}\rho^{s} \\
&\leq \frac{2\,c}{1-\rho}
\end{aligned}$$

where the first inequality uses the triangle inequality, the second inequality uses Jensen's inequality and the third inequality the result of Theorem 7.2. The proof of Theorem 7.5 is complete.

Appendix 2: Proof of Corollary 7.1 (end)

We are now ready to prove Corollary 7.1.

Let $\varphi : \mathcal{Y} \to \mathbb{R}$ and $\mu(\varphi) = \sum_{i=1}^{m} \varphi(i)\mu_i$.

Then ($\mathbb{P}$-a.s.),

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t) - \sum_{i=1}^{m}\varphi(i)\mu_i &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{m}\varphi(i)\big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\\ &= \sum_{i=1}^{m}\varphi(i)\frac{1}{T}\sum_{t=1}^{T}\big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\end{aligned}$$

so that, using Cauchy-Schwartz's inequality, we have ($\mathbb{P}$-a.s.)

$$\Big(\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t) - \sum_{i=1}^{m}\varphi(i)\mu_i\Big)^2 \le \sum_{i=1}^{m}\varphi(i)^2\sum_{i=1}^{m}\Big(\frac{1}{T}\sum_{t=1}^{T}\big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\Big)^2$$

and therefore

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\varphi(Y_t) - \sum_{i=1}^{m}\varphi(i)\mu_i\Big)^2\Big] \le \sum_{i=1}^{m}\varphi(i)^2\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Y_t) - \mu_i\Big)^2\Big]$$

and the result follows from Theorem 7.5.