Notes on Nonlinear Time Series

Xunzhao Yin

February 25, 2019

1 Mathematical Background

1.1 Linear Algebra

Property 1.1. *[Properties of the Determinant]*

1. $\det(A^T) = \det(A)$.
2. $\det(A^{-1}) = \det(A)^{-1}$.
3. *For square matrices A and B of equal size,*

$$\det(AB) = \det(A) \cdot \det(B).$$

4. *For $n \times n$ matrix A,* $\det(cA) = c^n A$.
5. $\det(A + BC) = \det(A) \cdot \det(I + CA^{-1}B)$.

1.2 Probability

Definition 1.2 (Log-Normal Distribution)**.** Let Z be a normal variable with mean μ and variance σ^2, and $X = e^Z$. Then X is subject to a log-normal distribution with mean

$$\mathrm{E}[X] = e^{\mu + \frac{\sigma^2}{2}}.$$

Corollary 1.3. *Let Z be a normal variable with mean μ and variance σ^2, and $X = e^{\frac{Z}{2}}$. Then its mean*

$$\mathrm{E}[X] = \sqrt{e^{\mu + \frac{\sigma^2}{4}}}.$$

Proof. Since $X = e^{\frac{Z}{2}}$, we know $\log X = \frac{Z}{2}$. So

$$\log X \sim \mathrm{N}\left(\frac{\mu}{2}, \frac{\sigma^2}{4}\right).$$

Then applying Definition 1.2, we know X is log-normally distributed with mean

$$\mathrm{E}[X] = \exp\left\{\frac{\mu}{2} + \frac{\sigma^2}{8}\right\}.$$

□

Corollary 1.4. *Let Z be a normal variable with mean μ and variance σ^2, and $X = e^Z$. Then all moments of X exist and*

$$\mathrm{E}[X^n] = e^{n\mu + \frac{n^2\sigma^2}{2}}.$$

Property 1.5. *Let Z be a normal variable with mean μ and standard deviation σ, and $X = e^{iZ}$, then*

$$\mathrm{E}[X] = e^{i\mu - \frac{\sigma^2}{2}}.$$

Property 1.6 (Transformation Theorem)**.** *Let X be a n-dimensional continuous random variable with density $f_X(x)$ defined on $S \subset \mathbb{R}^n$. Let $g = (g_1, \ldots, g_n)$ be a bijection from S to some set $T \subset \mathbb{R}^n$. Consider the n-dimensional random variable*

$$Y = g(X).$$

Assume that g is 1st-order continuously differentiable. Then the density of Y is

$$f_Y(y) = f_X(h_1(y), h_2(y), \ldots, h_n(y)) \cdot |J|^{-1}, \quad y \in T,$$

where $h = (h_1, \ldots, h_n)$ is the inverse function of g, and J is the Jacobian (the determinant of the Jacobian matrix)

$$J = \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix}.$$

1.3 Stochastic Process

Definition 1.7 (Strict/Strong Stationarity)**.** Let $\{X_t\}$ be a stochastic process, and let $F_X(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_k+\tau})$ be the cumulative distribution function of the joint distribution of $\{X_t\}$ at times $t_1+\tau, t_2+\tau, \ldots, t_k+\tau$. Then, $\{X_t\}$ is said to be strictly or strongly stationary if, for all k, for all τ, and for all $t_1, \ldots, t_k$,

$$F_X(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, x_{t_2}, \ldots, x_{t_k}).$$

Definition 1.8 (Weak-Sense/Wide-Sense Stationarity)**.** $\{X_t\}$ is said to be weakly or widely stationary if the 1st moment and autocovariance do not vary with respect to time. Weak-sense stationarity is also called covariance stationarity.

Definition 1.9 (Weak-sense/Wide-sense Ergodicity Process)**.** Let $X(t)$ be a wide-sense stationary process with constant mean

$$\mu_X = \mathrm{E}[X(t)],$$

and autocovariance

$$\gamma_X(\tau) = \mathrm{E}\left[(X(t) - \mu_X)(X(t+\tau) - \mu_X)\right].$$

$X(t)$ of said to be *mean-ergodic* or *mean-square ergodic in the first moment* if the time average

$$\hat{\mu} = \frac{1}{T}\int_0^T X(t)\mathrm{d}t$$

converges in mean square to the ensemble average μ_X as $T \to \infty$, i.e.

$$\hat{\mu} \xrightarrow{L^r} \mu.$$

Likewise, $X(t)$ is said to be *autocovariance-ergodic* or *mean-square ergodic in the second moment* if the time average

$$\hat{\gamma}_X(\tau) = \frac{1}{T}\int_0^T \left[X(t) - \mu_X\right]\left[X(t+\tau) - \mu_X\right]\mathrm{d}t$$

converges in mean square to the ensemble autocovariance $\gamma_X(\tau)$ as $T \to \infty$.

If $X(t)$ is ergodic in mean and autocovariance, then it is called *ergodic in the wide sense*.

1.4 Statistics

SUFFICIENCY PRINCIPLE: (Casella & Berger, 2002) If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the reference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{Y} = \mathbf{y}$ is observed.

Definition 1.10 (Sufficient Statistics)**.** (Casella & Berger, 2002) A statistic $T(\mathbf{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

Theorem 1.11. *(Casella & Berger, 2002) If $p(\mathbf{x}|\theta)$ is the pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ is the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{X})|\theta)$ is constant as a function of θ.*

Theorem 1.12 (Factorization Theorem)**.** *(Casella & Berger, 2002) Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}).$$

Theorem 1.13 (Rao-Blackwell Theorem)**.** *Let $\hat{\theta}(\mathbf{X})$ be an estimator of θ with $\mathrm{E}[\hat{\theta}(\mathbf{X})] < \infty$ for all θ. Suppose that $T(\mathbf{X})$ is sufficient for θ, and let $\theta^* = \mathrm{E}[\hat{\theta}(\mathbf{X})|T(\mathbf{X})]$. Then, for all θ,*

$$\mathrm{E}[\theta^*(\mathbf{X}) - \theta]^2 \leq \mathrm{E}[\hat{\theta}(\mathbf{X}) - \theta]^2.$$

The inequality is strict unless $\hat{\theta}$ is a function of T.

Definition 1.14 (Rao-Blackwellization)**.** The estimator $\theta^*(\mathbf{X})$ in Theorem 1.13 is called a *Rao-Blackwell estimator*. And the process of transforming an estimator into a Rao-Blackwell estimator is called *Rao-Blackwellization*.

1.5 Ring

Definition 1.15 (Ring)**.** A family $\mathcal{R}$ of sets is called a *ring* if $A \in \mathcal{R}$ and $B \in \mathcal{R}$ implies

$$A \cup B \in \mathcal{R}, \quad A - B \in \mathcal{R}.$$

Corollary 1.16. *If $\mathcal{R}$ is a ring, $A \in \mathcal{R}$ and $B \in \mathcal{R}$, then $A \cap B \in \mathcal{R}$.*

Definition 1.17 (σ-ring)**.** A ring $\mathcal{R}$ is a σ-ring if

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{R}$$

whenever $A_n \in \mathcal{R}$, $n = 1, 2, 3, \ldots$.

Corollary 1.18. *If a ring $\mathcal{R}$ is a σ-ring, and $A_n \in \mathcal{R}$, $n = 1, 2, 3, \ldots$, then*

$$\bigcap_{i=1}^{n} A_n \in \mathcal{R}.$$

Definition 1.19 (Set Function)**.** We say that ϕ is a *set function* defined on $\mathcal{R}$ if ϕ assigns to every $A \in \mathcal{R}$ a number $\phi(A)$ of the extended real number system.

Definition 1.20 (Additivity and Countable Additivity)**.** ϕ is *additive* if $A \cap B = \emptyset$ implies

$$\phi(A \cup B) = \phi(A) + \phi(B).$$

ϕ is *countably additive* if $A_i \cap A_j = 0 (i \neq j)$ implies

$$\phi\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \phi(A_n).$$

1.6 Lebesgue Measure

Definition 1.21 (Interval)**.** Let $\mathbb{R}^p$ be a p-dimensional Euclidean space. An *interval* in $\mathbb{R}^p$ is a set of points $\mathbf{x} = \{(x_1, \ldots, x_p)\}$ such that

$$a_i \leq x_i \leq b_i, \quad i = 1, \ldots, p.$$

The possibility that $a_i = b_i$ for any i is not ruled out; the empty set is included among the intervals.

Note that $\mathbb{R}^p$ is a σ-ring.

Definition 1.22 (Elementary Set)**.** $A \in \mathbb{R}^p$ is said to be an *elementary set* if A is the union of a finite number of intervals.

We let $\mathcal{E}$ denote the family of all elementary subsets of $\mathbb{R}^p$. Then, $\mathcal{E}$ is a ring, but not a σ-ring.

Corollary 1.23. *If $A \in \mathcal{E}$, then A is the union of a finite number of* disjoint *intervals.*

If I is an interval, we define

$$m(I) = \sum_{i=1}^{p} (b_i - a_i). \tag{1}$$

If $A = I_1 \cup \cdots \cup I_n$, and if these intervals are disjoint, we set

$$m(A) = m(I_1) + \cdots + m(I_n). \tag{2}$$

Definition 1.24 (Regularity)**.** A nonnegative additive set function ϕ defined on $\mathcal{E}$ is said to be *regular* if the following is true: to every $A \in \mathcal{E}$ and to every $\varepsilon > 0$ there exist sets $F \in \mathcal{E}$, $G \in \mathcal{E}$ such that F is closed, G is open, $F \subset A \subset G$, and

$$\phi(G) - \varepsilon \leq \phi(A) \leq \phi(F) + \varepsilon.$$

Note that the set function m defined in (1) and (2) is regular.

Definition 1.25 (Outer Measure)**.** Let the set function μ be additive, regular, nonnegative, and finite on $\mathcal{E}$. Consider countable coverings of any set $E \subset \mathbb{R}^p$ by open elementary sets A_n:

$$E \subset \bigcup_{n=1}^{\infty} A_n.$$

Define

$$\mu^*(E) = \inf \sum_{n=1}^{\infty} \mu(A_n),$$

the inf being taken over all open coverings of E by open elementary sets. $\mu^*(E)$ is called the *outer measure* of E corresponding to μ.

Note that μ^* is the extension of μ from $\mathcal{E}$ to the family of all subsets of $\mathbb{R}^p$.

Theorem 1.26.

a *For every* $A \in \mathcal{E}$, $\mu^*(A) = \mu(A)$.

b *If* $E = \bigcup_{n=1}^{\infty} E_n$, *then*

$$\mu^*(E) \leq \sum_{n=1}^{\infty} \mu^*(E_n).$$

Definition 1.27 (Symmetric Difference)**.** For any $A \subset \mathbb{R}^p$ and $B \subset \mathbb{R}^p$, we define the *symmetric difference* as

$$S(A, B) = (A - B) \cup (B - A).$$

Definition 1.28 (Finite Measurability)**.** If there is a sequence $\{A_n\}$ of elementary sets such that $A_n \to A$, we say that A is *finitely μ-measurable* and write $A \in \mathcal{M}_F(\mu)$.

Definition 1.29 (Measurability)**.** If A is the union of a countable collection of finitely μ-measurable sets, we say that A is *μ-measurable* and write $A \in \mathcal{M}(\mu)$.

Theorem 1.30. *$\mathcal{M}(\mu)$ is a σ-ring, and μ^* is countably additive on $\mathcal{M}(\mu)$.*

Definition 1.31 (Measure)**.** The extended countably additive set function μ^* defined on the σ-ring $\mathcal{M}(\mu)$ is called a *measure*.

Definition 1.32 (Lebesgue Measure)**.** When $\mu = m$ which is defined in (1) and (2), μ^* is called a *Lebesgue measure* on $\mathbb{R}^p$.

1.7 Measure Spaces and Probability

What is a measure? What is a measurable function? Note that the μ in the following definition is just a measure, not a Lebesgue measure. A Lebesgue measure is defined on a Euclidean space.

Definition 1.33 (Lebesgue Integral)**.** Let $(\mathsf{X}, \mathcal{X}, \mu)$ be a measure space, where $\mu \in \mathbb{M}_+(\mathcal{X})$.

1. Let $S \in \mathcal{X}$, and $1_S \in \mathbb{F}(\mathsf{X}, \mathcal{X})$ be the indicator function of S. The Lebesgue integral of 1_S

$$\int 1_S \, \mathrm{d}\mu = \mu(S).$$

2. Let $K \in \mathbb{N}$. For $k \in \{1, 2, \ldots, K\}$, let $a_k \in \mathbb{R}$, and $S_k \in \mathcal{X}$ such that $\mu(S_k) < \infty$ when $a_k \neq 0$. Then $\sum_k a_k 1_k \in \mathbb{F}(\mathsf{X}, \mathcal{X})$ is called a measurable simple function. The Lebesgue integral of $\sum_{k=1}^{K} a_k 1_k$

$$\int \left(\sum_k a_k 1_k \right) \mathrm{d}\mu = \sum_{k=1}^{K} \mu(S_k).$$

The convention $0 \times \infty = 0$ must be used, and the result could be infinite.

3. Let $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$ and s be a measurable simple function. Its Lebesgue integral

$$\int f \mathrm{d}\mu = \sup \left\{ \int s \mathrm{d}\mu : 0 \leq s \leq f \right\}.$$

For some function, this integral is infinite.

4. Let $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$. We write

$$f = f^+ - f^-$$

where

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$f^-(x) = \begin{cases} -f(x) & \text{if } f(x) < 0 \\ 0 & \text{otherwise} \end{cases}$$

Then $f^+, f^- \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$, and

$$|f| = f^+ + f^-.$$

If

$$\min \left(\int f^+ \mathrm{d}\mu, \int f^- \mathrm{d}\mu \right) < \infty,$$

or, in other words, at least one of them is finite, we say that the Lebesgue integral of f *exists*, and we define

$$\int f \mathrm{d}\mu = \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu.$$

If

$$\int |f| \mathrm{d}\mu < \infty,$$

we say that f is *Lebesgue integrable* with respect to μ. We write $f \in L(\mathsf{X}, \mathcal{X}, \mu)$.

Property 1.34 (Lebesgue Integrability)**.** *(Rudin (1976), P315)*

1. *If $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$, and $\mu \in \mathbb{M}_+(\mathcal{X})$ satisfying $\mu(\mathsf{X}) < \infty$, then $f \in L(\mathsf{X}, \mathcal{X}, \mu)$.*

2. *If $a \leq f(x) \leq b$ for $x \in \mathsf{X}$, and $\mu(\mathsf{X}) < \infty$, then*

$$a\mu(\mathsf{X}) \leq \int f \mathrm{d}\mu \leq b\mu(\mathsf{X}).$$

Definition 1.35 (L^p Spaces)**.** For $p \in [1, \infty)$, $L^p(\mathsf{X}, \mathcal{X}, \mu)$ is the set of $\mathbb{R}$-valued functions with finite p-norms, i.e.

$$\|f\|_p = \left(\int |f|^p \mathrm{d}\mu \right)^{\frac{1}{p}} < \infty.$$

Definition 1.36 (Uniform/Supremum Norm)**.** Let $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$. The *uniform norm* or *supremum norm* assigns to f the non-negative number

$$\|f\|_\infty = \sup\{f(x) : x \in \mathsf{X}\}.$$

Definition 1.37 (Sup Norm Distance)**.** Let $f, g \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$. The *sup norm distance* between f and g is defined as $d_\infty(f, g) = \|f - g\|_\infty = \sup_{x \in \mathsf{X}} |f(x) - g(x)|$.

Definition 1.38 (Oscillation Seminorm)**.** Let $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$. Its *oscillation seminorm* is defined as

$$\mathrm{ocs}(f) = \sup_{(x,y) \in \mathsf{X}^2} |f(x) - f(y)| = 2 \inf_{c \in \mathbb{R}} \|f - c\|_\infty.$$

Definition 1.39 (Absolute Continuity (Corbae, Stinchcombe, and Zeman (2009), P349))**.** Let $(\mathsf{X}, \mathcal{X})$ be a measurable space, μ and ν be two measures on $(\mathsf{X}, \mathcal{X})$. μ is absolutely continuous with respect to ν or ν dominates μ, written $\mu \ll \nu$, if $\mu(A) = 0$ whenever $\nu(A) = 0$ for $A \in \mathcal{X}$.

Theorem 1.40 (Radon-Nikodym (Corbae et al. (2009), P349))**.** *Let* $(\mathsf{X}, \mathcal{X}, P)$ *be a probability space,* $\mu \in \mathbb{M}_+(\mathcal{X})$*. If* $\mu \ll P$*, then* μ *has a density with respect to P, that is, there is a non-negative P-integrable* $w : \mathsf{X} \to \mathbb{R}_+$ *such that* $\mu(A) = \int_A w \mathrm{d}P$ *for all* $A \in \mathcal{X}$*, and w is unique up to a set of P-measure 0.*

Definition 1.41 (Filtered Probability Space)**.** A probability space $(\mathsf{X}, \mathcal{X}, P)$ is a filtered probability space if there exists an increasing sequence $\{\mathcal{X}_t, t \in \mathbb{N}\}$ of σ-algebra included in $\mathcal{X}$, $\{\mathcal{X}_s \subset \mathcal{X}_t\}$ for $s \leq t$. A filtered probability space is denoted by $(\mathsf{X}, \mathcal{X}, \{\mathcal{X}_t, t \in \mathbb{N}\}, P)$.

Definition 1.42 (Adapted Stochastic Process)**.** A stochastic process $\{X_t, t \in \mathbb{N}\}$ defined on a filtered probability space $(\mathsf{X}, \mathcal{X}, \{\mathcal{X}_t, t \in \mathbb{N}\}, P)$ is called adapted if X_t is $\mathcal{X}_t$-measurable for any $t \in \mathbb{N}$. An adapted stochastic process is denoted by $\{(X_t, \mathcal{X}_t), t \in \mathbb{N}\}$.

Definition 1.43 (Almost Sure Equality)**.** Let X, Y be two real valued measurable function on $(\mathsf{X}, \mathcal{X}, \mu)$. We say that $X = Y$ *almost surely* if $P(\{|X - Y| = 0\}) = 1$.

Definition 1.44 (Convergence)**.** Let $(\mathsf{X}, \mathcal{X}, P)$ be a probability space, $\{X_n\}$ and X be real valued random variables on the probability space. We say the sequence $\{X_n\}$ of random variables converges

1. *in distribution* towards the random variable X if, for all x at which F is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x),$$

   where F_n and F are the cumulative distribution functions of X_n and X respectively.

2. *in probability* towards the random variable X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0;$$

3. *almost surely* or *almost everywhere* or *in probability 1* or *strongly* towards X if

$$P(\lim_{n\to\infty} X_n = X) = 1,$$

   or equivalently, if

$$P\big(x \in \mathsf{X} : \lim_{n\to\infty} X_n(x) = X(x)\big) = 1$$

4. *in the r-th mean* or *in the L^r-norm* towards X if, given a real number $r \geq 1$, the r-th absolute moments $\mathrm{E}[|X_n|^r]$ and $\mathrm{E}[r^n]$ of $\{X_n\}$ and X exist, and

$$\lim_{n\to\infty} \mathrm{E}[|X_n - X|^r] = 0.$$

This type of convergence is often denoted by

$$X_n \xrightarrow{L^r} X.$$

When $r = 1$, we say that $\{X_n\}$ converges *in mean* to X. And when $r = 2$, we say that $\{X_n\}$ converges *in mean square* to X.

