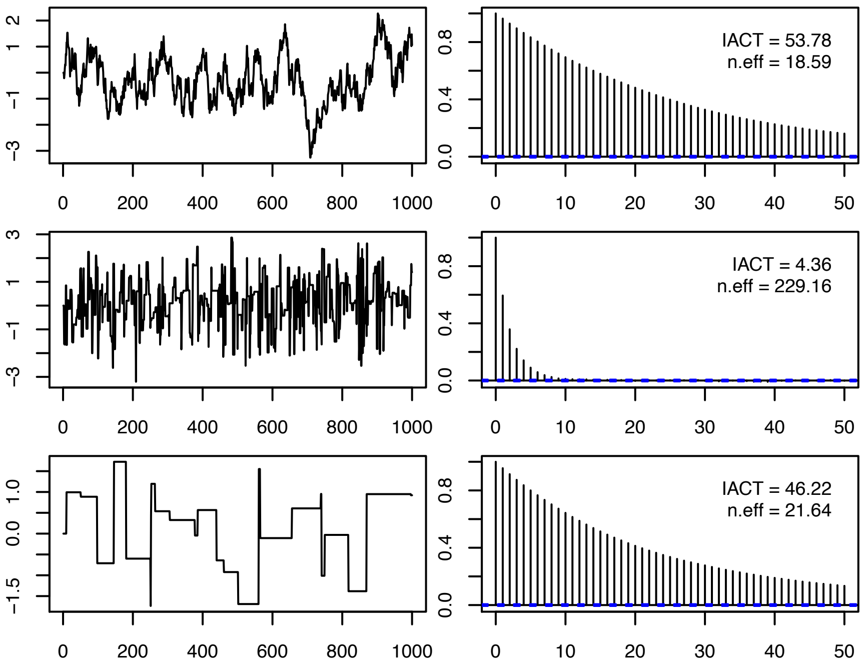


Figure 18: Sample paths and correlations of MH in Example 6.27 with $a = 0.5$ (top), $a = 5$ (middle) and $a = 50$ (bottom); here $f(x) := x$.

where $\hat{\sigma}_n^2 := (n-1)^{-1}\sum_{k=1}^n \big[f(X_k) - I_{p,q,\mathrm{MH}}(f)\big]^2 \xrightarrow{n\to\infty} \mathrm{Var}_p\big(f(X)\big)$ and β is the desired Normal quantile; cf. Proposition 1.13.

Remember to discard the burn-in samples before proceeding to (iii) and (iv). Remember also that both ACF and n_{eff} depend on the function!

7.4 Optimising MCMC (*)

Usually asymptotic variance cannot be calculated in a closed form, but comparison of asymptotic variances may be possible.

Theorem 7.11 (Peskun [21], Tierney [28]). *Suppose that P and Q are transition probabilities both reversible wrt. common distribution π. Suppose that*

$$\sum_{x,y\in\mathbb{X}} \pi(x)P(x,y)[f(x)-f(y)]^2 \geq \sum_{x,y\in\mathbb{X}} \pi(x)Q(x,y)[f(x)-f(y)]^2, \tag{19}$$

for all $f : \mathbb{S} \to \mathbb{R}$ with $\mathbb{E}_\pi[f^2(X)] < \infty$. Then, P is always better than Q in the following sense: for any function $f : \mathbb{S} \to \mathbb{R}$ with $\mathbb{E}_\pi[f^2(X)] < \infty$,

$$\lim_{n\to\infty} n\mathrm{Var}\bigg(\frac{1}{n}\sum_{k=1}^n f(X_k^{(P)})\bigg) \leq \lim_{n\to\infty} n\mathrm{Var}\bigg(\frac{1}{n}\sum_{k=1}^n f(X_k^{(Q)})\bigg),$$

where $(X_k^{(P)})_{k\geq 0}$ and $(X_k^{(Q)})_{k\geq 0}$ are stationary Markov chains with transition probabilities P and Q, respectively.

Remark 7.12. It is easy to see that

$$P(x,y) \geq Q(x,y) \qquad \text{for all } x \neq y, \tag{20}$$

implies (19). The condition (20) is referred to as the *off-diagonal order* or the *Peskun order* and (19) is known as the *covariance order*.

Remark 7.13. In the continuous case, if P and Q are in the form (14) with $k_P(x,y)$ and $k_Q(x,y)$, respectively, then the covariance order (19) corresponds to

$$\iint \pi(x)k_P(x,y)[f(x)-f(y)]^2 \mathrm{d}x\mathrm{d}y \geq \iint \pi(x)k_Q(x,y)[f(x)-f(y)]^2 \mathrm{d}x\mathrm{d}y,$$

which holds if the analogous off-diagonal order holds:

$$k_P(x,y) \geq k_Q(x,y) \qquad \text{for all } x \neq y.$$

The covariance order is equivalent with order $\mathcal{E}_P(f) \geq \mathcal{E}_Q(f)$ of Dirichet forms

$$\mathcal{E}_P(f) := \langle f, (I-P)f\rangle_\pi, \qquad \langle f, g\rangle_\pi := \int \pi(x)f(x)g(x)\mathrm{d}x,$$

where I is identity operator so $(If)(x) = f(x)$ and $(Pf)(x) = \int P(x,\mathrm{d}y)f(y)\mathrm{d}y$.

Example 7.14. In the Ising model Example 6.39, we have a choice of the proposal distribution $q_i(x, y \mid x^{(-i)})$. Note that here $x, y \in \{0,1\}$. The best choice in terms of asymptotic variance is to take $q_i(x,y \mid x^{(-i)}) = \mathbf{1}\,(y = 1-x)$, because any other choice would be worse in terms of the off-diagonal order (20).

Example 7.15 (Barker's algorithm). In the Metropolis-Hastings algorithm, we could use an alternative acceptance probability

$$\alpha_B(x,y) := \frac{r(x,y)}{r(x,y)+1}, \qquad r(x,y) := \frac{p(y)q(y,x)}{p(x)q(x,y)}.$$

Similarly as with Metropolis-Hastings, it is direct to check that

$$p(x)q(x,y)\alpha_B(x,y) = p(y)q(y,x)\alpha_B(y,x),$$

so the resulting algorithm is still reversible wrt. p.

Direct calculation shows that $\alpha_B(x,y) \leq \alpha(x,y) = \min\{1, r(x,y)\}$, which implies an off-diagonal order, so the Barker's algorithm using α_B acceptance rate is never better than Metropolis-Hastings. (There are certain situations where α_B is easier to calculate, though.)

8 Sequential Monte Carlo

We shall focus next on algorithms which operate on a *sequence* of distributions $\pi_1, \pi_2, \ldots, \pi_T$, which gradually evolve towards the distribution of interest $p = \pi_T$. The samples are often called *particles* in this context, and the key algorithm in this context is known as the *particle filter*.

We will motivate the algorithms in a time-series context, which was their original motivation, and where they have been applied extensively. We present the methods with densities on an Euclidean space; discrete case follows similarly.

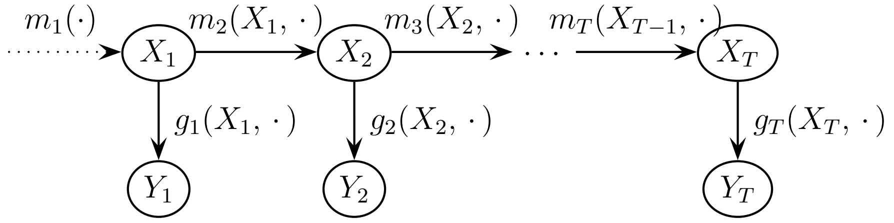


Figure 19: General state-space model.

In this section, we denote for $a \le b$ the vector $x_{a:b} = (x_a, \ldots, x_b)$. We also exceptionally denote *'time' indices* in subscript (not Monte Carlo samples as before), and superscript contain *sample indices* (not coordinates as before).