5. *pointwise* if for all $x \in \mathsf{X}$,

$$\lim_{n\to\infty} X_n(x) = X(x).$$

Property 1.45 (Properties of Convergence in the r-th Mean)**.**

1. *Convergence in the r-th mean, for $r \geq 1$, implies convergence in probability.*
2. *If $r > s \geq 1$, convergence in r-th mean implies convergence in the s-th mean.*
3. *If $X_n \xrightarrow{L^r} X$, then*

$$\lim_{n\to\infty} \mathrm{E}[|X_n|^r] = \mathrm{E}[|X|^r].$$

Definition 1.46 (Conditional Expectation)**.** (P84, Walkden (2017)) Let $(\mathsf{X}, \mathcal{X}, P)$ be a probability space, and $\mathcal{A} \subset \mathcal{X}$ be a sub-σ-algebra. Then $P \in \mathbb{M}_+(\mathcal{A})$.

1. For any $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$, define a measure $\nu \in \mathbb{M}_+(\mathcal{A})$ such that, for any $A \in \mathcal{A}$,

$$\nu(A) = \int_A f \mathrm{d}P.$$

Since $\nu \ll P$ on $(\mathsf{X}, \mathcal{A})$, by the Radon-Nikodym Theorem 1.40, there is a unique function $\mathrm{E}[f|\mathcal{A}] \in \mathbb{F}_+(\mathsf{X}, \mathcal{A})$ such that

$$\nu(A) = \int_A \mathrm{E}[f|\mathcal{A}] \mathrm{d}P.$$

2. For any $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$, we split f into positive and negative parts $f = f_+ - f_-$ and define

$$\mathrm{E}[f|\mathcal{A}] = \mathrm{E}[f_+|\mathcal{A}] - \mathrm{E}[f_-|\mathcal{A}].$$

Definition 1.47 (Measure-Preserving Transformation)**.** Let $(\mathsf{X}, \mathcal{X}, \mu)$ be a measure space, and $T : \mathsf{X} \to \mathsf{X}$ be a *measurable transformation* on $(\mathsf{X}, \mathcal{X}, \mu)$. Then, T is called measure preserving if for any $A \in \mathcal{X}$, $\mu(T^{-1}A) = \mu(A)$.

Definition 1.48 (Ergodic Transformation)**.** Let $(\mathsf{X}, \mathcal{X}, P)$ be a probability space, and let $T : \mathsf{X} \to \mathsf{X}$ be a measure-preserving transformation on $(\mathsf{X}, \mathcal{X}, P)$. Then T is *ergodic* if for all $A \in \mathcal{X}$ with $T^{-1}A = A$, either $P(A) = 0$ or $P(A) = 1$.

Theorem 1.49 (Birkhoff's Ergodic Theorem)**.** *(P89, Walkden (2017)) Let $(\mathsf{X}, \mathcal{X}, P)$ be a probability space, and $T : \mathsf{X} \to \mathsf{X}$ be a measure-preserving transformation on $(\mathsf{X}, \mathcal{X}, P)$, and let $\mathcal{I} = \{A \in \mathcal{X} : T^{-1}A = A a.e.\}$ denote the σ-algebra of T-invariant sets* [1]*. Then for any $f \in L^1(\mathsf{X}, \mathcal{X}, P)$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} f\left(T^{i-1}x\right) \xrightarrow{P\text{-}a.e.} \mathrm{E}(f|\mathcal{I})(x).$$

Corollary 1.50 (Birkhoff's Ergodic Theorem for an Ergodic Transformation)**.**
Let $(\mathsf{X}, \mathcal{X}, P)$ be a probability space, and let $T : \mathsf{X} \to \mathsf{X}$ be an ergodic measure-preserving transformation on $(\mathsf{X}, \mathcal{X}, P)$. Then for any $f \in L^1(\mathsf{X}, \mathcal{X}, P)$,

$$\lim_{n\to\infty} \sum_{i=1}^{n} f\left(T^{i-1}x\right) \xrightarrow{P\text{-}a.e.} \int f \mathrm{d}P.$$

1.8 Kernels

Definition 1.51 (Kernel)**.** Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A kernel with source $(\mathsf{X}, \mathcal{X})$ and target $(\mathsf{Y}, \mathcal{Y})$ is a map $N : \mathsf{X} \times \mathcal{Y} \to [0, \infty]$ satisfying the following conditions:

1. for every $B \in \mathcal{Y}$, the map $N(\cdot, B) : x \to N(x, B)$ is a measurable function from $(\mathsf{X}, \mathcal{X})$ to $[0, \infty]$,

[1] It is straightforward to check that $\mathcal{I}$ is a σ-algebra.

2. for every $x \in \mathsf{X}$, the map $N(x, \cdot) : B \to N(x, B)$ is a measure on $\mathcal{Y}$.

A kernel N is said to be finite if $N(x, \mathsf{Y}) < \infty$ for all $x \in \mathsf{X}$.

A kernel N is said to be bounded if $\sup_{x \in \mathsf{X}} N(x, \mathsf{Y}) < \infty$.

A kernel N is said to be Markovian if $N(x, \mathsf{Y}) = 1$, for all $x \in \mathsf{X}$.

Or equivalently, a Markov kernel could be define as follows.

Definition 1.52 (Markov Kernel)**.** Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A Markov kernel with source $(\mathsf{X}, \mathcal{X})$ and target $(\mathsf{Y}, \mathcal{Y})$ is a map $N : \mathsf{X} \times \mathcal{Y} \to [0, 1]$ with the following properties:

1. For every $B \in \mathcal{Y}$, the map $N(\cdot, B) : x \to \kappa(x, B)$ with $x \in \mathsf{X}$ is $\mathcal{X}$-measurable.

2. For every $x \in \mathsf{X}$, the map $N(x, \cdot) : B \to \kappa(x, B)$ with $B \in \mathcal{Y}$ is a probability measure on $(\mathsf{Y}, \mathcal{Y})$.

Definition 1.53 (Tensor Product)**.** Let $(\mathsf{X}, \mathcal{X})$ be a measurable space, $\nu \in \mathbb{M}_+(\mathcal{X})$ be a measure and $Q : \mathsf{X} \times \mathcal{X} \to [0 : \infty]$ be a kernel. The *tensor product* of the measure ν and the kernel Q is a measure defined on $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ such that, for any $C \in \mathcal{X}^{\otimes 2}$,

$$\nu \otimes Q(C) = \iint \nu(\mathrm{d}x) Q(x, \mathrm{d}x') 1_C(x, x').$$

1.9 Martingales

Definition 1.54 (Martingales)**.** Let $(\mathsf{X}, \mathcal{X}, \{\mathcal{F}_n, n \in \mathbb{N}\}, P)$ be a filtered probability space and $\{(X_n, \mathcal{F}_n), n \in \mathbb{N}\}$ be a real integrable adapted process. We say that $\{(X_n, \mathcal{F}_n), n \in \mathbb{N}\}$ is

1. a martingale if, for all $0 \leq m < n$, $\mathrm{E}(X_n | \mathcal{F}_m) = X_m$, P-a.s.

2. a submartingale if, for all $0 \leq m < n$, $\mathrm{E}(X_n | \mathcal{F}_m) \geq X_m$, P-a.s.

3. a supermartingale if, for all $0 \leq m < n$, $\mathrm{E}(X_n | \mathcal{F}_m) \leq X_m$, P-a.s.

Theorem 1.55 (Martingale Convergence Theorem)**.**

1. *If* $\{(X_n, \mathcal{F}_n), n \in \mathbb{N}\}$ *is a submartingale that satisfies* $\sup \mathrm{E}[X_n^+] < \infty$*, then* $X_n \overset{P\text{-a.e.}}{\longrightarrow} X$ *and* $\mathrm{E}[X] < \infty$.

2. *If* $\{(X_n, \mathcal{F}_n), n \in \mathbb{N}\}$ *is a positive supermartingale, then* $X_n \xrightarrow{P\text{-a.e.}} X$ *as* $n \to \infty$ *and* $\mathrm{E}[X] \leq \mathrm{E}[X_0]$.

Lemma 1.56. *Let* $\{(Z_n, \mathcal{F}_n), n \in \mathbb{N}\}$ *be an adapted sequence of nonnegative random variables. For all* $\varepsilon > 0$*, if for* $n \in \mathbb{N}$ *we denote*

$$\xi_n \equiv \mathrm{E}\left[\sum_{i=1}^{n} Z_i \cdot 1\Big(\sum_{j=1}^{i} \mathrm{E}[Z_j|\mathcal{F}_{j-1}] \leq \varepsilon\Big)\right],$$

then

$$\xi_n \leq \varepsilon.$$

Lemma 1.57. *Let* $\{(Z_n, \mathcal{F}_n), n \in \mathbb{N}\}$ *be an adapted sequence of nonnegative random variables. Then, for all* $\varepsilon > 0$*,* $\alpha > 0$ *and* $n \in \mathbb{N}$*,*

$$P(\max_{1 \leq i \leq n} Z_i > \varepsilon) \leq \alpha + P\Big(\sum_{i=1}^{n} P(Z_i > \varepsilon|\mathcal{F}_{i-1}) > \alpha\Big).$$

Lemma 1.58. *Let* $\mathcal{G}$ *be a* σ*-field and* X *a random variable such that* $\mathrm{E}[X^2|\mathcal{G}] < \infty$*. Then, for any* $\varepsilon > 0$*,*

$$\begin{aligned} &4\mathrm{E}\left[|X|^2 \cdot 1(|X| > \epsilon)|\mathcal{G}\right] \\ &\geq \mathrm{E}\left[\left|X - \mathrm{E}[X|\mathcal{G}]\right|^2 \cdot 1\left(\left|X - \mathrm{E}[X|\mathcal{G}]\right| \geq 2\varepsilon\right)\Big|\mathcal{G}\right] \end{aligned}$$

Theorem 1.59 (Limit Theorem 1 for Triangular Arrays)**.** *(Douc, Moulines, and Stoffer (2014), P479)*

Let $(\mathsf{X}, \mathcal{X}, P)$ *be a probability space,* $\{M_N\}_{N \in \mathbb{N}^+}$ *be a sequence of positive integers satisfying* $M_N \to \infty$ *as* $N \to \infty$ *,* $\{(U_{N,i})_{i \in \{1,\ldots,M_N\}}\}_{N \in \mathbb{N}^+} \equiv \{U_j\}_{j \in \mathbb{N}^+}$ *be a triangular array of random variables on* $(\mathsf{X}, \mathcal{X})$*, and* $\{(\mathcal{F}_{N,i})_{i \in \{1,\ldots,M_N\}}\}_{N \in \mathbb{N}^+} \equiv \{\mathcal{F}_j\}_{j \in \mathbb{N}^+}$ *be a triangular array of sub-*σ*-fields of* $\mathcal{X}$*. For each* $j \in \mathbb{N}^+$*,* $\mathcal{F}_{j-1} \subset \mathcal{F}_j$*. For each* $N \in \mathbb{N}^+$ *and* $i = 1, \ldots, M_N$*,* $U_{N,i}$ *is* $\mathcal{F}_{N,i}$*-measurable.*

Assume that, for any $N \in \mathbb{N}^+$ *and* $i = 1, \ldots, M_N$*,* $\mathrm{E}\left[|U_{N,i}|\Big|\mathcal{F}_{N,i-1}\right] < \infty$*, and*

$$\sup_N P\Big(\sum_{i=1}^{M_N} \mathrm{E}\left[|U_{N,i}|\Big|\mathcal{F}_{N,i-1}\right] \geq \lambda\Big) \to 0 \quad \text{as } \lambda \to \infty, \tag{3}$$

$$\forall \varepsilon > 0, \ \sum_{i=1}^{M_N} \mathrm{E}\left[|U_{N,i}| \cdot 1(|U_{N,i}| > \varepsilon)\Big|\mathcal{F}_{N,i-1}\right] \xrightarrow{P} 0 \quad \text{as } N \to \infty. \tag{4}$$

Then,

$$\max_{i\in\{1,\ldots,M_N\}}\left|\sum_{k=1}^{i}U_{N,k}-\sum_{k=1}^{i}\mathrm{E}[U_{n,k}|\mathcal{F}_{N,k-1}]\right|\xrightarrow{P}0\quad as\ N\to 0.$$

Lemma 1.60. *Assume that for all N,* $\sum_{i=1}^{M_N}\mathrm{E}\left[U_{N,i}^2\middle|\mathcal{F}_{N-1}\right]=1$, *and* $\mathrm{E}[U_{N,i}|\mathcal{F}_{N,i-1}]=0$ *for* $i=1,\ldots,M_N$, *and for all* $\varepsilon>0$,

$$\sum_{i=1}^{M_N}\mathrm{E}\left[U_{N,i}^2\cdot 1(|U_{N,i}|\geq\varepsilon)\middle|\mathcal{F}_{N,0}\right]\xrightarrow{P}0.$$

Then, for any real u,

$$\mathrm{E}\left[\exp\left(i\cdot u\sum_{j=1}^{M_N}U_{N,j}\right)\middle|\mathcal{F}_{N,0}\right]-\exp\left(-\frac{u^2}{2}\right)\xrightarrow{P}0,$$

where i is the imaginary number.

Theorem 1.61 (Limit Theorem 2 for Triangular Arrays)**.** *Assume that for each N and* $i=1,\ldots,M_N$, $\mathrm{E}\left[U_{N,i}^2\middle|\mathcal{F}_{N,i-1}\right]<\infty$, *and*

$$\sum_{i=1}^{M_N}\left(\mathrm{E}[U_{N,i}^2|\mathcal{F}_{N,i-1}]-\mathrm{E}^2[U_{N,i}|\mathcal{F}_{N,i-1}]\right)\xrightarrow{P}\sigma^2,\quad for\ some\ \sigma^2>0, \tag{5}$$

$$\sum_{i=1}^{M_N}\mathrm{E}\left[U_{N,i}^2\cdot 1\left(|U_{N,i}|>\varepsilon\right)\middle|\mathcal{F}_{N,i-1}\right]\xrightarrow{P}0,\quad for\ any\ \varepsilon>0. \tag{6}$$

Then, for any real u,

$$\mathrm{E}\left[\exp\left(i\cdot u\sum_{i=1}^{M_N}\left(U_{N,i}-\mathrm{E}[U_{N,i}|\mathcal{F}_{N,i-1}]\right)\right)\middle|\mathcal{F}_{N,0}\right]\xrightarrow{P}\exp\left(-\frac{u^2}{2}\sigma^2\right).$$

Theorem 1.62. *Assume that* $\{X_k,k\in\mathbb{N}\}$ *is a strict-sense stationary, ergodic process such that* $\mathrm{E}[X_1^2]$ *is finite, and* $\mathrm{E}[X_k|\mathcal{F}_{k-1}^X]=0$, *where* $\{\mathcal{F}_k^X,k\in\mathbb{N}\}$ *is the natural filtration. Then,*

$$n^{-1/2}\sum_{k=1}^{n}X_k\xrightarrow{D}\mathrm{N}(0,\mathrm{E}[X_1^2]).$$

2 Notation

$\mathbb{N}$: the set of natural numbers including zero.

$\mathbb{N}_+$: the set of natural numbers excluding zero.

$\mathbb{F}(\mathsf{X}, \mathcal{X})$: the space of measurable functions from $(\mathsf{X}, \mathcal{X})$ to $[-\infty, \infty]$.

$\mathbb{F}_b(\mathsf{X}, \mathcal{X})$: the subset of $\mathbb{F}(\mathsf{X}, \mathcal{X})$ of bounded functions.

$\mathbb{F}_+(\mathsf{X}, \mathcal{X})$: the set of bounded measurable functions from $(\mathsf{X}, \mathcal{X})$ to $[0, \infty)$.

$\mathscr{L}(\mu)$: the set of Lebesgue integrable functions on $(\mathsf{X}, \mathcal{X}, \mu)$.

$\mathbb{M}(\mathcal{X})$: the set of finite signed measures on the measurable space $(\mathsf{X}, \mathcal{X})$.

$\mathbb{M}_+(\mathcal{X})$: the set of measures on the measurable space $(\mathsf{X}, \mathcal{X})$.

$\mathbb{M}_1(\mathcal{X})$: the set of probability measures on the measurable space $(\mathsf{X}, \mathcal{X})$.

$\mu(f)$: for any $\mu \in \mathbb{M}_+(\mathcal{X})$ and $f \in \mathbb{F}(\mathsf{X}, \mathcal{X})$, $\mu(f) \equiv \int f \mathrm{d}\mu$ is the measure of a function, or the Lebesgue integral of a function.

Nf: (Douc et al. (2014), P135) let N be a kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$, and $f \in \mathbb{F}_+(\mathcal{Y})$. A function $Nf : \mathsf{X} \to \mathbb{R}^+$ is defined as follows:

$$Nf : x \to \int_{\mathsf{Y}} N(x, \mathrm{d}y) f(y).$$

μN: a measure on $(\mathsf{X}, \mathcal{X})$ defined as

$$\mu N(A) = \iint \mu(\mathrm{d}x) N(x, \mathrm{d}x') 1(x' \in A), \quad \text{for any } A \in \mathcal{X}. \tag{7}$$

$[\mu]_C$: for $\mu \in \mathbb{R}$, $C \in \mathbb{R}$, $[\mu]_C = \mu \cdot 1(\mu \geq C)$.

3 Markov Chain Monte Carlo Methods

3.1 Metropolis-Hastings Algorithm

Assume that π, which is known up to a normalizing constant, is a probability density of interest on $\mathbb{R}^d$. Let $q(x, \cdot)$ be a *proposal* density that is easy to sample from. All these densities are considered with respect to the Lebesgue measure.

The algorithm proceeds as follows. An initial value x_0 is chosen. Given x_t, a candidate y_{t+1} is sampled from $q(x_t, \cdot)$, and is then accepted with probability $\alpha(X_t, Y_{t+1})$, given by

$$\alpha(x, y) = \begin{cases} \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{if } \pi(x)q(x, y) = 0. \end{cases} \tag{8}$$

If accepted, we set $x_{t+1} = y_{t+1}$. Otherwise, y_{t+1} is not accepted and we set $x_{t+1} = x_t$.

This procedure produces a Markov chain, $\{X_t, t \in \mathbb{N}_+\}$, with transitional kernel P given by

$$\begin{aligned} &P(x, A) \\ &= \int \alpha(x, y)q(x, y)1_A(y)\text{Leb}(\text{d}y) + 1_A(x) \int q(x, y)[1 - \alpha(x, y)]\text{Leb}(\text{d}y), \end{aligned} \tag{9}$$

where Leb(dy) represents a d-dimensional infinitesimal interval in $\mathbb{R}^d$.