8.1 Motivation: General state-space models/hidden Markov models

Figure 19 illustrates a general state-space model. It consists of two parts:

- 'Latent' Markov chain $(X_t)_{t\ge 1}$ evolving in $\mathbb{S} = \mathbb{R}^d$ with initial density $X_1 \sim m_1$, and with conditional densities $m_t(x_{t-1}, x_t)$ of $X_t \mid (X_t = x_t)$. (Note that the transition densities may depend on time t.)
- Conditionally independent observed process $(Y_t)_{t\ge 1}$ following the observation densities $Y_t \mid X_t \sim g_t(X_t,\,\cdot\,)$.

More precisely, the model defines the joint density of the form $\hat{p}(x_{1:T}, y_{1:T}) := m_1(x_1)g_1(x_1, y_1)\prod_{t=2}^T m_t(x_{t-1}, x_t)g_t(x_t, y_t)$.

We are interested in Bayesian inference of $X_{1:T}$ having observed $Y_{1:T} = y_{1:T}$, that is, we focus on the conditional density p of $\hat{p}$:

$$p(x_{1:T}) \propto p_u(x_{1:T}) := m_1(x_1)g_1(x_1, y_1)\prod_{t=2}^T m_t(x_{t-1}, x_t)g_t(x_t, y_t), \tag{21}$$

where $y_{1:T}$ are the observed values, which are constant in our case, and omitted from the notation.

Remark 8.1. What we call state-space models (SSM), some other authors call *hidden Markov models* (HMM) [e.g. 4, 12]. Some authors reserve HMM to mean the case where X_k are discrete, taking values on a finite set. Some others reserve SSM to mean only linear(-Gaussian) models.

Example 8.2 (Noisy AR(1) process). Let $\sigma_1^2, \sigma_x^2, \sigma_y^2 \in (0,\infty)$ and $\rho \in \mathbb{R}$ be known parameters. Then, let $m_1 = N(0, \sigma_1^2)$ and for $k \ge 2$, assume $(Z_k)_{k\ge1}, (W_k)_{k\ge1} \overset{\text{i.i.d.}}{\sim} N(0,1)$, and define

$$\begin{aligned} X_k &:= \rho X_{k-1} + \sigma_x Z_k \\ Y_k &:= X_k + \sigma_y W_k. \end{aligned}$$

This corresponds to setting

$$\begin{aligned} m_k(x_{k-1}, x_k) &:= N(x_k; \rho x_{k-1}, \sigma_x^2) \\ g_k(x_k, y_k) &:= N(y_k; x_k, \sigma_y^2). \end{aligned}$$

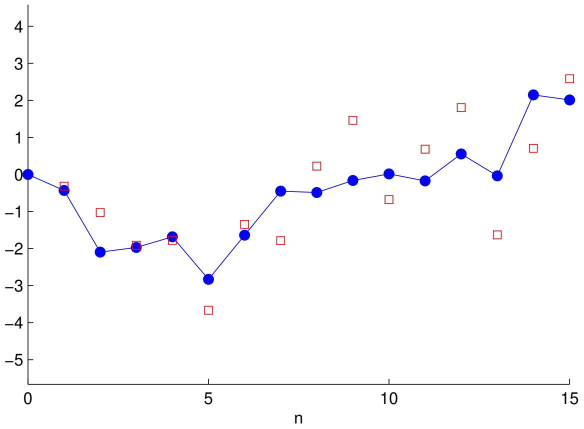


Figure 20: Sample path of the noisy AR(1) process in Example 8.2 with $\rho = 1$ and $\sigma_1^2 = \sigma_x^2 = 1 = \sigma_y^2$: The Markov chain $X_{1:15}$ in blue and the noisy observations $Y_{1:15}$ in red.

In other words, $(X_k)_{k\geq 1}$ is an AR(1) process.[15] Given a realisation of the process $(X_1, \ldots, X_T)$, the observations are conditionally independent and perturbed by Gaussian increments with variance σ_y^2. Figure 20 shows an example realisation of the process.

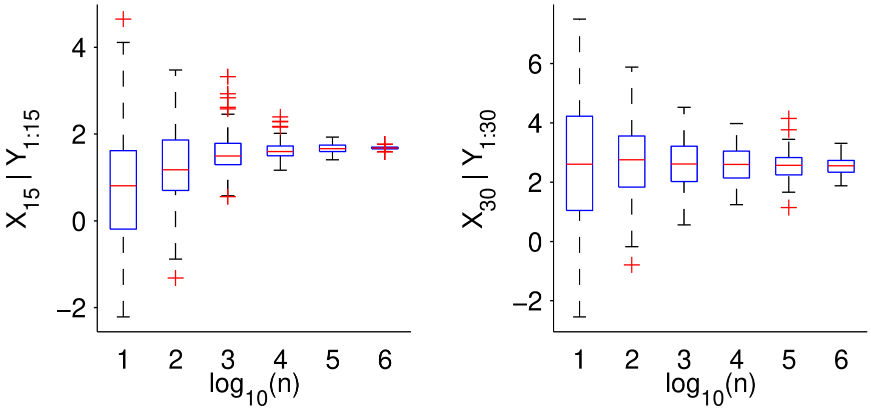
Remark 8.3. The generic methods such as importance sampling and MCMC (Random-walk Metropolis, Metrpolis-within-Gibbs, Hamiltonian Monte Carlo. . .) are, in theory, directly applicable in the SSM context. However, when T is large, the space $\mathbb{S}^T$ is high-dimensional, and there are substantial correlations in the model, which often lead to poor performance. . .

Remark 8.4 (*). Exact SSM inference (i.e. when the conditional distribution is available in a closed form) is possible only in some specific cases, most notably [e.g. 4]:

- When $\mathbb{S}$ is finite, exeact inference is possible through the *forward-backward* algorithm.
- If $\mathbb{S} = \mathbb{R}^d$ and and the conditional distributions m_t and g_t are linear Gaussian, that is, $g_t(x_t, \,\cdot\,)$ is a Gaussian density with mean $L_t x_t$ and some covariance matrix R_t, and similarly for m_t, then, the smoothing density (and consequently all the marginals) are Gaussian. Then, the mean & covariance parameters can be computed by simple matrix formulae (the Kalman filter and smoother).

In most other cases, inference need to be based on an approximation, such as SMC.

[15] Stationary iff $|\rho| < 1$ and $\sigma_1^2 = \frac{\sigma_x^2}{1-\rho^2}$.



(a) (b)

Figure 21: Box plot of estimates from Example 8.6 with up to one million samples, and 100 repetitions. (a) $T = 15$ (true value: 1.685) (b) $T = 30$ (true value: 2.508).

8.2 First attempt: Sequential importance sampling

Let us see what happens if we apply self-normalised importance sampling in the context of SSMs.

Generic self-normalised importance sampling is straightforward to apply here, because assuming $q(x_{1:T})$ is a proposal density on $\mathbb{S}^T$, with support covering that of $p(x_{1:T})$, we could just draw $X_{1:T}^{(k)} \overset{\text{i.i.d.}}{\sim} q$ and approximate

$$\mathbb{E}_p[f(X_{1:T})] \approx \frac{\sum_{k=1}^n w_u(X_{1:T}^{(k)}) f(X_{1:T}^{(k)})}{\sum_{j=1}^n w_u(X_{1:T}^{(j)})}, \qquad \text{where} \qquad w_u(x_{1:T}) := \frac{p_u(x_{1:T})}{q(x_{1:T})}.$$

Remark 8.5. Note that also the proposal q may depend on the observations $y_{1:T}$, in an arbitrary manner. Recall also that the notation differs here from the notation in Section 4.3: we write the sample index in superscript.

Example 8.6 (Noisy AR(1) with prior as q). Consider Example 8.2 and let q be the prior of $X_{1:T}$, that is,

$$q(x_{1:T}) = m_1(x_1) \prod_{t=2}^{T} m_t(x_{t-1}, x_t).$$

This means that we simulate $X_{1:T}$ to be the trajectories of T steps of a random walk with independent Gaussian $N(0, 1)$ increments.