Definition 3.1 (Reversible Markov Kernel)**.** The Markov kernel P is reversible with respect to a probability measure μ if for any function f and g on $\mathbb{R}^p$,

$$\iint \mu(\text{d}x)P(x, \text{d}y)f(x)g(y) = \iint \mu(\text{d}x)P(x, \text{d}y)g(x)f(y).$$

If P is reversible with respect to μ, take $g(\cdot) = 1$, and we get

$$\int \mu(\text{d}x)f(x) = \iint \mu(\text{d}x)P(x, \text{d}y)f(y).$$

Therefore, the following lemma holds:

Lemma 3.2. *If P is reversible with respect to μ, then μ is a stationary distribution for P.*

And because of (8),

Lemma 3.3. *For any* $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$*, we have*

$$\pi(x)\alpha(x,y)q(x,y) = \pi(x)q(x,y) \wedge \pi(y)q(y,x) = \pi(y)\alpha(y,x)q(y,x).$$

Proof. This is obvious from the definition equation (8). □

Theorem 3.4. *The transition kernel* P *define in* (9) *is reversible with respect to* π*.* π *is a stationary distribution for* P*.*

Proof. For any function f and g defined on $\mathbb{R}^d$,

$$\begin{aligned}
&\iint \pi(x)P(x,\mathrm{d}y)f(x)g(y)\mathrm{Leb}(\mathrm{d}x) \\
&= \iint \pi(x)\left[\alpha(x,y)q(x,y) + 1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)\right] \\
&\quad \cdot f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(x)\alpha(x,y)q(x,y)f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&\quad + \iint \pi(x)1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y). \qquad (10)
\end{aligned}$$

The first term, from Lemma 3.3,

$$\begin{aligned}
&\iint \pi(x)\alpha(x,y)q(x,y)f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(y)\alpha(y,x)q(y,x)f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(x)\alpha(x,y)q(x,y)f(y)g(x)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y),
\end{aligned}$$

and, because the second term exists only when $x = y$,

$$\begin{aligned}
&\iint \pi(x)1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)f(x)g(y)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(x)1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)f(y)g(x)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y).
\end{aligned}$$

So (10) equals

$$\begin{aligned}
&\iint \pi(x)\alpha(x,y)q(x,y)f(y)g(x)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&\quad + \iint \pi(x)1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)f(y)g(x)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(x)\left[\alpha(x,y)q(x,y) + 1(x=y)\int q(y,z)[1-\alpha(y,z)]\mathrm{Leb}(\mathrm{d}z)\right] \\
&\quad \cdot f(y)g(x)\mathrm{Leb}(\mathrm{d}x)\mathrm{Leb}(\mathrm{d}y) \\
&= \iint \pi(x)P(x,\mathrm{d}y)f(y)g(x)\mathrm{Leb}(\mathrm{d}x)
\end{aligned}$$

This shows that P is reversible with respect to π. □

Example 3.5. [Random Walk Metropolis Algorithm] Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) introduced a symmetric random walk algorithm. If the current state is X_t, an increment Z_{t+1} is drawn and the candidate $Y_{t+1} = X_t + Z_{t+1}$ is proposed. If the distribution of Z_{t+1} is symmetric with respect to 0, then for all $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ we have $q(x, y) = q(y, x)$, and

$$\alpha(x,y) = 1 \wedge \frac{\pi(y)}{\pi(x)}.$$

A classical choice is the multivariate normal distribution with zero-mean and covariance matrix Γ. It is well known that either too small or too large a covariance matrix will result in highly positively related Markov chains. When the covariance is too small, almost all the moves are accepted but the chain mixes slowly *(because $\pi(y)/\pi(x)$ is close to 1)*. When the scale is too large, most of the moves get rejected *(when covariance is large, y tends to be too large or small and, therefore, $\pi(y)$ tends to be small)* and the chain remains stuck for long intervals. In practice, this covariance matrix Γ is determined by trial and error.

Example 3.6 (Independent Sampler)**.** Another possibility is to set the transition density to be $q(x, y) = \bar{q}(y)$. In this case, the next candidate is drawn independently of the current state of the chain. This method is closely related to the Acceptance-Rejection method introduced in Section 6.3.1.

3.2 Gibbs Sampling

Each step of the Gibbs sampling algorithm corresponds to a very special type of Metropolis-Hastings move where the acceptance possibility is equal to 1, due to the choice of the proposal distribution.

The name stems from Geman and Geman (1984).

Assume that $\mathsf{X} = \mathsf{X}_1 \times \cdots \times \mathsf{X}_m$ be a space equipped with the σ-algebra $\mathcal{X} = \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_m$. Let λ_k, $k = 1, \ldots, m$, be finite measures on $(\mathsf{X}, \mathcal{X})$, and let $\lambda = \lambda_1 \otimes \cdots \otimes \lambda_m$ be the product measure.

Suppose we have a joint distribution with probability density function π with respect to the measure λ *(meaning that the joint distribution is absolutely continuous with respect to λ, and the Radon-Nikodym density is π)* .

An element of X $x = (x^{[1]}, \ldots, x^{[m]})$, where $x^{[k]} \in \mathsf{X}_k$. Denote $\pi_k(\cdot|x^{[-k]})$ the conditional probability density function defined as

$$\pi_k(x^{[k]}|x^{[-k]}) = \frac{\pi(x^{[1]}, \ldots, x^{[m]})}{\int \pi(x^{[1]}, \ldots, x^{[m]})\lambda(\mathrm{d}x^{[k]})}.$$

Starting from an arbitrary state X_0, the *deterministic scan Gibbs* sampler is the MCMC which updates the current state $X_i = (X_i^{[1]}, \ldots, X_i^{[m]})$ to a new state X_{i+1} as follows. For $k = 1, \ldots, m$, simulate $X_{i+1}^{[k]}$ from $\pi_k(\cdot|x_{i+1}^{[1]}, \ldots, x_{i+1}^{[k-1]}, x_i^{[k+1]}, \ldots, x_i^{[m]})$. A complete cycle of the m conditional simulations is usually referred to as a *sweep* of the algorithm. We denote by K_k the corresponding Markov kernel:

$$K_k(x, A) = \int \cdots \int \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}x'^{[j]}) \pi_k(x'^{[k]}|x^{[-k]}) 1(x' \in A) \lambda(\mathrm{d}x'^{[k]}).$$

Theorem 3.7. *Each of these m individual kernels K_k, $k = 1, \ldots, m$, is reversible with respect to π and thus admits π as a stationary probability density function.*

Proof. For any $f, g \in \mathbb{F}(\mathsf{X}, \mathcal{X}, \lambda)$,

$$
\begin{aligned}
&\int \cdots \int \pi(x) K_k(x, \mathrm{d}y) f(x) g(y) \lambda(\mathrm{d}x) \\
&= \int \cdots \int \pi(x) \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}y^{[j]}) \pi_k(y^{[k]} | x^{[-k]}) f(x) g(y) \lambda_k(\mathrm{d}y^{[k]}) \lambda(\mathrm{d}x) \\
&= \int \cdots \int \pi(x^{[k]} | x^{[-k]}) \pi(x^{[-k]}) \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}y^{[j]}) \pi_k(y^{[k]} | y^{[-k]}) f(x) g(y) \lambda_k(\mathrm{d}y^{[k]}) \lambda(\mathrm{d}x) \\
&= \int \cdots \int \pi(x^{[k]} | y^{[-k]}) \pi(y^{[-k]}) \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}y^{[j]}) \pi_k(y^{[k]} | y^{[-k]}) f(x) g(y) \lambda_k(\mathrm{d}y^{[k]}) \lambda(\mathrm{d}x) \\
&= \int \cdots \int \pi(x^{[k]} | y^{[-k]}) \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}y^{[j]}) \pi_k(y) f(x) g(y) \lambda_k(\mathrm{d}y^{[k]}) \lambda(\mathrm{d}x)
\end{aligned}
$$

Because $\delta_{x^{[j]}}(\mathrm{d}y^{[j]}) = \delta_{y^{[j]}}(\mathrm{d}x^{[j]})$ when $j \neq k$,

$$
\begin{aligned}
&\int \cdots \int \pi(x^{[k]} | y^{[-k]}) \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}y^{[j]}) \pi_k(y) f(x) g(y) \lambda_k(\mathrm{d}y^{[k]}) \lambda(\mathrm{d}x) \\
&= \int \cdots \int \pi(x^{[k]} | y^{[-k]}) \prod_{j \neq k} \delta_{y^{[j]}}(\mathrm{d}x^{[j]}) \pi_k(y) f(x) g(y) \lambda_k(\mathrm{d}y) \lambda(\mathrm{d}x^{[k]}) \\
&= \int \cdots \int \pi_k(y) \prod_{j \neq k} \delta_{y^{[j]}}(\mathrm{d}x^{[j]}) \pi(x^{[k]} | y^{[-k]}) f(x) g(y) \lambda_k(\mathrm{d}y) \lambda(\mathrm{d}x^{[k]}) \\
&= \int \cdots \int \pi_k(y) K_k(y, \mathrm{d}x) f(x) g(y) \lambda(\mathrm{d}y) \\
&= \int \cdots \int \pi_k(x) K_k(x, \mathrm{d}y) f(y) g(x) \lambda(\mathrm{d}x)
\end{aligned}
$$

□

4 State-Space Models

Based on Douc et al. (2014). In this section, state-space models are also called hidden Markov models.

4.1 Definitions

Definition 4.1 (Hidden Markov Model)**.** Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{X}, \mathcal{Y})$ be two measurable spaces and let M and G denote, respectively, a Markov kernel on $(\mathsf{X}, \mathcal{X})$ and a Markov kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Denote by K the Markov kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$:

$$K(x; C) = \iint_C M(x, \mathrm{d}x') G(x', \mathrm{d}y'), \quad x, x' \in \mathcal{X}, C \in \mathcal{X} \otimes \mathcal{Y}. \tag{11}$$

If $\{X_t, t \in \mathbb{N}\}$ is not observable, then the Markov chain $\{(X_t, Y_t), t \in \mathbb{N}\}$ with Markov kernel K and initial distribution $\Xi \otimes G$, where Ξ is a probability measure on $(\mathsf{X}, \mathcal{X})$, is called a hidden Markov model (HMM).

Denote by $\mathbb{P}_\Xi$ the probability measure associated with the HMM.

Property 4.2. *Let $\{(X_t, Y_t), t \in \mathbb{N}\}$ be a Markov chain over $\mathsf{X} \times \mathsf{Y}$ with the transition kernel K given by Equation (11). Then, for any integer p and any ordered set $\{t_1 < \ldots < t_p\}$ of indexes, the variables $Y_{t_1}, \ldots, Y_{t_p}$ are $\mathbb{P}_\Xi$-conditionally independent given $X_{t_1}, \ldots, X_{t_p}$, i.e. for all functions $f_{t_1}, \ldots, f_{t_p} \in \mathbb{F}_b(\mathsf{Y}, \mathcal{Y})$,*

$$\mathrm{E}_\Xi \left[\prod_{i=1}^{p} f_i(Y_{t_i}) \middle| X_{t_1}, \ldots, X_{t_p} \right] = \prod_{i=1}^{p} \int_{\mathsf{Y}} G(X_{t_i}, \mathrm{d}y_{t_i}) f_i(y_{t_i}).$$

An HMM is said to be *partially dominated* if there exists a probability measure μ on $(\mathsf{Y}, \mathcal{Y})$ such that for all $x \in \mathsf{X}$, $G(x, \cdot) \ll \mu(\cdot)$ with transition density function $g(x, \cdot)$. Then, for $A \in \mathcal{Y}$, $G(x, A) = \int_A g(x, y) \mu(\mathrm{d}y)$. The joint kernel

$$K(x; C) = \iint_C M(x, \mathrm{d}x') g(x', y') \mu(\mathrm{d}y'), \quad C \in \mathcal{X} \otimes \mathcal{Y}.$$

For a partially dominated HMM, the joint probability of the unobservable states and

observations up to time t is such that, for any function $h \in \mathbb{F}_b((\mathsf{X}\times\mathsf{Y})^{t+1}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes(t+1)})$,

$$\mathrm{E}_\Xi\left[h(X_0, Y_0, \ldots X_t, Y_t)\right] = \int\cdots\int_{(\mathsf{X}\times\mathsf{Y})^{t+1}} h(x_0, y_0, \ldots, x_t, y_t) \times \Xi(\mathrm{d}x_0)g(x_0, y_0)\prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s)g(x_s, y_s)\prod_{s=0}^{t}\mu(\mathrm{d}y_s),$$

and the joint probability of the observations

$$\mathrm{p}_\Xi(y_{0:t}) = \int\cdots\int_{\mathsf{X}^{t+1}} \Xi(\mathrm{d}x_0)g(x_0, y_0)\prod_{s=1}^{t} M(x_{s-1}, \mathrm{d}x_s)g(x_s, y_s).$$

A partially dominated HMM is fully dominated if there exists a probability measure λ on $(\mathsf{X}, \mathcal{X})$ such that $\Xi \ll \lambda$ with density ξ and, for all $x \in \mathsf{X}$, $M(x, \cdot) \ll \lambda(\cdot)$ with transition density $m(x, \cdot)$. Then, for all $x \in \mathsf{X}$, the product $\lambda \otimes \mu$ is a probability measure on $(\mathsf{X}\otimes\mathsf{Y}, \mathcal{X}\otimes\mathcal{Y})$. And for all $x \in \mathsf{X}$, the joint kernel $K \ll \lambda \otimes \mu$ with joint density

$$k(x; x', y') = m(x, x')g(x', y'), \quad (x', y') \in \mathsf{X}\times\mathsf{Y}.$$

4.2 Filtering and Smoothing

Statistical inference for state-space models involves computing the posterior distribution, $\phi_{\xi,s:s'|t}$, of a collection of state variables $X_{s:s'} = (X_s, \ldots, X_{s'})$, with $s < s'$ conditional on a batch of observations, $Y_{0:t}$. We abbreviate $\phi_{\xi,t|t}$ to $\phi_{\xi,t}$.

4.2.1 Filtering

We denote

$$g_t(x_t) \equiv g(x_t, Y_t), \tag{12}$$

$$Q_t(x_{t-1}, A) \equiv \int_A M(x_{t-1}, \mathrm{d}x_t)g_t(x_t), \quad A \in \mathcal{X}. \tag{13}$$

So $Q_t : \mathsf{X} \to \mathcal{X}$, $t \in \mathbb{N}_+$, is a kernel.

Let Ξ denote the prior distribution of X_0. A measure $\gamma_{\Xi,t} \in \mathbb{M}_+(\mathcal{X})$ is defined such that, for any $A \in \mathcal{X}$,

$$\gamma_{\Xi,t}(A) = \int\cdots\int_{\mathsf{X}^t\times A} \Xi(\mathrm{d}x_0)g_0(x_0)\prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s),$$

For $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$. we write

$$\gamma_{\Xi,t}(f) = \int \cdots \int_{\mathsf{X}^{t+1}} \Xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s) f(x_t).$$

This distribution may be computed recursively as follows

$$\begin{aligned}\gamma_{\Xi,0}(f) &= \int \Xi(\mathrm{d}x_0) g_0(x_0) f(x_0) \\ &= \Xi(g_0 f),\end{aligned}$$

and

$$\begin{aligned}\gamma_{\Xi,t}(f) &= \iint \gamma_{\Xi,t-1}(\mathrm{d}x_{t-1}) Q_t(x_{t-1}, \mathrm{d}x_t) f(x_t) \\ &= \gamma_{\Xi,t-1} Q_t(f).\end{aligned}$$

The joint distribution of the observations $(Y_0, \ldots, Y_t)$

$$\begin{aligned}\mathrm{p}_{\Xi,t}(Y_{0:t}) &= \int \gamma_{\Xi,t}(\mathrm{d}x_t) \\ &= \gamma_{\Xi,t}(1).\end{aligned}$$

The filtering distribution, i.e. the conditional distribution of X_t given $Y_{0:t}$, is obtained by dividing the joint distribution of $(X_t, Y_0, \ldots, Y_t)$ by $\mathrm{p}_{\Xi,t}(Y_{0:t})$

$$\begin{aligned}\phi_{\Xi,t}(f) &= \frac{\gamma_{\Xi,t}(f)}{\gamma_{\Xi,t}(1)} \\ &= \frac{\int \cdots \int_{\mathsf{X}^{t+1}} \Xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s) f(x_t)}{\int \cdots \int_{\mathsf{X}^{t+1}} \Xi(\mathrm{d}x_0) g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s)} \\ &= \frac{\iint \phi_{\Xi,t-1}(\mathrm{d}x_{t-1}) Q_t(x_{t-1}, \mathrm{d}x_t) f(x_t)}{\iint \phi_{\Xi,t-1}(\mathrm{d}x_{t-1}) Q_t(x_{t-1}, \mathrm{d}x_t)} \\ &= \frac{\phi_{\Xi,t-1} Q_t(f)}{\phi_{\Xi,t-1} Q_t(1)}. \end{aligned} \tag{14}$$

So $\phi_{\Xi,t} \in \mathbb{M}_1(\mathcal{X})$, $t \in \mathbb{N}$. Note that $\phi_{\Xi,t-1} Q_t$ is the distribution of (x_t, Y_t) conditional on $Y_{0:t-1}$, and $\phi_{\Xi,t-1} Q_t(1)$ is the distribution of Y_t conditional on $Y_{0:t-1}$.

To highlight a two-step procedure involving both the predictive and filtering distributions, with the convention that $\phi_{\Xi,0|-1} = \Xi$, (14) may be decomposed as

$$\phi_{\Xi,t|t-1}(f) = \phi_{\Xi,t-1} M(f), \tag{15}$$

$$\phi_{\Xi,t}(f) = \frac{\phi_{\Xi,t|t-1}(f g_t)}{\phi_{\Xi,t|t-1}(g_t)}, \tag{16}$$

where $\phi_{\Xi,t|t-1} \in \mathbb{M}_1(\mathcal{X})$, $t \in \mathbb{N}$, is the distribution of x_t conditional on $Y_{0:t-1}$; $\phi_{\Xi,t|t-1}(fg_t)$ is the joint distribution of (x_t, Y_t) conditional on $Y_{0:t-1}$; and $\phi_{\Xi,t|t-1}(g_t)$ is the distribution of Y_t conditional on $Y_{0:t-1}$.

4.2.2 Smoothing

The joint smoothing distribution $\phi_{\Xi,0:t|t}$ then satisfies, for $f \in \mathbb{F}_+(\mathcal{X}^{\otimes(t+1)})$,

$$\phi_{\Xi,0:t|t}(f) = \left(\mathrm{p}_{\Xi,t}(Y_{0:t})\right)^{-1} \int \cdots \int f(x_0, \ldots, x_t)\Xi(\mathrm{d}x_0)g_0(x_0) \prod_{s=1}^{t} Q_s(x_{s-1}, \mathrm{d}x_s). \tag{17}$$

Likewise, for indices $p \geq 0$,

$$\phi_{\Xi,0:t+p|t}(f) = \left(\mathrm{p}_{\Xi,t}(Y_{0:t})\right)^{-1} \int \cdots \int f(x_0, \ldots, x_{t+p})\phi_{\Xi,0:t|t}(\mathrm{d}x_0, \ldots, \mathrm{d}x_t) \prod_{s=t+1}^{t+p} M(x_{s-1}, \mathrm{d}x_s) \tag{18}$$

for all functions $f \in \mathbb{F}_+(\mathcal{X}^{\otimes(t+p+1)})$.