Figures 21 and 22 show simulation results of Example 8.6.

The problem with Example 8.6 is that, even if the weights are bounded, the discrepancy of p and q increases very rapidly as T increases. In intuitive terms, most samples from q fall into low density area of p, and consequently the variance of the weights is large.

Let us have another attempt with more carefully chosen q:

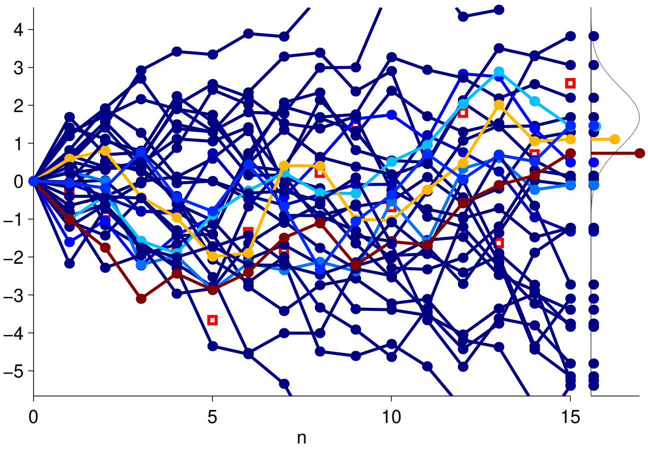
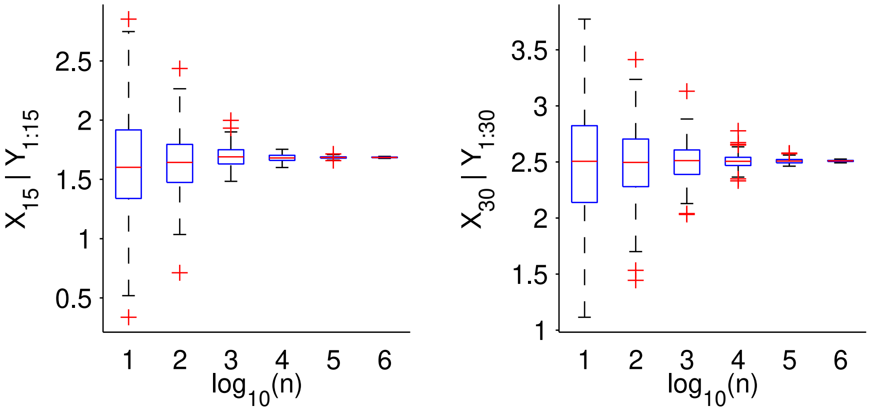


Figure 22: Some samples corresponding Example 8.6. Note that the weight distribution is very unequal. The true posterior density is shown on the right.



(a)

(b)

Figure 23: Box plot of estimates from Example 8.7; compare with 21.

Example 8.7 (Noisy AR(1) with a 'one-step optimal' q)*.* Consider Example 8.6, but choose now

$$q(x_{1:T}) = q_1(x_1)\prod_{t=2}^{T} q_t(x_t \mid x_{t-1}), \qquad q_t(x_t \mid x_{t-1}) = N\left(x_t; \frac{x_{t-1}+y_t}{2}, \frac{1}{2}\right) \qquad \text{(with } x_0 \equiv 0\text{).}$$

In fact, this choice of q_t corresponds to the conditional distribution of X_t given $X_{t-1} = x_{t-1}$ and $Y_t = y_t$. The conditional distribution is, in a certain sense, the best choice we can have (if we restrict on q_t that can only depend on $y_{1:t}\ldots$). It is direct to check that the unnormalised weights $w_u(z_{1:T})$ resulting from this choice are also bounded (exercise).

Figures 23 and 24 show simulation results corresponding Example 8.7.