The joint smoothing distribution (17) implicitly defines all other particular cases of smoothing distributions. For instance, the marginal smoothing distribution $\phi_{\Xi,t|n}$ for $0 \leq t \leq n$ is such that, for $f \in \mathbb{F}_+(\mathcal{X})$,

$$\phi_{\Xi,t|n}(f) = \int \cdots \int f(x_t)\phi_{\Xi,0:n|n}(\mathrm{d}x_0, \ldots, \mathrm{d}x_n).$$

Similarly, the p-step predictive distribution $\phi_{\Xi,n+p|n}$ may be obtained by marginalization of the joint distribution $\phi_{\Xi,0:n+p|n}$. (18) suggests that

$$\phi_{\Xi,n+p|n}(f) = \phi_{\Xi,n}M^p(f).$$

Definition 4.3 (Backward Kernel)**.** Let $\eta \in \mathbb{M}_1(\mathcal{X})$ and B_η be a kernel on $(\mathsf{X}, \mathcal{X})$. If for all $h \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

$$\iint h(x, x')\eta(\mathrm{d}x)M(x, \mathrm{d}x') = \iint h(x, x')\eta M(\mathrm{d}x')B_\eta(x', \mathrm{d}x), \tag{19}$$

then B_η is referred to as the *backward kernel*, where $\eta M(\cdot)$ is a probability measure on $(\mathsf{X}, \mathcal{X})$ defined as

$$\eta M(A) = \int \eta(\mathrm{d}x) \int M(x, \mathrm{d}x')1(x' \in A), \quad \text{for } A \in \mathcal{X}.$$

The backward kernel $B_\eta(x', x)$ denotes the conditional distribution of X given X'. When the HMM is fully dominated(see Section 4.1), then the backward kernel may be explicitly written as

$$B_\eta(x', A) = \frac{\int \eta(\mathrm{d}x) m(x, x') 1(x \in A)}{\int \eta(\mathrm{d}x) m(x, x')}. \tag{20}$$

Note that $B_\eta(x', x)$ denotes the conditional distribution of X_t given X_{t+1}.

Property 4.4. *Given a positive integer n, initial distribution Ξ, and $t = 0, 1, \ldots, n-1$,*

$$\mathrm{E}[f(X_t)\big|X_{t+1:n}, Y_{0:n}] = \mathrm{E}[f(X_t)\big|X_{t+1}, Y_{0:t}] = B_{\phi_{\Xi,t}} f(X_{t+1}) \tag{21}$$

for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$, where $B_{\phi_{\Xi,t}} f$ is as defined in Section 2. In addition,

$$\phi_{\Xi,t|n} = \phi_{\Xi,t+1|n} B_{\phi_{\Xi,t}}, \tag{22}$$

and

$$\phi_{\Xi,0:n|n}(f(X_{0:n})) = \int \cdots \int f(x_{0:n}) \phi_{\Xi,n}(\mathrm{d}x_n) \prod_{s=0}^{n-1} B_{\phi_{\Xi,s}}(x_{s+1}, \mathrm{d}x_s) \tag{23}$$

for any $f \in \mathbb{F}(\mathsf{X}^{n+1}, \mathcal{X}^{\otimes(n+1)})$.

Proof. We will prove (22). For $A \in \mathcal{X}$,

$$\begin{aligned}
\phi_{t+1|n} B_{\phi_t}(A) &= \iint 1(x \in A) \phi_{t+1|n}(\mathrm{d}x') B_{\phi_t}(x', \mathrm{d}x) \\
&= \iint 1(x \in A) \phi_{t|n} M(\mathrm{d}x') B_{\phi_t}(x', \mathrm{d}x),
\end{aligned}$$

where, from (21),

$$B_{\phi_t} = B_{\phi_{t|n}}.$$

So we have

$$\begin{aligned}
\phi_{t+1|n} B_{\phi_t}(A) &= \iint 1(x \in A) \phi_{t|n} M(\mathrm{d}x') B_{\phi_{t|n}}(x', \mathrm{d}x) \\
&= \iint 1(x \in A) \phi_{t|n}(\mathrm{d}x) M(x, \mathrm{d}x')
\end{aligned}$$

by Definition 4.3, and

$$\begin{aligned}\iint 1(x \in A)\phi_{t|n}(\mathrm{d}x)M(x, \mathrm{d}x') &= \int 1(x \in A)\phi_{t|n}(\mathrm{d}x) \int M(x, \mathrm{d}x') \\ &= \int 1(x \in A)\phi_{t|n}(\mathrm{d}x) \\ &= \phi_{t|n}(A).\end{aligned}$$

□

Note that the backward kernel $B_{\phi_{\Xi,t}}$ denotes the conditional distribution $X_t|X_{t+1}, y_{0:t}$, and $B_{\phi_{\Xi,t|n}}$ denotes the conditional distribution $X_t|X_{t+1}, y_{0:n}$.

Algorithm 4.1 (Forward Filtering Backward Smoothing)**.**

1. ***Forward Filtering*** Compute, forward in time, the filtering distributions $\phi_{\Xi,0}, \ldots, \phi_{\Xi,n}$ using the recursion (14). At each index t, the backward transition kernel maybe computed according to (19).

2. ***Backward Smoothing*** From $\phi_{\Xi,n}$, compute

   $$\phi_{\Xi,t|n} = \phi_{\Xi,t+1|n} B_{\phi_{\Xi,t}}.$$

   for $t = n-1, n-2, \ldots, 0$.

5 Particle Filtering

Sequential Monte Carlo (SMC) is a class of methods designed to approximate a sequence of probability distributions. SMC uses a set of particles, of which each has an assigned weight and is updated recursively.

SMC methods are a combination of the sequential importance sampling method introduced in Handschin and Mayne (1969) and the sampling importance resampling algorithm proposed in Rubin (1987).

5.1 Importance Sampling

μ denotes a probability measure on a measurable space $(\mathsf{X}, \mathcal{X})$, which is referred to as the *target distribution*. The aim of importance sampling is to approximate integrals of the form

$$\mu(f) = \int f(x)\mu(\mathrm{d}x), \quad f \in (\mathsf{X}, \mathcal{X}).$$

The plain Monte Carlo approach consists in drawing an i.i.d. sample $\{X^i\}_{i=1}^N$ from the target distribution μ and then evaluating the sample mean $N^{-1}\sum_{i=1}^N f(X^i)$.

In certain situation, it is more appropriate to sample from a *proposal distribution* $\nu \in \mathbb{M}_1(\mathcal{X})$. Importance sampling is based on that idea. Assume that the target distribution μ is absolutely continuous with respect to ν, and denote by $w = \mathrm{d}\mu/\mathrm{d}\nu$ the Radon-Nikodym derivative (Theorem 1.40) of μ with respect to ν, referred to in the sequel as the *weight function*.

Then, for $f \in L^1(\mu)$,

$$\mu(f) = \int f(x)\mu(\mathrm{d}x) = \int f(x)w(x)\nu(\mathrm{d}x) = \nu(fw).$$

If $\{X^i\}_{i=1}^N$ is an i.i.d. sample from ν, the above equation suggests the following estimator of $\mu(f)$:

$$N^{-1}\sum_{i=1}^N f(X^i)w(X^i),$$

which is called the *importance sampling estimator*.

When applying importance sampling ideas to solve problems in NLSS, in many situations, the target probability measure μ is known only up to a normalizing/scaling factor. That is, we only know $\mu' \in \mathbb{M}_+(\mathcal{X})$, which is absolutely continuous to ν and such that $\mu'/\mu'(1) = \mu$. The weigh function $w = \mathrm{d}\mu'/\mathrm{d}\nu$ is then known up to a scaling factor only. Then the following self-normalizing form of importance sampling estimator of $\mu(f)$ can be used

$$\sum_{i=1}^{N} \frac{w(X^i)}{\sum_{j=1}^{N} w(X^j)} f(X^i),$$

because

$$\frac{1}{N}\sum_{i=1}^{N} w(X^i) f(X^i) \xrightarrow{P} \mu'(f),$$
$$\frac{1}{N}\sum_{i=1}^{N} w(X^i) \xrightarrow{P} \mu'(1).$$

$\{(X^i, w(X^i))_{i=1}^N\}$ is called a *weighted sample*.

Importance sampling introduces little restrictions on the choice of the proposal distribution. The choice is typically guided by two requirements: the proposal distribution should be easy to simulate and should lead to efficient estimator.

Example 5.1. Assume the target distribution is a Gaussian mixture, with density $p(x) = \alpha \mathrm{g}(x; m_1, \sigma_1^2) + (1-\alpha)\mathrm{g}(x; m_2, \sigma_2^2)$. To see what is a Gaussian mixture, refer to Example 5.13. A natural choice for the proposal distribution is the Student t_κ-distribution,

$$q_\kappa(x) = \frac{\Gamma\left(\frac{\kappa+1}{2}\right)}{\sqrt{\kappa\pi}\Gamma\left(\frac{\kappa}{2}\right)}\left(1 + \frac{x^2}{\kappa}\right)^{-\frac{\kappa+1}{2}},$$

where κ is the number of degrees of freedom and Γ is the Gamma function.

Alternative proposal distributions are $s^{-1}q_\kappa(s^{-1}x)$, where $s \in \mathbb{R}_+$ is the scale.

Note that in this example, we know the density of the target distribution. Therefore, the weight function w can be easily derived, which is the ratio of the two densities $\mathrm{d}\mu/\mathrm{d}\nu$.

Definition 5.2 (Consistent Weighted Sample)**.** A weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ of size

N is consistent for the probability measure $\mu \in \mathbb{M}_1(\mathcal{X})$ if, for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$, as $N \to \infty$,

$$\sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} f(X^{N,i}) \xrightarrow{P} \mu(f),$$

$$\max_{1 \le i \le N} \frac{\omega^{N,i}}{\Omega^N} \xrightarrow{P} 0,$$

where Ω^N is the sum of the importance weights

$$\Omega^N = \sum_{i=1}^{N} \omega^{N,i}. \tag{24}$$

Definition 5.3. A weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ of sample size N, $N \in \mathbb{N}$, is said to be adapted to $\mathcal{F}^N \subset \mathcal{F}$ if $\sigma(\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N) \subset \mathcal{F}^N$.

Now suppose that we already have a weighted sample $\{(X^{N,i}, \omega^{N,i})_{i=1}^N\}$ which is consistent for the probability measure $\nu \in \mathbb{M}_1(\mathcal{X})$. How do we get a weight sample consistent for a dominated measure $\mu \in \mathbb{M}_+(\mathcal{X})$?

Property 5.4. *Let $(\mathsf{X}, \mathcal{X}, \nu)$ be a probability space, $\mu \in \mathbb{M}_+(\mathcal{X})$, and $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ be a weighted sample consistent for ν. If $\mu \ll \nu$, and $w \in L^1(\mathsf{X}, \mathcal{X}, \nu)$ is the Radon-Nikodym derivative of μ with respect to ν, setting $\tilde{\omega}^{N,i} = w(X^{N,i})\omega^{N,i}$, then $\{(X^{N,i}, \tilde{\omega}^{N,i})_{i=1}^N\}$ is a weighted sample consistent for μ.*

Now suppose that we already have a weighted sample $\{(X^{N,i}, \omega^{N,i})_{i=1}^N\}$ which is consistent for the probability measure $\nu \in \mathbb{M}_1(\mathcal{X})$. We consider constructing a weighted sample consistent for the following probability measure

$$\mu = \frac{\nu Q}{\nu Q(1)}, \tag{25}$$

where Q is a bounded [2] kernel (not necessarily Markov) on $\mathsf{X} \times \mathcal{X}$; νQ is as defined by (7). Therefore, $\mu \in \mathbb{M}_1(\mathcal{X})$.

Consider a Markov kernel R on $\mathsf{X} \times \mathcal{X}$. Assume there exists a function $w : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ such that for all $x \in \mathsf{X}$ and for all $A \in \mathcal{X}$,

$$Q(x, A) = \int_A w(x, x') R(x, \mathrm{d}x'). \tag{26}$$

[2] On Page 324, Douc et al. (2014), Q is only required to be finite. However, the proof of Property 5.5 clearly needs Q to be bounded.

If, for all $x \in \mathcal{X}$, Q and R have densities denoted by q and r with respect to the same dominating measure, then we set

$$w(x, x') = \begin{cases} q(x, x')/r(x, x') & \text{if } r(x, x') \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

A new weighted sample $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})_{i=1}^N\}$ is constructed as follows.

Firstly, for $i = 1, \ldots, N$, we draw $\tilde{X}^{N,i}$ from the proposal Markov kernel $R(X^{N,i}, \cdot)$. By construction, for any $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$,

$$\begin{aligned} \mathrm{E}[f(\tilde{X}^{N,i}|\mathcal{F}^N)] &= \int_{\mathsf{X}} R(X^{N,i}, \mathrm{d}x) f(x) \\ &= Rf(X^{N,i}), \end{aligned} \tag{27}$$

where $\mathcal{F}^N$ is a σ-algebra such that $\sigma(\{(X^{N,i}, \omega^{N,i})_{i=1}^N\}) \subset \mathcal{F}^N$. We can take the σ-algebra $\mathcal{F}^N = \sigma\left(\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N\right)$. But $\mathcal{F}^N$ can be chosen larger than that, because $R(\cdot, \cdot)$ is a Markov kernel (we will see example of this later when we apply these results sequentially).

We then associate each particle position with the following weight:

$$\tilde{\omega}^{N,i} = \omega^{N,i} w(X^{N,i}, \tilde{X}^{N,i}), \quad i = 1, \ldots, N. \tag{28}$$

Property 5.5. *Assume that the weighted sample $\{(X^{N,i}, \omega^{N,i})_{i=1}^N\}$ is adapted to $\mathcal{F}^N$ and consistent for ν. Then, the weighted sample $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})_{i=1}^N\}$ defined by (27) and (28) is consistent for μ defined by (25).*

Proof. We show first that for any $f \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

$$\frac{1}{\Omega^N} \sum_{i=1}^N \tilde{\omega}^{N,i} f(X^{N,i}, \tilde{X}^{N,i}) \xrightarrow{\mathbb{P}} \nu \otimes Q(f), \tag{29}$$

where $\Omega^N = \sum_i \omega^{N,i}$; the tensor product $\nu \otimes Q$ is a measure on $\mathcal{X}^{\otimes 2}$, and

$$\nu \otimes Q(f) \equiv \iint \nu(\mathrm{d}x) Q(x, \mathrm{d}x') f(x, x').$$

The definition (26) implies that

$$\begin{aligned} \mathrm{E}[\tilde{\omega}^{N,i} f(X^{N,i}, \tilde{X}^{N,i})|\mathcal{F}^N] &= \omega^{N,i} \int w(X^{N,i}, x) f(X^{N,i}, x) R(X^{N,i}, \mathrm{d}x) \\ &= \omega^{N,i} \int f(X^{N,i}, x) Q(X^{N,i}, \mathrm{d}x). \end{aligned} \tag{30}$$

Define a probability measure $\delta_{X^{N,i}} \in \mathbb{M}_+(\mathsf{X}, \mathcal{X})$ such that

$$\delta_{X^{N,i}}(x) = \begin{cases} 1 & \text{if } x = X^{N,i} \\ 0 & \text{otherwise} \end{cases}.$$

Then

$$\int f(X^{N,i}, x) Q(X^{N,i}, \mathrm{d}x) = \delta_{X^{N,i}} \otimes Q(f),$$

and (30) implies that

$$\mathrm{E}[\tilde{\omega}^{N,i} f(X^{N,i}, \tilde{X}^{N,i}) | \mathcal{F}^N] = \omega^{N,i} \delta_{X^{N,i}} \otimes Q(f).$$

Therefore, we get

$$\sum_{i=1}^{N} \mathrm{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) \middle| \mathcal{F}^N \right] = \sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} \delta_{X^{N,i}} \otimes Q(f).$$

Note that $\delta_x \otimes Q(f)$ is a measurable function on $(\mathsf{X}, \mathcal{X})$. And because Q is a bounded kernel [3], according to Property 1.34, the function $x \to \delta_x \otimes Q(f)$ is bounded, i.e. $\delta_x \otimes Q(f) \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$. Recall that $\nu \in \mathbb{M}_1(\mathcal{X})$, $\{(X^{N,i}, \omega^{N,i})_{i=1}^N\}$ is consistent for ν, then, by definition 5.2,

$$\sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} \delta_x \otimes Q(f) \xrightarrow{\mathbb{P}} \nu(\delta_x \otimes Q(f)),$$

where

$$\begin{aligned} \nu(\delta_x \otimes Q(f)) &= \int \delta_x \otimes Q(f) \nu(\mathrm{d}x) \\ &= \iint f(x, x') \nu(\mathrm{d}x) Q(x, \mathrm{d}x') \\ &= \nu \otimes Q(f). \end{aligned}$$

Therefore,

$$\sum_{i=1}^{N} \mathrm{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) \middle| \mathcal{F}^N \right] \xrightarrow{\mathbb{P}} \nu \otimes Q(f). \tag{31}$$

Now we will show that

$$\sum_{i=1}^{N} \left\{ \frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) - \mathrm{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) \middle| \mathcal{F}^N \right] \right\} \xrightarrow{\mathbb{P}} 0. \tag{32}$$

[3] This is where the boundedness Q is required, as mentioned before in Footnote 2.

appealing to Theorem 1.59. Take $U_{N,i} = \left(\tilde{\omega}^{N,i}/\Omega^N\right) f(X^{N,i}, \tilde{X}^{N,i})$ for $i = 1, \ldots, N$. Then $\{(U_{N,i}|\mathcal{F}^N)_{i=1,\ldots,N}\}_{N\in\mathbb{N}^+}$ is a triangular array of random variables on $(\mathsf{X}, \mathcal{X})$, $\{\mathcal{X}\}$ is a triangular array of σ-fields, and, for each $N \in \mathbb{N}^+$ and each $i = 1, \ldots, N$, $U_{N,i}|\mathcal{F}^N$ is $\mathcal{X}$-measurable. Note that, for each $N \in \mathbb{N}^+$ and each $i = 1, \ldots, N$, from Property 1.34,

$$\begin{aligned}
\mathrm{E}\left[|U_{N,i}|\middle|\mathcal{F}_N\right] &= \mathrm{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N}\left|f(X^{N,i}, \tilde{X}^{N,i})\right|\middle|\mathcal{F}^N\right] \\
&= \frac{\omega^{N,i}}{\Omega^N}\int f(X^{N,i}, x)Q(X^{N,i}, \mathrm{d}x) \\
&< \infty,
\end{aligned}$$

and, from (31),

$$\begin{aligned}
\sum_{i=1}^{N}\mathrm{E}\left[|U_{N,i}|\middle|\mathcal{F}\right] &\xrightarrow{P} \nu\otimes Q(|f|) \\
&< \infty.
\end{aligned}$$

Denote $Y_N = \sum_{i=1}^{N}\mathrm{E}\left[|U_{N,i}|\middle|\mathcal{F}\right]$. Then, $\{Y_N\}_{N\in\mathbb{N}^+}$ is a monotonically increasing sequence with limit $\nu\otimes Q(|f|)$. So, when $\lambda > \nu\otimes Q(|f|)$, for any $N \in \mathbb{N}^+$,

$$P(Y_N \geq \lambda) = 0.$$

For that reason,

$$\sup_N P(Y_N \geq \lambda) = 0,$$

showing that the tightness condition (3) holds. To check the asymptotic negligibility condition (4), for all $\varepsilon > 0$, denote $A_N = \sum_{i=1}^{N}\mathrm{E}\left[|U_{N,i}|\cdot 1(|U_{N,i}| > \varepsilon)\middle|\mathcal{F}^N\right]$. Then,

$$\begin{aligned}
A_N &= \sum_{i=1}^{N}\int\left[\frac{\omega^{N,i}}{\Omega^N}w(X^{N,i}, x)\cdot\left|f(X^{N,i}, x)\right|\right]_{\varepsilon} R(X^{N,i}, \mathrm{d}x) \\
&= \sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\int\left[w(X^{N,i}, x)\cdot\left|f(X^{N,i}, x)\right|\right]_{(\varepsilon\Omega^N)/\omega^{N,i}} R(X^{N,i}, \mathrm{d}x)
\end{aligned}$$

So for all $C > 0$,

$$\begin{aligned}
&A_N \cdot 1\left(\max_{1\le i\le N}\frac{\omega^{N,i}}{\Omega^N} \le \frac{\varepsilon}{C}\right)\\
&= \sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\int \left[w(X^{N,i},x)\cdot\left|f(X^{N,i},x)\right|\right]_{(\varepsilon\Omega^N)/\omega^{N,i}} 1\left(\max_{1\le i\le N}\frac{\omega^{N,i}}{\Omega^N} \le \frac{\varepsilon}{C}\right) R(X^{N,i},\mathrm{d}x)\\
&\le \sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\int \left[w(X^{N,i},x)\cdot\left|f(X^{N,i},x)\right|\right]_{C} R(X^{N,i},\mathrm{d}x)\\
&= \sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\delta_{X^{N,i}}\otimes R([w\cdot|f|]_C)
\end{aligned}$$

As $N\to\infty$,

$$A_N \cdot 1\left(\max_{1\le i\le N}\frac{\omega^{N,i}}{\Omega^N} \le \frac{\varepsilon}{C}\right) \to A_N,$$

and,

$$\sum_{i=1}^{N}\frac{\omega^{N,i}}{\Omega^N}\delta_{X^{N,i}}\otimes R([w\cdot|f|]_C) \xrightarrow{P} \nu\otimes R([w\cdot|f|]_C).$$

So

$$A_N \le \nu\otimes R([w\cdot|f|]_C) \quad \forall\ C>0.$$

Because

$$\nu\otimes R([w\cdot|f|]_C)\to 0 \quad \text{as } C\to 0,$$

we have

$$A_N \xrightarrow{P} 0 \quad \text{as } N\to\infty,$$

showing (4) holds. Thus, Theorem 1.59 applies, and (32) holds.

Equations (31) together with (32) shows that Equation (29) holds.

From Equation (29), we have

$$\sum_{i=1}^{N}\frac{\tilde{\omega}^{N,i}}{\Omega^N} \xrightarrow{P} \nu\otimes Q(1),$$

i.e. $\tilde{\Omega}^N/\Omega^N \xrightarrow{P} \nu \otimes Q(1)$, which, combined with (29), shows that

$$\begin{aligned}\sum_{i=1}^{N} \frac{\tilde{\omega}^{N,i}}{\tilde{\Omega}^N} f(X^{N,i}, \tilde{X}^{N,i}) &\xrightarrow{P} \frac{\nu \otimes Q(f)}{\nu \otimes Q(1)} \\ &= \mu(f).\end{aligned}$$

It remains to prove that $\max_{1\leq i\leq N} \tilde{\omega}^{N,i}/\tilde{\Omega}^N \xrightarrow{P} 0$. Because $\tilde{\Omega}^N\Omega^N \xrightarrow{P} \nu \otimes Q(1)$, it suffices to show that $\max_{1\leq i\leq N} \tilde{\omega}^{N,i}/\Omega^N \xrightarrow{P} 0$. For any $C > 0$, we have

$$\begin{aligned}\max_{1\leq i\leq N} \frac{\tilde{\omega}^{N,i}}{\Omega^N} \cdot 1(w(X^{N,i}, \tilde{N}^{N,i}) \leq C) &\leq C\frac{\tilde{\omega}^{N,i}}{\Omega^N} \\ &\xrightarrow{P} 0,\end{aligned}$$

and, by applying Equation (29),

$$\begin{aligned}\max_{1\leq i\leq N} \frac{\tilde{\omega}^{N,i}}{\Omega^N} \cdot 1(w(X^{N,i}, \tilde{N}^{N,i}) > C) &\leq \sum_{i=1}^{N} \frac{\tilde{\omega}^{N,i}}{\Omega^N} \cdot 1(w(X^{N,i}, \tilde{N}^{N,i}) > C) \\ &\xrightarrow{P} \nu \otimes Q(1(w(X^{N,i}, \tilde{N}^{N,i}) > C)) \\ &\xrightarrow{P} 0 \quad \text{as } C \to 0.\end{aligned}$$

□

Next, we discuss the asymptotic normality of the estimator. We first need to extend our definition of consistent weighted samples.

Definition 5.6 (Asymptotically Normal Weighted Sample)**.** Let $\mu \in \mathbb{M}_1(\mathcal{X})$ and $\sigma, \zeta \in \mathbb{M}_+(\mathcal{X})$. A weighted sample $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ on X is said to be asymptotically normal for (μ, σ, ζ) if, for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,

$$N^{1/2} \sum_{i=1}^{N} \frac{\omega^{N,i}}{\Omega^N} \left[f(X^{N,i}) - \mu(f)\right] \xrightarrow{D} \mathrm{N}\left(0, \sigma^2(f)\right), \tag{33}$$

$$N \sum_{i=1}^{N} \left(\frac{\omega^{N,i}}{\Omega^N}\right)^2 f(X^{N,i}) \xrightarrow{P} \zeta(f), \tag{34}$$

$$N^{1/2} \max_{1\leq i\leq N} \frac{\omega^{N,i}}{\Omega^N} \xrightarrow{P} 0, \tag{35}$$

where Ω^N is defined in (24).

Property 5.7. *Suppose the assumptions of Property 5.5 hold. Assume in addition that the weighted sample* $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ *is asymptotically normal for* (ν, σ, ζ)*. Then, the weighted sample* $\{(\tilde{X}^{N,i}, \tilde{\omega}^{N,i})\}_{i=1}^N$ *is asymptotically normal for* $(\mu, \tilde{\sigma}, \tilde{\zeta})$ *with*

$$\begin{aligned}\tilde{\zeta}(f) &= \frac{1}{[\nu Q(1)]^2} \iint \zeta(\mathrm{d}x) R(x, \mathrm{d}x') w^2(x, x') f(x') \\ &= \frac{1}{[\nu Q(1)]^2} \zeta \otimes R(w^2 f),\end{aligned}$$

and

$$\tilde{\sigma}^2(f) = \frac{\sigma^2\left(Q[f - \mu(f)]\right)}{[\nu Q(1)]^2} + \tilde{\zeta}([f - \mu(f)]^2) - \frac{\zeta\left(\{Q[f - \mu(f)]\}^2\right)}{[\nu Q(1)]^2},$$

where $Q[f - \mu(f)] : \mathsf{X} \to \mathbb{R}$ *denotes the following function:*

$$Q[f - \mu(f)] : x \to \int Q(x, \mathrm{d}x') \left[f(x') - \mu(f)\right].$$

Proof. Pick $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$ and assume, without loss of generality, that $\mu(f) = 0$. Write $\sum_{i=1}^N \tilde{\omega}^{N,i} / \tilde{\Omega}^N f(\tilde{X}^{N,i}) = (\Omega^N / \tilde{\Omega}^N)(A_N + B_N)$, with

$$\begin{aligned}A_N &= \sum_{i=1}^N \mathrm{E}\left[\frac{\tilde{\omega}^{N,i}}{\Omega^N} f(\tilde{X}^{N,i}) \middle| \mathcal{F}^N\right] \\ &= \sum_{i=1}^N \int \frac{\tilde{\omega}^{N,i}}{\Omega^N} f(x') R(X^{N,i}, \mathrm{d}x') \\ &= \sum_{i=1}^N \int \frac{\omega^{N,i}}{\Omega^N} f(x') Q(X^{N,i}, \mathrm{d}x') \\ &= \sum_{i=1}^N \frac{\omega^{N,i}}{\Omega^N} Q f(X^{N,i}),\end{aligned}$$

and

$$B_N = \sum_{i=1}^N \frac{\tilde{\omega}^{N,i}}{\Omega^N} f(X^{N,i}, \tilde{X}^{N,i}) - A_N.$$

Because $\tilde{\Omega}^N / \Omega^N \xrightarrow{P} \nu Q(1)$, the conclusion of the theorem follows if we prove that $N^{1/2}(A_N + B_N) \xrightarrow{D} \mathrm{N}(0, \sigma^2(Qf) + \eta^2(f))$ where

$$\eta^2(f) = \zeta \otimes R(w^2 f^2) - \zeta([Qf]^2).$$

Because $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^{N}$ is asymptotically normal for (ν, σ, η), $N^{1/2}A_N \xrightarrow{D} \mathrm{N}(0, \sigma^2(Qf))$. Next, we prove that for any real u,

$$\mathrm{E}\left[\exp\left(i \cdot uN^{1/2}B_N\right)\middle|\mathcal{F}^N\right] \xrightarrow{P} \exp\left(-\frac{u^2}{2}\eta^2(f)\right) \tag{36}$$

using Theorem 1.61.

Take $U_{N,i} = N^{1/2}(\tilde{\omega}^{N,i}/\Omega^N)f(X^{N,i}, \tilde{X}^{N,i})$, $i = 1, \ldots, N$. Then,

$$\begin{aligned}\sum_{i=1}^{N}\mathrm{E}\left[U_{N,i}^2\middle|\mathcal{F}^N\right] &= \int N\sum_{i=1}^{N}\left(\frac{\omega^{N,i}w(X^{N,i},x)}{\Omega^N}\right)^2 f^2(X^{N,i},x)R(X^{N,i},\mathrm{d}x)\\ &\xrightarrow{P} \zeta \otimes R(w^2f^2),\end{aligned}$$

because of Equation (34). And we have

$$\begin{aligned}\sum_{i=1}^{N}\left(\mathrm{E}\left[U_{N,i}\middle|\mathcal{F}^N\right]\right)^2 &= \sum_{i=1}^{N}\left(\int N^{1/2}\frac{\omega^{N,i}w(X^{N,i},x)}{\Omega^N}f(X^{N,i},x)R(X^{N,i},\mathrm{d}x)\right)^2\\ &\xrightarrow{P} \zeta([Qf]^2).\end{aligned}$$

So

$$\sum_{i=1}^{M_N}\left(\mathrm{E}[U_{N,i}^2|\mathcal{F}_{N,i-1}] - \mathrm{E}^2[U_{N,i}|\mathcal{F}_{N,i-1}]\right) \xrightarrow{P} \eta^2(f),$$

showing Equation (5) holds. It then remains for us to check Equation (6). Note that this condition is similar to Equation (4). So proceeding like in the preceding proof, we can show that (6) is also satisfied. So, Theorem 1.61 applies, and (36) holds.

(36) suggests a possibility that $N^{1/2}B_N \xrightarrow{D} \mathrm{N}(0, \eta^2(f))$, in light of Property 1.5. Therefore, $N^{1/2}(A_N + B_N) \xrightarrow{P} \mathrm{N}(0, \sigma^2(Qf) + \eta^2(f))$.

Now consider Equation (34). Recalling that $\tilde{\Omega}^N/\Omega^N \xrightarrow{P} \nu \otimes Q(1)$, it is sufficient to show that for $f \in \mathbb{F}_b(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$,

$$N\sum_{i=1}^{N}\left(\frac{\tilde{\omega}^{N,i}}{\Omega^N}\right)^2 f(X^{N,i}, \tilde{X}^{N,i}) \xrightarrow{P} \zeta \otimes R(w^2 f). \tag{37}$$

Define $U_{N,i} = N\left(\tilde{\omega}^{N,i}/\Omega^N\right)^2 f(X^{N,i}, \tilde{X}^{N,i})$. Then, note that, from Theorem 1.59,

$$\sum_{i=1}^{N} U_{N,i} \xrightarrow{P} \sum_{i=1}^{N}\mathrm{E}\left[U_{N,i}\middle|\mathcal{F}^N\right],$$

of which the proof is along the same lines as above in the preceding proof. And note that

$$\sum_{i=1}^{N} \mathrm{E}\left[U_{N,i} \middle| \mathcal{F}^N \right] = N \sum_{i=1}^{N} R[w^2 f](X^{N,i}) \xrightarrow{P} \zeta \otimes R(w^2 f),$$

showing that condition (34) holds.

Finally consider (35). It suffices to show that

$$C_N \equiv N \max_{1 \le i \le N} \left(\frac{\tilde{\omega}^{N,i}}{\Omega^N} \right)^2 \xrightarrow{P} 0.$$

For any $C > 0$,

$$C_N \le N \max_{1 \le i \le N} \left(\frac{\tilde{\omega}^{N,i} C}{\Omega^N} \right)^2 + N \sum_{i=1}^{N} \left(\frac{\tilde{\omega}^{N,i}}{\Omega^N} \right)^2 \cdot 1\left(w(X^{N,i}, \tilde{X}^{N,i}) > C \right)$$

where, because $\{(X^{N,i}, \omega^{N,i})\}_{i=1}^N$ is asymptotically normal for (ν, σ, ζ),

$$N \max_{1 \le i \le N} \left(\frac{\tilde{\omega}^{N,i}}{\Omega^N} \right)^2 \xrightarrow{P} 0;$$

and, from (37),

$$N \sum_{i=1}^{N} \left(\frac{\tilde{\omega}^{N,i}}{\Omega^N} \right)^2 \cdot 1\left(w(X^{N,i}, \tilde{X}^{N,i}) > C \right) \xrightarrow{P} \zeta \otimes R(w^2 \cdot 1\,(w > C)),$$

which can be made as small as possible by taking C large enough. □

5.2 Sequential Importance Sampling

According to (14), the filtering distribution (*the target distribution*) ϕ_t is given by, for all $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$,

$$\phi_0(f) = \frac{\Xi(g_0 f)}{\Xi(g_0)}, \tag{38}$$

$$\phi_t(f) = \frac{\phi_{t-1} Q_t(f)}{\phi_{t-1} Q_t(1)}, \quad t \ge 1, \tag{39}$$

where g_t and Q_t are as defined by (12) and (13).

Let $\{R_t, t \geq 1\}$ be a family of Markov kernels (the *proposal kernels*) on $(\mathsf{X}, \mathcal{X})$ and $R_0 \in \mathbb{M}_1(\mathcal{X})$ such that there exist weight functions $w_0 : \mathsf{X} \to \mathbb{R}_+$ and $w_t : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ satisfying [4] , for any $f \in \mathbb{F}_+(\mathsf{X}, \mathcal{X})$,

$$\Xi(f) = R_0(w_0 f), \tag{40}$$

$$Q_t f(x) = \int w_t(x, x') R_t(x, \mathrm{d}x') f(x'), \quad \text{for } t \geq 1. \tag{41}$$

Then, if we draw an i.i.d. sample $\{X_0^{N,i}\}_{i=1}^N$ from R_0, then $\{(X_0^{N,i}, w_0(X^{N,i}))\}_{i=1}^N$ will be a weighted sample consistent for Ξ, and

$$\{(X_0^{N,i}, g_0(X_0^{N,i}) w_0(X_0^{N,i}))\}_{i=1}^N \tag{42}$$

will be a weighted sample consistent for ϕ_0.

Assume that the weighted sample $\{(X_{t-1}^{N,i}, \omega_{t-1}^{N,i})\}_{i=1}^N$ is consistent for ϕ_{t-1}. We construct a weighted sample $\{(X_t^{N,i}, \omega_t^{N,i})\}_{i=1}^N$ as follows. In the proposal step, each particle $X_{t-1}^{N,i}$ gives a single offspring $X_t^{N,i}$, $i \in \{1, \ldots, N\}$, of which the distribution is specified by the proposal kernel $R_t(X_{t-1}^{N,i}, \cdot)$. From here we know that the Markov kernel R_t is some conditional distribution of X_t. This is different from Q_t, which denotes the joint distribution of X_t and Y_t given X_{t-1}.

Next, we assign to the new particle $X_t^{N,i}$ the importance weight

$$\omega_t^{N,i} = \omega_{t-1}^{N,i} w_t(X_{t-1}^{N,i}, X_t^{N,i}).$$

The first obvious choice is to set $R_t = M$, which is called the *prior kernel*. The weight function w_t then, according to (13), simplifies to

$$w_t(X_{t-1}^{N,i}, X_t^{N,i}) = g(X_t^{N,i}, Y_t), \tag{43}$$

which does not depend on $X_{t-1}^{N,i}$. This kernel is often convenient: sampling particles from M is often straightforward, and computing the incremental weight amounts to evaluating the conditional likelihood of the new observation given the current particle.

Algorithm 5.1 (Sequential Importance Sampling (SIS))**.**

[4] The equation below is different from what is one the book. The corresponding equation on the book is not compatible with the following algorithm 5.1.

Initial State Draw an i.i.d. sample $\{X_0^{N,i}\}_{i=1}^N$ from R_0, and set

$$\omega_0^{N,i} = g_0(X_0^{N,i}) w_0(X_0^{N,i}) \quad \text{for } i = 1, \ldots, N.$$

Recursion For $t = 1, 2, \ldots$,

- Draw $\{X_t^{N,i}\}_{i=1}^N$ conditionally independently from the distribution $R_t(X_{t-1}^{N,i}, \cdot)$.
- Compute the updated importance weights

$$\omega_t^{N,i} = \omega_{t-1}^{N,i} w_t(X_{t-1}^{N,i}, X_t^{N,i}), \quad i = 1, \ldots, N.$$

The *optimal proposal kernel* is given by

$$P_t^*(x, A) = \frac{Q_t(x, A)}{Q_t(x, \mathsf{X})}, \tag{44}$$

which may be interpreted as the conditional distribution of the hidden state X_t given X_{t-1} and the current observation Y_t. And the associated weight function

$$w_t(x, x') = Q_t(x, \mathsf{X}) \quad \text{for} (x, x') \in \mathsf{X}^2,$$

is the conditional likelihood of the observation Y_t given the previous state X_{t-1}. The optimal kernel was introduced in Zaritskii, Svetnik, and Shimelevich (1976) and Akashi and Kumamoto (1977). When the observation equation is linear, the optimal kernel is convenient.

Example 5.8 (ARCH(1) Observed in Additive Noise)**.**

$$X_t = \sigma_w(X_{t-1}) W_t, \quad W_t \sim \text{i.i.d. N}(0, 1),$$
$$Y_t = X_t + \sigma_v V_t, \quad V_t \sim \text{i.i.d. N}(0, 1),$$

with $\sigma_w^2(x) = \alpha_0 + \alpha_1 x^2$, where α_0 and α_1 are positive.