Using a better proposal distribution in Example 8.7 improved significantly. It made reliable inference possible for up to $T = 30$ with around one million samples. This is achieved by better approximation of p by q, which shows in

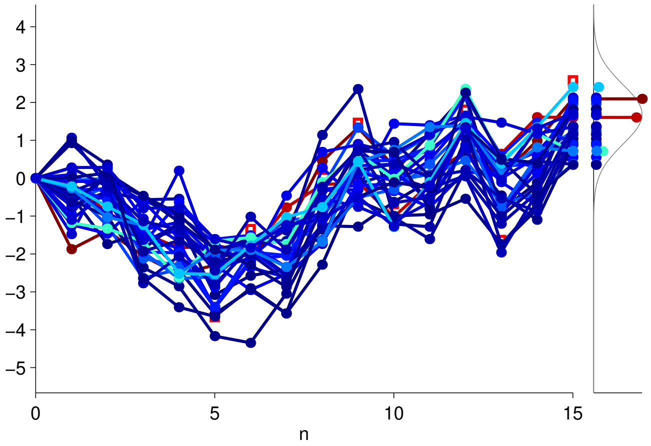
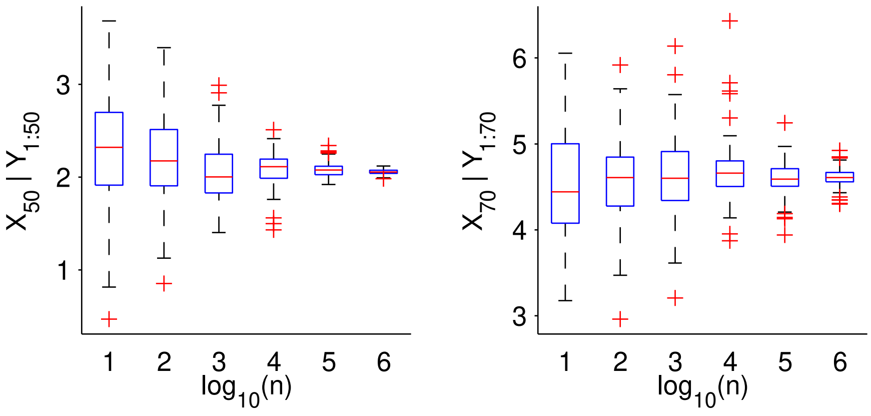


Figure 24: Some samples corresponding Example 8.7; compare with Figure 22.



(a)

(b)

Figure 25: Box plot of estimates from Example 8.7 with $T = 50$ (true: 2.058) and $T = 70$ (true: 4.606).

Figure 24 by concentration of the samples around the measured values.

However, if we increase T a bit more, we see that even the very good proposal distribution in 8.7 is insufficient for efficient inference; see Figure 25. In fact, the variance typically increases exponentially in T (cf. [4, Example 7.3.1]).

The particle filter algorithm, which we discuss next, is a simple algorithmic modification of the SIS, which resolves the 'mismatch' by further randomisation…

8.3 Generic form of sequential importance sampling

Suppose now that $M_t(x_t \mid x_{1:t-1})$ for $t = 2, \ldots, T$ determines a distribution on $\mathbb{S}$ for x_t for any $x_{1:t-1} \in \mathbb{S}^{t-1}$, and that $G_t(x_{1:t}) \geq 0$ are some 'potential' functions, for which:

$$\boxed{M_1(x_1)G_1(x_1)\prod_{t=2}^{T} M_t(x_t \mid x_{1:t-1})G_t(x_{1:t}) \equiv p_u(x_{1:T}).} \tag{22}$$

Remark 8.8. Note that in the SSM context, (22) is equivalent with $q(x_{1:T}) = M_1(x_1)\prod_{t=2}^T M_t(x_t \mid x_{1:t-1})$ satisfying the SNIS support condition (10) and G_t forming a *factorisation* of the unnormalised importance weight:

$$\prod_{t=1}^T G_t(x_{1:t}) = w_u(x_{1:T}) = \frac{m_1(x_1)g_1(x_1,y_1)\prod_{t=2}^T m_t(x_{t-1},x_t)g_t(x_t,y_t)}{M_1(x_1)\prod_{t=2}^T M_t(x_t \mid x_{1:t-1})}, \qquad \text{when } q(x_{1:T}) > 0.$$

We may choose $G_t(x_{1:t}) = \frac{m_t(x_{t-1},x_t)g_t(x_t,y_t)}{M_t(x_t|x_{1:t-1})}$, which satisfies (22), but other choices are possible.