The prior kernel density

$$r_t(x, x') = \mathfrak{g}(x'; 0, \sigma_w^2(x)), \tag{45}$$

with weight function, according to Equation (43),

$$w_t(x, x') = \mathfrak{g}(y'; x', \sigma_v^2).$$

The optimal kernel density

$$p_t^*(x, x') = \mathfrak{g}\left(x'; \frac{\sigma_w^2(x) Y_t}{\sigma_w^2(x) + \sigma_v^2}, \frac{\sigma_w^2(x)\sigma_v^2}{\sigma_w^2(x) + \sigma_v^2}\right), \tag{46}$$

and the associated weight function

$$Q(x, \mathsf{X}) = \mathfrak{g}\left(y_t; 0, \sigma_w^2(x) + \sigma_v^2\right).$$

Proof.

$$f(x_t|x_{t-1}, y_t) = \frac{f(y_t|x_t) f(x_t|x_{t-1})}{f(y_t|x_{t-1})},$$

where

$$\begin{aligned} f(y_t|x_t) &\propto \exp\left\{-\frac{(y_t - x_t)^2}{2\sigma_v^2}\right\}, \\ f(x_t|x_{t-1}) &\propto \exp\left\{-\frac{x_t^2}{2\sigma_w^2(x_{t-1})}\right\}, \\ f(y_t|x_{t-1}) &\propto \exp\left\{-\frac{y_t^2}{2\left[\sigma_w^2(x_{t-1}) + \sigma_v^2\right]}\right\}. \end{aligned}$$

So

$$\begin{aligned} f(x_t|x_{t-1}, y_t) &\propto \exp\left\{-\frac{(y_t - x_t)^2}{2\sigma_v^2} - \frac{x_t^2}{2\sigma_w^2(x_{t-1})} - \frac{y_t^2}{2\left[\sigma_w^2(x_{t-1}) + \sigma_v^2\right]}\right\} \\ &= \exp\left\{-\frac{\left[x_t - \frac{\sigma_w^2(x) Y_t}{\sigma_w^2(x) + \sigma_v^2}\right]^2}{2\left[\frac{\sigma_w^2(x) Y_t}{\sigma_w^2(x) + \sigma_v^2}\right]}\right\}. \end{aligned}$$

□

Assume the initial distribution of X_0

$$\Xi = \mathrm{N}(0, \sigma_{w0}^2), \quad \text{with } \sigma_{w0}^2 = \frac{\alpha_0}{1 - \alpha_1},$$

because $\mathrm{Var}(X_t|X_{t-1}) = \alpha_0 + \alpha_1 X_{t-1}^2$. The optimal proposal distribution for Ξ

$$R_0^* = \mathrm{N}\left(\frac{\sigma_{w0}^2 Y_0}{\sigma_{w0}^2 + \sigma_v^2}, \frac{\sigma_{w0}^2 \sigma_v^2}{\sigma_{w0}^2 + \sigma_v^2}\right),$$

with the following weight function

$$\frac{\mathfrak{g}(x_0; 0, \sigma_{w0}^2)}{\mathfrak{g}\left(x_0; \sigma_{w0}^2 Y_0 \big/ (\sigma_{w0}^2 + \sigma_v^2), \sigma_{w0}^2 \sigma_v^2 \big/ (\sigma_{w0}^2 + \sigma_v^2)\right)}.$$

Therefore, according to Algorithm 5.1, a weighted sample consistent for the filtering distribution ϕ_0 can be constructed by sampling $\{X_0^i\}_{i=1}^N$ from R_0^* and setting weights $\omega_0^i = g_0(X_0^i)w_0(X_0^i)$.

When $\sigma_v^2 \gg \sigma_w^2$, the observation is non-informative, and the optimal kernel (46) is no better than the prior kernel (45). One the other hand, when $\sigma_v^2 \ll \sigma_w^2$, the optimal kernel (46)

$$p_t^*(x, x') \to \mathrm{N}(x'; Y_t, 0),$$

is markedly different from the prior kernel (45), and is expected to display better performance.

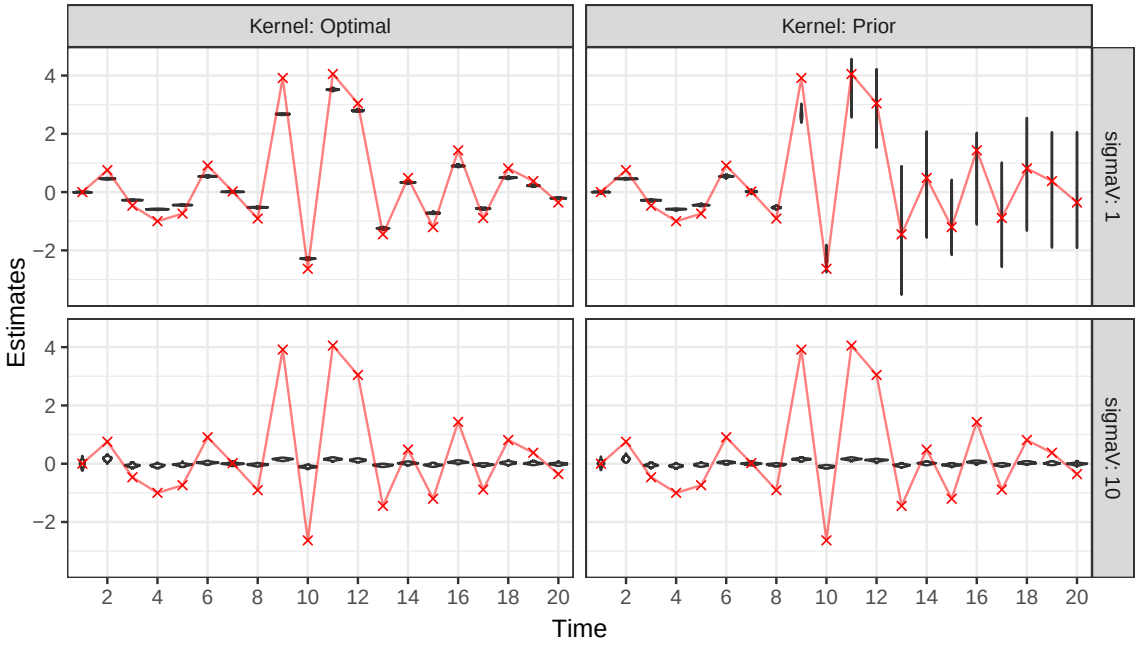

Figure 1: Boxplots of the posterior mean estimates of 100 independent SIS runs.

One advantage of the optimal kernel over the prior kernel is a higher precision, which is illustrated by Figure 1. SIS was run with 5000 particles for each of the two proposal kernels, for two distinct values of σ_v. The procedure was repeated independently 100 times, to obtain a sample of posterior mean estimates in all four cases. As exhibited in Figure 1, for informative observations, i.e. $\sigma_v = 1$, while the SIS estimates for both kernels are centered around a same value, the variance is much smaller when using the optimal

kernel.

The weight ω_t^i measures the adequacy of the particle X_t^i to the target distribution ϕ_t. A particle with small associated weight does not contribute to the estimator. If there are too many ineffective particles, the particle approximation is inefficient. Unfortunately, this situation is the rule rather than the exception. The importance weights will degenerate as the time index t *(not the sample size N)* increases, with most of the normalized importance weights ω_t^i/Ω_t^i close to 0 except for a few *(Most of them approach 0, even after being normalized? It does not make sense.)*.

Example 5.9 (Example 5.8, cont.)**.** Figure 2 displays the Lorenz curve of the normalized importance weights after 5, 10, 25 and 50 time steps for the ARCH(1) with $\sigma_v = 1$ for the prior kernel and the optimal kernel. The number of particles is set to 5000.

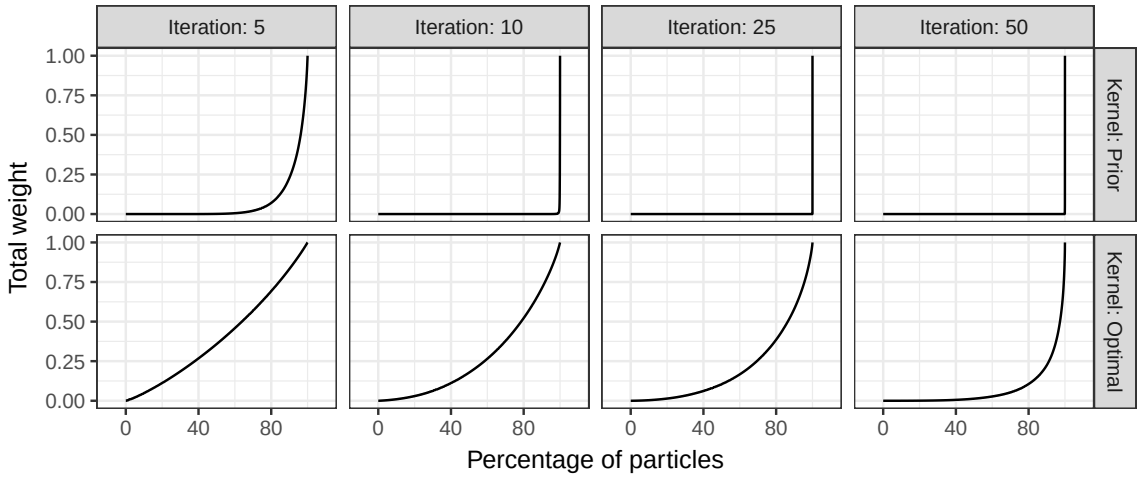

Figure 2: Lorenz Curves of the Normalized Importance Weights

The Lorentz curve show the total sum of the importance weights that the particles of the lowest $x\%$ possesses. Figure 2 shows that the normalized weights for the prior kernel quickly degenerates. For the optimal kernel, the degeneracy is much slower. This is the second advantage of the optimal kernel.

A simple criterion to quantify the degeneracy of a set of importance weights $\{\omega^i\}_{i=1}^N$ is

the *coefficient of variation* used by Kong, Liu, and Wong (1994), which is defined by

$$\mathrm{CV}\left(\{\omega^i\}_{i=1}^N\right) = \left[\frac{1}{N}\sum_{i=1}^{N}\left(N\frac{\omega^i}{\Omega^N} - 1\right)^2\right]^{\frac{1}{2}}. \tag{47}$$

The coefficient of variation is minimal, $\mathrm{CV} = 0$, when the normalized weights are all equal to $1/N$, and is maximized, $\mathrm{CV} = \sqrt{N-1}$, when one of the normalized weights is one and all others null.

A related criterion is the *effective sample size*, or ESS (Liu, 1996), defined as

$$\mathrm{ESS}\left(\{\omega\}_{i=1}^N\right) = \left[\sum_{i=1}^{N}\left(\frac{\omega^i}{\Omega^N}\right)^2\right]^{-1} = \frac{N}{1 + \left[\mathrm{CV}\left(\{\omega^i\}_{i=1}^N\right)\right]^2}. \tag{48}$$

Example 5.10 (Example 5.9, cont.)**.** Figure 3 shows the effective sample size curves of the importance weights with 100 time points. Note how the optimal kernel performs better than the prior kernel, but eventually degenerates.

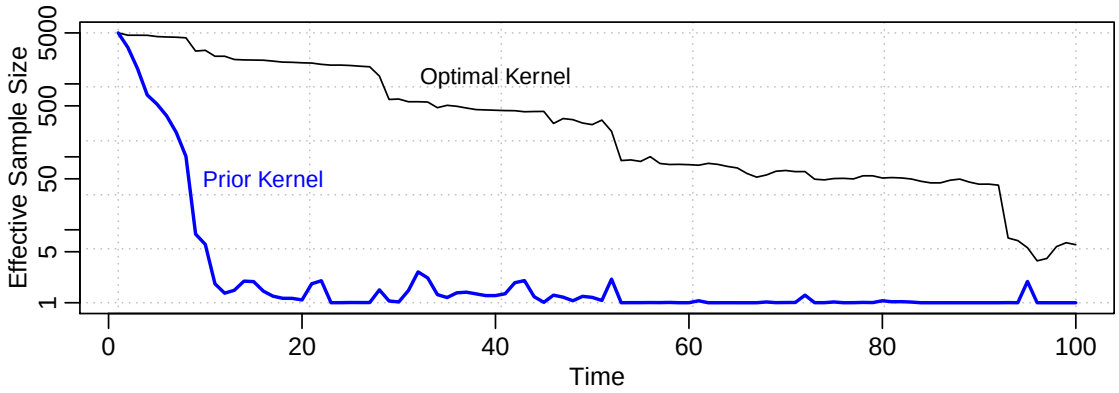

Figure 3: Effective Sample Size curves

5.3 Sampling Importance Resampling

To avoid the degeneracy of the importance weights, Gordon, Salmond, and Smith (1993) proposed a solution based on *resampling* using the normalized weights as probabilities of selection.

That resampling method is rooted in the *sampling importance resampling*, or SIR, method to sample a distribution μ, introduced by Rubin (1987, 1988).

In the setting of a single step importance estimator, the SIR process proceeds in two stages. In the *sampling stage*, an i.i.d. sample $\{X^i\}_{i=1}^N$ is drawn from the proposal distribution ν. The importance weights are then evaluated at particle positions, $\omega^i = w(X^i)$. In the *resampling* state, a sample of size N denoted by $\{\tilde{X}^i\}_{i=1}^N$ is drawn from the set of points $\{X^i\}_{i=1}^N$, with probability proportional to the weights ω^i. Doing so we obtain an equally weighted sample $\{\tilde{X}^i, 1\}_{i=1}^N$ also targeting μ.

Denote by M^i the number of times that the particle X^i is resampled. Then, $\{M^i\}_{i=1}^N$ is multinomial

$$M^1, \ldots, M^N \left| \{(X^i, \omega^i)\}_{i=1}^N \right. \sim \text{Mult}\left(N; \frac{\omega^1}{\Omega^N}, \ldots, \frac{\omega^N}{\Omega^N}\right), \tag{49}$$

and

$$\frac{1}{N}\sum_{i=1}^N f(\tilde{X}^i) = \sum_{i=1}^N \frac{M^i}{N} f(X^i).$$

Assume that the weighted sample $\{(X^i, \omega^i)\}_{i=1}^N$ is adapted to $\mathcal{F}^N$. A resampling procedure is said to be *unbiased* if

$$\mathrm{E}\left[M^i \middle| \mathcal{F}^N \right] = \frac{\omega^i}{\Omega^N} N, \quad i = 1, \ldots, N.$$

Unbiasedness implies that

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^N f(\tilde{X}^i) \middle| \mathcal{F}^N \right] = \sum_{i=1}^N \frac{\omega^i}{\Omega^N} f(X^i).$$

Therefore, the mean square error [5] of the SIR estimator is always larger than that of the importance sampling estimator:

$$\begin{aligned} &\mathrm{E}\left(\frac{1}{N}\sum_{i=1}^N f(\tilde{X}^i) - \mu(f) \right)^2 \\ &= \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^N f(\tilde{X}^i) - \sum_{i=1}^N \frac{\omega^i}{\Omega^N} f(X^i) \right)^2 + \mathrm{E}\left(\sum_{i=1}^N \frac{\omega^i}{\Omega^N} f(X^i) - \mu(f) \right)^2 . \end{aligned} \tag{50}$$

Theorem 5.11. *Assume that the weighted sample $\{(X^i, \omega^i)\}_{i=1}^N$ is adapted to $\mathcal{F}^N$ and is consistent for μ. Then, the uniformly weighted sample $\{(\tilde{X}^i, 1)\}_{i=1}^N$ obtained using (49) is consistent for μ.*

[5] *Error* or *deviation* is the difference between the estimator and what is estimated. *Mean square error* or *mean squared deviation* is the mean of the squares of the errors or deviations.

Theorem 5.12. *Assume that* $\{(X^i, \omega^i)\}_{i=1}^N$ *is adapted to* $\mathcal{F}^N$*, consistent for* μ*, and asymptotically normal for* (μ, σ, ζ)*. Then the uniformly weighted particle system* $\{(\tilde{X}^i, 1)\}_{i=1}^N$ *obtained using (*49*) is asymptotically normal for* $(\mu, \tilde{\sigma}, \tilde{\zeta})$ *with*

$$\tilde{\sigma}^2(f) = \mathrm{Var}_\nu(f) + \sigma^2(f),$$
$$\tilde{\zeta} = \mu.$$

5.4 Sequential Importance Sampling with Resampling

As shown in (50), the one-step effect of resampling seems to be negative, but resampling is required to guarantee that the particle approximation does not degenerate in the long run. This remark suggests that it may be advantageous to restrict the use of resampling to cases where the importance weights are unbalanced. The resulting algorithm is generally known under the name of *sequential importance sampling with resampling* (SISR):

Algorithm 5.2 (Sequential Importance Sampling with Resampling)**.**

Initial State Draw an i.i.d. sample $\{X_0^{N,i}\}_{i=1}^N$ from R_0, and set

$$\omega_0^{N,i} = g_0(X_0^{N,i}) w_0(X_0^{N,i}) \quad \text{for } i = 1, \ldots, N.$$

Recursion For $t = 1, 2, \ldots$,

- Draw $\{X_t^{N,i}\}_{i=1}^N$ conditionally independently from the distribution $R_t(X_{t-1}^{N,i}, \cdot)$.
- Compute the updated importance weights

$$\omega_t^{N,i} = \omega_{t-1}^{N,i} w_t(X_{t-1}^{N,i}, X_t^{N,i}), \quad i = 1, \ldots, N.$$

Resampling (Optional) Draw a multinomial trial $\{I_t^i\}_{i=1}^N$ with probabilities of success $\{\omega_t^i / \Omega_t^N\}_{i=1}^N$, and set $X_t^i = X_t^{N,I_t^i}$ and $\omega_t^i = 1$ for $i = 1, \ldots, N$.

As a special instance of Algorithm 5.2, the resampling procedure could be triggered when the coefficient of variation exceeds a threshold κ, i.e. $\mathrm{CV}(\{\omega_t^i\}_{i=1}^N) > \kappa$.

5.5 Auxiliary Particle Filter

The *auxiliary particle filter* (APF) has proven to be one of the most useful implementations of the particle filtering methodology. It was proposed by Pitt and Shephard (1999). The consistency and asymptotic normality of the auxiliary particle filter is discussed in Douc, Moulines, and Olsson (2009) and Johansen and Doucet (2008). Non-asymptotic bounds for the auxiliary particle filter is given in Douc, Garivier, Moulines, Olsson, et al. (2011).

5.5.1 Auxiliary Particle Filter

Assume that we at time $t-1$ has a weighted sample $\{X_{t-1}^i, \omega_{t-1}^i\}_{i=1}^N$ providing an approximation, ϕ_{t-1}^N, of the filtering distribution ϕ_{t-1}:

$$\phi_{t-1}^N = \sum_{i=1}^N \frac{\omega_{t-1}^i}{\Omega_{t-1}^N} \delta_{X_{t-1}^i}. \tag{51}$$

This is one example of how to approximate the probability density function of a distribution.

When the observation Y_t becomes available, an approximation of the filtering distribution ϕ_t may be obtained by plugging ϕ_{t-1}^N into the recursion (14), yielding, for $A \in \mathcal{X}$,

$$\begin{aligned}\phi_t^{N,\text{tar}}(A) &= \frac{\phi_{t-1}^N Q_t(A)}{\phi_{t-1}^N Q_t(\mathsf{X})} \\ &= \sum_{i=1}^N \frac{\omega_{t-1}^i Q_t(X_{t-1}^i, A)}{\sum_{j=1}^N \omega_{t-1}^j Q_t(X_{t-1}^j, \mathsf{X})}.\end{aligned}$$

One way to produce a weighted sample approximating $\phi^{N,\text{tar}}$ is to consider the *auxiliary target* distribution

$$\phi_t^{N,\text{aux}}(i \times A) \equiv \frac{\omega_{t-1}^i Q_t(X_{t-1}^i, A)}{\sum_{j=1}^N \omega_{t-1}^i Q_t(X_{t-1}^i, \mathsf{X})}, \quad i \in \{1, \ldots, N\},\ A \in \mathcal{X}.$$

By construction, the target distribution $\phi_t^{N,\text{tar}}$ is the auxiliary distribution marginalized with respect to the particle index i. Therefore, we may construct a weighted sample targeting $\phi_t^{N,\text{tar}}$ by sampling from the auxiliary distribution $\phi_t^{N,\text{aux}}$.

To sample from the auxiliary distribution $\phi_t^{N,\text{aux}}$, we use an importance sampling strategy. Note that $\phi_t^{N,\text{aux}}$ is absolutely continuous with respect to the following proposal distribution

$$\phi_t^{N,\text{prop}}(i \times A) = \frac{\omega_{t-1}^i \vartheta_t(X_{t-1}^i)}{\sum_{j=1}^N \omega_{t-1}^j \vartheta_t(X_{t-1}^j)} R_t(X_{t-1}^i, A), \quad i \in \{1, \ldots, N\},\ A \in \mathcal{X}, \tag{52}$$

where $\vartheta_t \equiv \vartheta(X_{t-1}, Y_t)$ is the *adjustment multiplier weight* function; R_t is the Markov kernel defined in (41). Therefore, we draw pairs $\{(I_t^i, X_t^i)\}_{i=1}^N$ from $\phi_t^{N,\text{prop}}$, and, for each draw, we compute the importance weight

$$\omega_t^i = \frac{w_t(X_{t-1}^{I_t^i}, X_t^i)}{\vartheta_t(X_{t-1}^{I_t^i})}, \tag{53}$$

where w_t is the importance function defined in (41). You might wonder how to draw the pair from the proposal? Consider the following example.

Example 5.13. I is a random variable taking value 1 with probability p and 2 with probability q. And $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$. We are interested in the joint distribution of (I, X_I).

Note that

$$\mathrm{P}(I = i, X_i \le x_i) = F(x_i) \cdot \mathrm{P}(I = i).$$

Note that I is independent of the value of X_I. So if we want to draw the pair from the joint distribution, we first draw I and then draw X_I depending on I.

I guess if we marginalize the joint distribution with respect to I, we would get a Gaussian mixture.

So to draw the pair (I_t, X_t) from $\phi_t^{N,\text{prop}}$, we first draw

$$I_t \sim \left\{ \frac{\omega_{t-1}^i \vartheta_t(X_{t-1}^i)}{\sum_{j=1}^N \omega_{t-1}^j \vartheta_t(X_{t-1}^j)} \right\}_{i=1}^N,$$

and then draw

$$X_t \sim R_t(X_{t-1}^{I_t}, \cdot).$$

Finally, the indexes $\{I_t^i\}_{i=1}^N$ are discarded, and the weighted sample $\{(X_t^i, \omega_t^i)\}_{i=1}^N$ is taken as an approximation of $\phi_t^{N,\text{tar}}$, and thus an approximation of ϕ_t.

About the choice of $R_t(x, \cdot)$ and $\vartheta_t(x)$, the simplest choice consists of setting, for all $x \in \mathsf{X}$, $\vartheta_t \equiv 1$ and $R_t(x, \cdot) = M(x, \cdot)$. In this case, $\omega_t^i = w_t(X_{t-1}^{I_t^i}, X_t^i) = g(X_t^i, y_t)$. As the proposal density is independent of y_t, the particles are drawn without any knowledge of

the observation. Low-importance particles can be drawn. Therefore, this algorithm can be inefficient, and, as resampling is applied at every iteration, this can result in rapid path degeneracy (Section 6.1).

As suggested by Pitt and Shephard (1999), a more appealing choice consists of setting $R_t(x, \cdot) = M(x, \cdot)$ and

$$\vartheta_t(x) = g\left(\int x' M(x, \mathrm{d}x'), Y_t\right).$$

This algorithm generates particles from the $t-1$ sample conditioned on the current observation. Therefore, the particles generated are more likely to be close to the true state (Arulampalam, Maskell, Gordon, & Clapp, 2002).

A more computationally costly choice consists of setting $\vartheta_t(x) = w_t^*(x) = Q_t(x, \mathsf{X})$ and $R_t(x, \cdot) = P_t^*(x, \cdot)$, where $P_t^*(x, \cdot)$ denotes the optimal proposal kernel (44) and $w_t^*(\cdot)$ denotes the weight function associated with $P_t^*(x, \cdot)$. In this case, $\omega_t^i = 1$ for $i = 1, \ldots, N$, and the auxiliary particle filter is said to be *fully adapted*. Except in some specific models, the implementation is computationally impractical.

The difference between the APF and the SISR before is that the APF uses, not $\{\omega\}_{i=1}^N$, but $\{\vartheta_t(X_{t-1}^i)\}_{i=1}^N$ to resample and preselect from the existing set of particles. The particles that are mostly likely to have offsprings are chosen to give offsprings.

5.5.2 Convergence of the Auxiliary Particle Filter

Assumption 5.14.

1. for all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $g(x, y) > 0$, and for all $0 \leq t \leq n$, $\sup_{0 \leq t \leq n} \|g_t\|_\infty < \infty$.

2. $\sup_{0 \leq t \leq n} \|\vartheta_t\|_\infty < \infty$ and $\sup_{0 \leq t \leq n} \sup_{(x,x') \in \mathsf{X}^2} w_t(x, x')/\vartheta_t(x) < \infty$.

Theorem 5.15. *Assume 5.14. Then, for all $t \in \{1, \ldots, n\}$, the weighted sample $\{(X_t^i, \omega_t^i)\}_{i=1}^N$ is consistent for ϕ_t and asymptotically normal for $(\phi_t, \sigma_t, \zeta_t)$ where, for any $f \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,*

$$\zeta_t(f) = \frac{1}{[\phi_{t-1} Q_t(1)]^2} \phi_{t-1}(\vartheta_t) \iint \phi_{t-1}(\mathrm{d}x) R_t(x, \mathrm{d}x') \frac{w_t^2(x, x')}{\vartheta_t(x)} f(x'),$$
$$\sigma_t^2(f) = \frac{1}{[\phi_{t-1} Q_t(1)]^2} \sigma_{t-1}^2(Q_t[f - \phi_t(f)]) + \zeta_t([f - \phi_t(f)]^2).$$

Corollary 5.16. *The asymptotic variance is minimized if the adjustment multiplier weight function is chosen to be*

$$\vartheta_t^{\mathrm{opt}}(x) = \left[\int R(x, \mathrm{d}x') w_t^2(x, x') \left[f(x') - \phi_t(f) \right]^2 \right]^{\frac{1}{2}}.$$

For a given sequence $\{R_t\}_{t=0}^n$ of proposal kernels, the use of the optimal adjustment weight function provides the most efficient of all auxiliary particle filters.

Nonasymptotic bounds for the particle approximation are given below.

Theorem 5.17. *Assume 5.14. For any $t \in \{0, \ldots, n\}$, there exist constants $0 < c_{1,t}, c_{2,t} < \infty$ such that, for all $N \in \mathbb{N}_+$, $\varepsilon > 0$, and $h \in \mathbb{F}_b(\mathsf{X}, \mathcal{X})$,*

$$\mathrm{p}\left[\left\{ X_t^i \right\}_{i=1}^N : \left| \phi_t^N(h) - \phi_t(h) \right| \geq \varepsilon \right] \leq c_{1,t} \exp\left\{ -\frac{c_{2,t} N \varepsilon^2}{\mathrm{osc}^2(h)} \right\},$$

where $\mathrm{osc}(\cdot)$ *is the oscillation seminorm function defined in Definition 1.38; ϕ_i^N is as defined in* (51)*,*

$$\phi_t^N(h) = \sum_{i=1}^N \frac{\omega_t^i}{\Omega_t^N} h(X_t^i),$$

and the weighted sample $\{X_t^i, \omega_t^i\}_{i=1}^N$ is defined in (53)*.*

The above result establishes the convergence, as the number of particles N tends to infinity, of the particle filter for any *finite* time horizon.

Furthermore, the convergence can be shown to be *uniform* in time under rather general conditions.

Assumption 5.18.

1. For all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, $g(x, y) > 0$ and $\sup_{(x,y) \in \mathsf{X} \times \mathsf{Y}} g(x, y) < \infty$.

2. $\sup_{t \geq 0} \sup_{x \in \mathsf{X}} \vartheta_t(x) < \infty$ and $\sup_{t \geq 0} \sup_{(x,x') \in \mathsf{X}^2} w_t(x, x') / \vartheta_t(x) < \infty$.

3. There exist constants $\sigma^+ > \sigma^- > 0$ and a probability measure ν on $(\mathsf{X}, \mathcal{X})$ such that for all $x \in \mathsf{X}$ and $A \in \mathcal{X}$,

$$\sigma^- \nu(A) \leq M(x, A) \leq \sigma^+ \nu(A).$$

4. There exist a constant $c_- > 0$ such that, $\Xi(g_0) > c_-$ and for all $t \geq 1$,

$$\inf_{x \in \mathsf{X}} Q_t 1(x) \geq c_-.$$

Theorem 5.19. *Assume 5.18. Then, the approximated filtering distribution satisfies a time-uniform exponential deviation inequality, i.e., there exist constants c_1 and c_2 such that, for all $N \in \mathbb{N}_+$, $t \geq 0$, all measurable functions $h \in \mathbb{F}(\mathsf{X}, \mathcal{X})$ and $\varepsilon > 0$,*

$$\mathrm{p}\left[\left\{X_t^i\right\}_{i=1}^N : \left|\phi_t^N(h) - \phi_t(h)\right| \geq \varepsilon\right] \leq c_1 \exp\left\{-\frac{c_2 N \varepsilon^2}{\mathrm{osc}^2(\mathrm{h})}\right\}.$$

5.6 Implicit Particle Filter

Based on Chorin, Morzfeld, and Tu (2010). The authors call their sampling "implicit" by analogy to the implicit schemes of solving differential equations, where the determination of a next value requires the solution of algebraic equations.

According to (39), the filtering distribution

$$\mathrm{p}(x_t|y_{1:t}) \propto \int \mathrm{p}(x_{t-1}|y_{1:t-1})\mathrm{p}(x_t|x_{t-1})\mathrm{p}(y_t|x_t)\mathrm{d}x_{t-1}.$$

Suppose $\mathrm{p}(x_{t-1}|y_{1:t-1})$ is approximated by the consistent weighted sample $\{(X_{t-1}^i, \omega_{t-1}^i)\}_{i=1}^N$. Then,

$$\int \mathrm{p}(x_{t-1}|y_{1:t-1})\mathrm{p}(x_t|x_{t-1})\mathrm{p}(y_t|x_t)\mathrm{d}x_{t-1} \approx \sum_{i=1}^N \omega_{t-1}^i \mathrm{p}(x_t|x_{t-1}^i)\mathrm{p}(y_t|x_t).$$

So, roughly speaking, when N is large,

$$\begin{aligned}\mathrm{p}(x_t|y_{1:t}) &\propto \sum_{i=1}^N \omega_{t-1}^i \mathrm{p}(x_t|X_{t-1}^i)\mathrm{p}(y_t|x_t) \\ &= \sum_{i=1}^N \omega_{t-1}^i \mathrm{p}(y_t|X_{t-1}^i)\frac{\mathrm{p}(x_t|X_{t-1}^i)\mathrm{p}(y_t|x_t)}{\mathrm{p}(y_t|X_{t-1}^i)}.\end{aligned}$$

which suggests that we could approximate $\mathrm{p}(x_t|y_{1:t})$ with the weighted sample

$$\{(X_t^i, \omega_{t-1}^i \mathrm{p}(y_t|X_{t-1}^i))\}_{i=1}^N$$

with X_t^i sampled from

$$X_t^i \sim \frac{\mathrm{p}(x_t|X_{t-1}^i)\mathrm{p}(y_t|x_t)}{\mathrm{p}(y_t|X_{t-1}^i)}, \quad i = 1, \ldots, N, \tag{54}$$

And, of course, direct sampling is usually infeasible for NLSS models unless the measurement equation is linear. That is why we resort to the importance sampling techniques as introduced in Section 5.1 and 5.2. We pick $\{X_t^i\}_{i=1}^N$ according to some proposal distribution, and then use weights to get the resulting pdf to agree with the target distribution. In general, many of the new positions will have low probabilities and therefore small weights.

So the question becomes how to sample highly probable particle positions. Note that, for each $i = 1, \ldots, N$, given X_{t-1}^i and y_t, we know both $\mathrm{p}(x_t|X_{t-1}^i)$ and $\mathrm{p}(y_t|x_t)$, and we can explicitly write

$$\mathrm{p}(x_t|X_{t-1}^i)\mathrm{p}(y_t|x_t) \propto \exp\left\{-F_t^i(x_t)\right\}. \tag{55}$$

Suppose X_t is an m-dimensional random variable. We sample φ from an m-dimensional standard normal distribution

$$\varphi \sim \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{\varphi^T\varphi}{2}\right\},$$

and solve the equation

$$F_t^i(x_t) - C_t^i = \frac{\varphi^T\varphi}{2}, \tag{56}$$

where C_t^i is a constant. The value of φ is most likely in the neighborhood of 0. Therefore, if the constant is chosen in a way such that $C_t^i = \min_{x_t} F_t^i(x_t)$, then Equation (56) ensures that the solution x_t is a high probability position for (54). *(But of course neither* $\min_{x_t} F_t^i$ *nor Equation (56) is easy to solve.)*

Suppose (56) is solved by $X_t^i(\varphi)$. Then, according to Property 1.6, each value x_t of X_t^i appears with probability density

$$\frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{\varphi^T(x_t)\varphi(x_t)}{2}\right\} |J|^{-1},$$

where J is the Jacobian of $X_t^i(\varphi)$, while the target density $\mathrm{p}(x_t|X_{t-1}^i)\mathrm{p}(y_t|x_t)$, from (55) and (56), is proportional to

$$\exp\left\{-C_t^i\right\} \exp\left\{-\frac{\varphi^T(x_t)\varphi(x_t)}{2}\right\}.$$

Therefore, the weight of the particle should be

$$\exp\left\{-C_t^i\right\} \cdot |J|(2\pi)^{m/2}.$$

(So we need to calculate the Jacobian J, too.)

Two algorithms are introduced below. The iteration solution Section 5.6.1 was introduced by Chorin and Tu (2009) and Chorin and Tu (2012). And the random map solution Section 5.6.2 was introduced by Morzfeld, Tu, Atkins, and Chorin (2012). I prefer the latter.

5.6.1 An Iteration Solution to Equation (56)

The Solution Note that (56) provides considerable latitude (freedom) in linking φ to x_t: it is a single equation of m unknown variables and can be solved by many solutions. One solution can be found sometimes via the following iteration, which converges whenever $F_t^i(\cdot)$ defined in (55) is convex in the authors' experience. But they have not been able to prove this (Chorin & Tu, 2012, P538).

Suppose that the state-space model is denoted by

$$X_t = f(X_{t-1}) + w_t, \quad w_t \sim \mathrm{N}(0, Q);$$
$$Y_t = h(X_t) + v_t, \quad v_t \sim \mathrm{N}(0, R).$$

Assume that the function $F_t^i(\cdot)$ defined in (55) is convex and $h(\cdot)$ is not too different from a linear function *(as is often the case, I guess)*.

The following part is revised according to Chorin and Tu (2012). Let $x_t^0 = 0$.

For $j \geq 1$, by Taylor-expanding the function $h(\cdot)$ to the first order around x_t^{j-1}, the measurement equation is approximated as a linear function of X_t,

$$Y_t \approx h(x_t^{j-1}) + H_t^{j-1}(X_t - x_t^{j-1}) + v_t, \tag{57}$$

where H_t^{j-1} denotes the Jacobian matrix evaluated at x_t^{j-1}. Denote

$$\tilde{Y}_t^j \equiv Y_t - h(x_t^{j-1}) + H_t^{j-1} x_t^{j-1}.$$

Then we have

$$\tilde{Y}_t^j \approx H_t^{j-1} X_t + v_t.$$

The function $F_t^i(\cdot)$ defined in (55) is approximated by a quadratic function around x_t^{j-1},

$$F_t^i(x_t) \approx \frac{1}{2}\left\{[x_t - f(X_{t-1}^i)]^T Q^{-1}[x_t - f(X_{t-1}^i)] + [\tilde{Y}_t^j - H^{j-1}x_t]^T R^{-1}[\tilde{Y}_t^j - H^{j-1}x_t]\right\} \tag{58}$$

$$= \frac{1}{2}\left(x_t - m_t^{ij}\right)^T \left(\Sigma_t^j\right)^{-1}\left(x_t - m_t^{ij}\right) + C_t^{ij}, \tag{59}$$

where

$$\Sigma_t^j = \left[Q^{-1} + \left(H^{j-1}\right)^T R^{-1} H^{j-1}\right]^{-1},$$

$$m_t^{ij} = \Sigma^j\left[Q^{-1} f(X_{t-1}^i) + \left(H^{j-1}\right)^T R^{-1}\tilde{Y}_t^j\right],$$

$$C_t^{ij} = \frac{1}{2}\left[\tilde{Y}_t^j - H^{j-1} f(X_{t-1}^i)\right]^T \left(K_t^j\right)^{-1}\left[\tilde{Y}_t^j - H^{j-1} f(X_{t-1}^i)\right],$$

with

$$K_t^j = H^{j-1} Q\left(H^{j-1}\right)^T + R.$$

Then x_t^j is found by solving the following equation,

$$\frac{1}{2}\left(x_t - m_t^{ij}\right)^T \left(\Sigma_t^j\right)^{-1}\left(x_t - m_t^{ij}\right) = \frac{\varphi^T\varphi}{2},$$

which, again, has multiple solutions. Since Σ_t^j is positive-definite, it has a unique Cholesky decomposition,

$$\Sigma_t^j = L_t^j\left(L_t^j\right)^T,$$

where L_t^j is a lower-triangular matrix with real and positive diagonal entries. So

$$\frac{1}{2}\left(x_t - m_t^{ij}\right)^T \left[L_t^j\left(L_t^j\right)^T\right]^{-1}\left(x_t - m_t^{ij}\right) = \frac{\varphi^T\varphi}{2}.$$

One solution can be found by letting

$$\left(L_t^j\right)^{-1}\left(x_t - m_t^{ij}\right) = \varphi.$$

(One questions could be whether L_t^j is always invertible.) That is,

$$x_t^j = m_t^{ij} + L_t^j\varphi.$$

In the authors' experience, whenever $F_t^i(\cdot)$ is convex, $x_t^j \to x_t$, the solution to (56). Intuitively we can illustrate the convergence with Figure 4.

In the limit, note that the linear approximate measurement equation (57) becomes the exact measurement equation. For that reason, (58) and (59) become exact, and C_t^{ij} becomes C_t^i.