Remark 8.9 (*). The model with ingredients of the form $M_{1:T}$ and $G_{1:T}$ is known as the *Feynman-Kac* model [7].

Algorithm 8.10 (Sequential importance sampling)**.** In each line of the algorithm, $i = 1, \ldots, n$:

(i) Sample $X_1^{(i)} \sim M_1(\,\cdot\,)$ and set $\mathbf{X}_1^{(i)} = X_1^{(i)}$.

(ii) Calculate $\omega_1^{(i)} := G_1(\mathbf{X}_1^{(i)})$.

For $t = 2, \ldots, T$, do:

(iii) Sample $X_t^{(i)} \sim M_t(\,\cdot \mid \mathbf{X}_{t-1}^{(i)})$ and set $\mathbf{X}_t^{(i)} = (\mathbf{X}_{t-1}^{(i)}, X_t^{(i)})$.

(iv) Calculate $\omega_t^{(i)} := G_t(\mathbf{X}_t^{(i)})$.

Report unnormalised sample $(V^{(1:n)}, \mathbf{X}^{(1:n)})$ where $V^{(j)} := \prod_{t=1}^T \omega_t^{(j)}$ and $\mathbf{X}^{(j)} := \mathbf{X}_T^{(j)}$.

Proposition 8.11. *Let* $t \in \{1{:}T\}$ *such that* $\int M_1(x_1)G_1(x_1)\prod_{k=2}^t M_k(x_k \mid x_{1:k-1})G_k(x_{1:k})\mathrm{d}x_{1:t} < \infty$*. Consider Algorithm 8.10, and and let* $\pi_t(x_{1:t}) \propto M_1(x_1)G_1(x_1)\prod_{k=2}^t M_k(x_k \mid x_{1:k-1})G_k(x'_{1:k})$ *be a probability density. Then, denoting* $V_t^{(i)} := \prod_{k=1}^t \omega_k^{(i)}$*,*

$$\frac{\sum_{i=1}^n V_t^{(i)} f(\mathbf{X}_t^{(i)})}{\sum_{j=1}^n V_t^{(j)}} \xrightarrow{n\to\infty} \mathbb{E}_{\pi_t}[f(X_{1:t})] \qquad \textit{(in distribution),}$$

whenever the expectation is well-defined and finite.

Proof. This is self-normalised IS, because $\mathbf{X}_t \sim q_t(x_{1:t}) = M_1(x_1)\prod_{k=2}^t M_k(x_k \mid x_{1:k-1})$ and $V_t^{(i)} \propto \pi_t(\mathbf{X}_t)/q_t(\mathbf{X}_t)$. The result follows from Theorem 4.19. □

Corollary 8.12. *If assumption* (22) *holds, then the output of Algorithm 8.10 satisfies:*

$$\mathrm{SIS}^{(n)}_{M_{1:T},G_{1:T}}(f) := \frac{\sum_{k=1}^n V^{(k)} f(\mathbf{X}^{(k)})}{\sum_{j=1}^n V^{(j)}} \xrightarrow{n\to\infty} \mathbb{E}_p[f(X_{1:T})] \qquad \textit{(in distribution)}$$

Proof. Direct application of Proposition 8.11, because $p = \pi_T$ and $V^{(i)} = V_T^{(i)}$. □

Remark 8.13. When $\pi_1, \ldots, \pi_T = p$ are all well-defined, Proposition 8.11 indicates that Algorithm 8.10 may be regarded as approximating these distributions sequentially, by re-using the approximation for π_{t-1} when building the approximation for π_t.