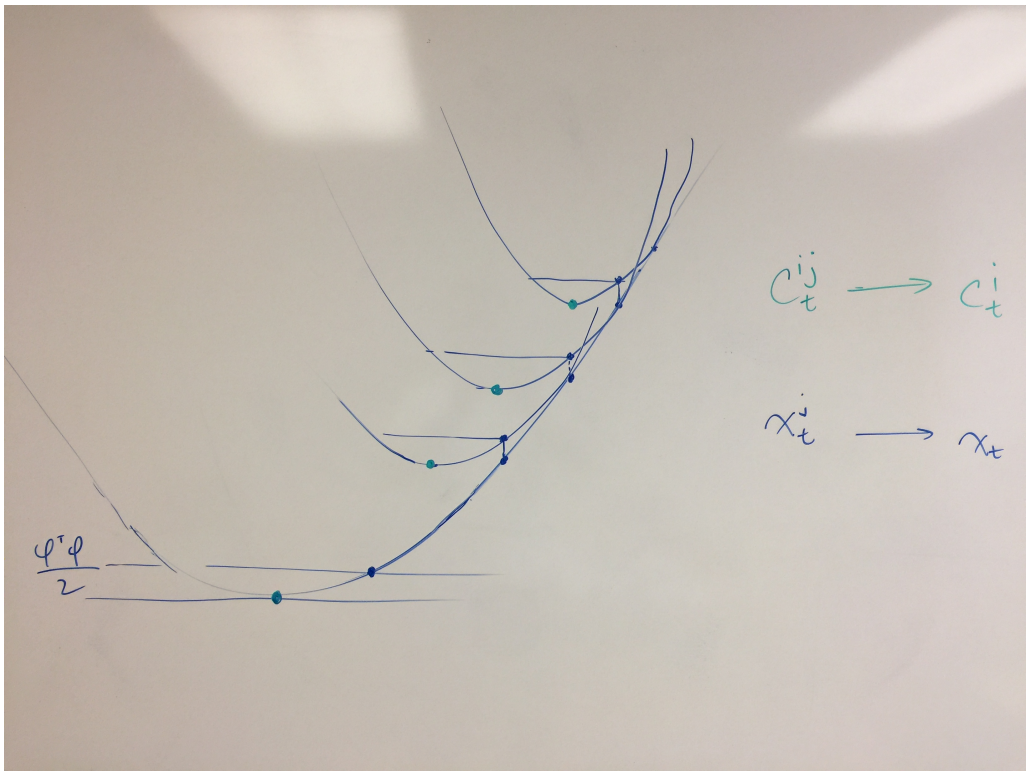

Figure 4: Convergence of x_t^j

The Jacobian Now we derive the Jacobian determinant. Denote

$$\lim_{j\to\infty} m_t^{ij} = m_t^i,$$
$$\lim_{j\to\infty} \Sigma_t^j = \Sigma_t.$$

Then in the limit according to (59),

$$F_t^i(x_t) - C_t^i = \frac{1}{2}\left(x_t - m_t^i\right)^T \Sigma_t^{-1}\left(x_t - m_t^i\right),$$

which, when combined with (56), shows that the Jacobian J can be solved from the following equation

$$\left(x_t - m_t^i\right)^T \Sigma_t^{-1}\left(x_t - m_t^i\right) = \varphi^T\varphi.$$

Note that for $i \in \{1, \ldots, m\}$,

$$\begin{aligned}\frac{\partial}{\partial x_{ti}}\left(x_t - m_t^i\right)^T \Sigma_t^{-1}\left(x_t - m_t^i\right) &= \frac{\partial}{\partial x_t}\left(x_t - m_t^i\right)^T \Sigma_t^{-1}\left(x_t - m_t^i\right)\cdot\frac{\partial x_t}{\partial x_{ti}} \\ &= 2\left(x_t - m_t^i\right)^T \Sigma_t^{-1} e_i,\end{aligned}$$

with e_i being a column vector such that $e_i^T = (0, \ldots, 0, 1, 0, \ldots, 0)$, and

$$\frac{\partial \varphi^T \varphi}{\partial \varphi} = 2\varphi^T.$$

So

$$\frac{\partial x_{ti}}{\partial \varphi} = \frac{\varphi^T}{\left(x_t - m_t^i\right)^T \Sigma_t^{-1} e_i},$$

and the Jacobian determinant

$$J = \det \begin{pmatrix} \frac{\partial x_{t1}}{\partial \varphi} \\ \vdots \\ \frac{\partial x_{tm}}{\partial \varphi} \end{pmatrix}.$$

5.6.2 A Random Map Solution

Refer to Morzfeld et al. (2012). In this section, we again assume that $F_t^i(\cdot)$ is convex.

Firstly, we need to solve the minimization problem, $\min_{x_t} F_t^i(x_t)$, using standard tools, e.g. Newton's method, a quasi-Newton method, or more sophisticated minimization strategies. Denote

$$m_t^i = \arg\min_{x_t} F_t^i(x_t),$$
$$C_t^i = F_t^i(m_t^i).$$

And if we use Newton's method, then in the neighborhood of m_t^i, we have

$$F_t^i(x_t) \approx C_t^i + \frac{1}{2}(x_t - m_t^i)^T H_t^i (x_t - m_t^i),$$

where H_t^i denotes the Hessian evaluated at the minimum. Cholesky decompose the Hessian $H_t^i = \left(U_t^i\right)^T U_t^i$.

Secondly, we solve (56). As we noted before, (56) has multiple solutions. One of them can be found by letting

$$x_t = m_t^i + \lambda_t^i U_t^i \varphi, \tag{60}$$

where U_t^i is some $m \times m$ matrix under our control and remains to be chosen. Substitute (60) into (56), and we get an algebraic equation with one single variable λ_t^i

$$F_t^i(m_t^i + \lambda_t^i U_t^i \varphi) - C_t^i = \frac{\varphi^T \varphi}{2}. \tag{61}$$

To initialize the numerical computation, choose $\lambda_t^{i0} = 1$. Intuitively, Equation (56) has a solution in various directions geometrically. We pick a direction, $U_t^i \varphi$, and determine how far we need to walk along the direction to reach the solution.

What remains to be done is to determine the Jacobian. From (60), we know

$$\begin{aligned}
\frac{\partial x_t}{\partial \varphi} &= \frac{\partial (m_t^i + \lambda_t^i U_t^i \varphi)}{\partial \phi} \\
&= \frac{\partial \lambda_t^i U_t^i \varphi}{\partial \phi} \\
&= \frac{\partial U_t^i \varphi \lambda_t^i}{\partial \phi} \\
&= U_t^i \frac{\partial \varphi \lambda_t^i}{\partial \phi} \\
&= U_t^i \left(\varphi \frac{\partial \lambda_t^i}{\partial \varphi} + \lambda_t^i I \right),
\end{aligned}$$

where $\partial \lambda_t^i / \partial \varphi$ is a row vector, which can be derived implicitly from (61) as follows:

$$\begin{aligned}
(\nabla F_t^i) \left(\lambda_t^i U_t^i + U_t^i \varphi \frac{\partial \lambda_t^i}{\partial \varphi} \right) &= \varphi^T, \\
\frac{\partial \lambda_t^i}{\partial \varphi} &= \frac{\varphi^T - \lambda_t^i (\nabla F_t^i) U_t^i}{(\nabla F_t^i) U_t^i \varphi}
\end{aligned}$$

with ∇F_t^i being a row vector denoting the gradient of F_t^i. So

$$\frac{\partial x_t}{\partial \varphi} = U_t^i \left[\frac{1}{(\nabla F_t^i) U_t^i \varphi} \varphi \left(\varphi^T - \lambda_t^i (\nabla F_t^i) U_t^i \right) + \lambda_t^i I \right].$$

Using Property 1.1 we calculate the Jacobian:

$$\begin{aligned}
J &= \det(U_t^i) \cdot \det \left(\frac{1}{(\nabla F_t^i) U_t^i \varphi} \varphi \left(\varphi^T - \lambda_t^i (\nabla F_t^i) U_t^i \right) + \lambda_t^i I \right) \\
&= \det(U_t^i) \cdot \det(\lambda_t^i I) \cdot \det \left(\frac{1}{(\nabla F_t^i) U_t^i \varphi} \left(\varphi^T - \lambda_t^i (\nabla F_t^i) U_t^i \right) \left(\lambda_t^i \right)^{-1} I \varphi + 1 \right) \\
&= \det(U_t^i) \cdot \det(\lambda_t^i I) \cdot \left[\frac{1}{(\nabla F_t^i) U_t^i \varphi} \left(\varphi^T - \lambda_t^i (\nabla F_t^i) U_t^i \right) \left(\lambda_t^i \right)^{-1} I \varphi + 1 \right] \\
&= \det(U_t^i) \cdot (\lambda_t^i)^{m-1} \cdot \frac{\varphi^T \varphi}{(\nabla F_t^i) U_t^i \varphi}.
\end{aligned}$$

6 Particle Smoothing

6.1 Poor Man's Particle Smoothing

Using the auxiliary particle filter, we generate the weighted samples $\{(X_t^i, \omega_t^i)\}_{i=1}^N$ targeting the filtering distribution ϕ_t.

I_t^i denotes the index of the ancestral particle from which the particle X_t^i originates. We call I_t^i the *one-step ancestor index* of the particle X_t^i.

We can construct the *ancestral path*, $X_{0:t}^i$, of X_t^i:

$$X_{0:t}^i = (X_0^{B_0^i}, \ldots, X_t^{B_t^i}),$$

where the *ancestor indexes* $B_{0:t}^i$ are given recursively by the one-step ancestor indexes

$$B_t^i = i,$$
$$B_s^i = I_{s+1}^{B_{s+1}^i}, \quad s = 0, \ldots, t-1.$$

Theorem 6.1. *Assume 5.14. Then, for all $t \in \{0, \ldots, n\}$, the weighted sample $\{(X_{0:t}^i, \omega_t^i)\}_{i=1}^N$ is consistent for the joint smoothing distribution $\phi_{0:t|t}$ and asymptotically normal for $(\phi_{0:t|t}, \sigma_t, \zeta_t)$ where for $f \in \mathbb{F}_b(\mathsf{X}^{t+1}, \mathcal{X}^{\otimes(t+1)})$,*

$$\zeta_f(f) = \frac{\phi_{t-1}(\vartheta_t)}{[\phi_{t-1}Q_t(1)]^2} \int \cdots \int \phi_{0:t|t}(\mathrm{d}x_0, \ldots, \mathrm{d}x_{t-1}) R_t(x_{t-1}, \mathrm{d}x_t) \frac{w_t^2(x_{t-1}, x_t)}{\vartheta_t(x_{t-1}) f(x_0, \ldots, x_t)},$$
$$\sigma_t^2(f) = \frac{\sigma_{t-1}^2(Q_t[f - \phi_{0:t|t}(f)])}{[\phi_{t-1}Q_t(1)]^2} + \zeta_t([f - \phi_{0:t|t}(f)]^2),$$

where

$$Q_t f(x_0, \ldots, x_{t-1}) = \int Q_t(x_{t-1}, \mathrm{d}x_t) f(x_0, \ldots, x_t).$$

Corollary 6.2. *For all $0 \le s \le t$, the weighted sample $\{(X_s^{B_s^i}, \omega_t^i)\}_{i=1}^N$ is consistent for the marginal smoothing distribution $\phi_{s|t}$.*

Despite these favorable statistical results, approximating the joint smoothing distribution with the ancestral path of $\{X_t^i\}_{i=1}^N$ is doomed to failure.

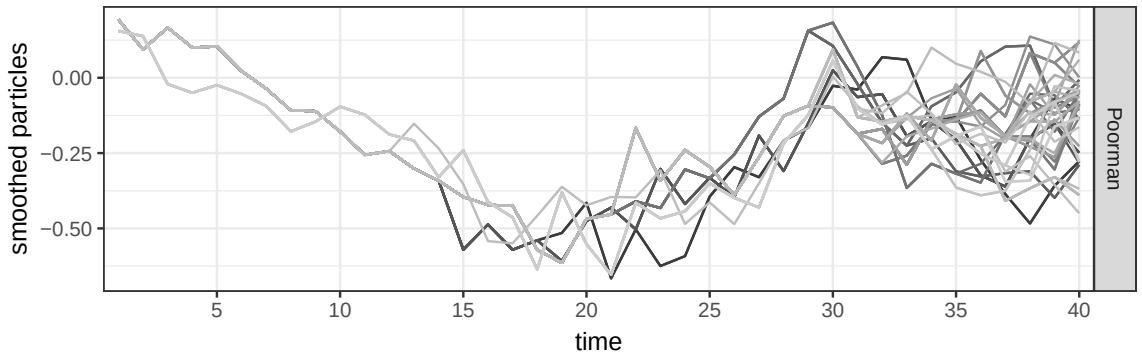

Figure 5: Sampled Trajectories from Poor Man's Smoother for 50 Particles

Example 6.3 (Noisy AR(1) - Poor Man's Smoother)**.** Consider a univariate state-space model where the observations are noisy,

$$Y_t = X_t + V_t,$$
$$X_t = \phi X_{t-1} + W_t,$$

where, for $t \in \mathbb{N}$, $|\phi| < 1$, $V_t \sim$ i.i.d. $\mathrm{N}(0,1)$, $W_t \sim$ i.i.d. $\mathrm{N}(0,1)$, $X_0 \sim \mathrm{N}(0, \sigma_w^2/(1-\phi^2))$; V_t, W_t and X_0 are independent.

Consider $n = 40$ simulated observations with $\phi = 0.9$. We use the poor man's smoother with $N = 50$ particles.

As illustrated by Figure 5, the path trajectories system collapse, a problem known as *path degeneracy*. At time 0, only 2 particles out of 50 survive, i.e. the marginal smoothing distribution $\phi_{0|n}$ is approximated with 2 particles. The estimator is therefore not reliable for sensible N values.

6.2 FFBSm

We have constructed a sequence of weighted samples $\{(X_t^i, \omega_t^i)\}_{i=1}^N$, $t = 1, \ldots, n$, approximating the filtering distributions ϕ_t, $t = 1, \ldots, n$.

For $t = 0, \ldots, n-1$, from (20), given $x' \in \mathsf{X}$, the backward kernel

$$B_{\phi_t}(x', A) = \frac{\phi_t(m(x, x')1(x \in A))}{\phi_t(m(x, x'))},$$

where, using the weighted sample $\{(X_t^i, \omega_t^i)\}_{i=1}^N$,

$$\phi_t(m(x, x')1(x \in A)) \approx \sum_{i=1}^N \frac{\omega_t^i}{\Omega_t^N} m(X_t^i, x')1(X_t^i \in A),$$
$$\phi_t(m(x, x')) \approx \sum_{i=1}^N \frac{\omega_t^i}{\Omega_t^N} m(X_t^i, x').$$

So we could approximate B_{ϕ_t} with

$$B_{\phi_t^N}(x', A) \equiv \frac{\sum_{i=1}^N \omega_t^i m(X_t^i, x')}{\sum_{i=1}^N \omega_t^i m(X_t^i, x')} 1(X_t^i \in A).$$

Therefore, given $x' \in \mathsf{X}$, the probability $B_{\phi_t}(x', \cdot)$ is approximated by the weighted sample $\{(X_t^i, \omega_t^i m(X_t^i, x'))\}_{i=1}^N$.

6.2.1 Joint Smoothing

According to (23),

$$\begin{aligned}
\phi_{n-1,n|n}(f) &= \iint f(x_{n-1}, x_n)\phi_n(\mathrm{d}x_n)B_{\phi_{n-1}}(x_n, \mathrm{d}x_{n-1}) \\
&= \int \phi_n(\mathrm{d}x_n) \int f(x_{n-1}, x_n)B_{\phi_{n-1}}(x_n, \mathrm{d}x_{n-1}) \\
&\approx \int \phi_n(\mathrm{d}x_n) \sum_{i=1}^N \frac{\omega_{n-1}^i m(X_{n-1}^i, x_n)}{\sum_i \omega_{n-1}^i m(X_{n-1}^i, x_n)} f(X_{n-1}^i, x_n) \\
&\approx \sum_{j=1}^N \frac{\omega_n^j}{\Omega_n^N} \sum_{i=1}^N \frac{\omega_{n-1}^i m(X_{n-1}^i, X_n^j)}{\sum_i \omega_{n-1}^i m(X_{n-1}^i, X_n^j)} f(X_{n-1}^i, X_n^j).
\end{aligned}$$

We denote the approximated smoothing distribution by $\phi_{n-1,n|n}^N$, i.e.

$$\phi_{n-1,n|n}^N(f) \equiv \sum_{j=1}^N \frac{\omega_n^j}{\Omega_n^N} \sum_{i=1}^N \frac{\omega_{n-1}^i m(X_{n-1}^i, X_n^j)}{\sum_i \omega_{n-1}^i m(X_{n-1}^i, X_n^j)} f(X_{n-1}^i, X_n^j). \tag{62}$$

Similarly, for $t = 0, \ldots, n-2$, the joint smoothing distribution $\phi_{t:n|n}$ could be approximated by

$$\phi_{t:n|n}^N(f) \equiv \sum_{i_t, i_{t+1}, \ldots, i_n} \frac{\omega_n^{i_n}}{\Omega_n^N} \prod_{s=t}^{n-1} \frac{\omega_s^{i_s} m(X_s^{i_s}, X_{s+1}^{i_{s+1}})}{\sum_{i_s} \omega_s^{i_s} m(X_s^{i_s}, X_{s+1}^{i_{s+1}})} f(X_t^{i_t}, X_{t+1}^{i_{t+1}}, \ldots, X_n^{i_n}).$$

The FFBSm keeps the filter particles $\{X_t^i\}$, $t = 0, \ldots, n$, unchanged. It updates the weights of these particles to target the smoothing distribution.

Evaluating the joint distribution is an $O(N^n)$ operation. It is not practically applicable on its own.

6.2.2 Marginal Smoothing

For $t = 0, \ldots, n-1$, assume we have available a particle system $\{(X_{t+1}^i, \omega_{t+1|n}^i)\}_{i=1}^N$ targeting $\phi_{t+1|n}$.

According to (22),

$$\begin{aligned}
\phi_{t|n}(f) &= \iint \phi_{t+1|n}(\mathrm{d}x_{t+1}) B_{\phi_{t|n}}(x_{t+1}, \mathrm{d}x_t) f(x_t) \\
&= \int \phi_{t+1|n}(\mathrm{d}x_{t+1}) \int B_{\phi_{t|n}}(x_{t+1}, \mathrm{d}x_t) f(x_t) \\
&= \int \phi_{t+1|n}(\mathrm{d}x_{t+1}) \int B_{\phi_t}(x_{t+1}, \mathrm{d}x_t) f(x_t) \\
&\approx \int \phi_{t+1|n}(\mathrm{d}x_{t+1}) \sum_{i=1}^N \frac{\omega_t^i m(X_t^i, x_{t+1})}{\sum_i \omega_t^i m(X_t^i, x_{t+1})} f(X_t^i) \\
&\approx \sum_{j=1}^N \omega_{t+1|n}^j \sum_{i=1}^N \frac{\omega_t^i m(X_t^i, X_{t+1}^j)}{\sum_i \omega_t^i m(X_t^i, X_{t+1}^j)} f(X_t^i).
\end{aligned}$$

So we could approximate $\phi_{t|n}$ by $\phi_{t|n}^N$, which is defined as

$$\phi_{t|n}^N \equiv \sum_{i=1}^N \omega_{t|n}^i f(X_t^i),$$

where

$$\omega_{t|n}^i \equiv \omega_t^i \sum_{j=1}^N \omega_{t+1|n}^j \frac{m(X_t^i, X_{t+1}^j)}{\sum_i \omega_t^i m(X_t^i, X_{t+1}^j)},$$

which are self-normalized, i.e. $\sum_i \omega_{t|n}^i = 1$.

Hence, we have obtained a particle system $\{(X_t^i, \omega_{t|n}^i)\}_{i=1}^N$ that targets $\phi_{t|n}$.

The complexity of this algorithm is $O(N^2)$ for one step, and $O(nN^2)$ for n steps altogether. The computational cost grows quadratically with the number of particles.

6.3 FFBSi

For $i \in \{0, \ldots, n-1\}$, consider the $N \times N$ Markov transition matrix, Λ_t^N, over the set of particle indexes, $\{1, \ldots, N\}$, whose (i, j) element is given by

$$\Lambda_t^N(i, j) = \frac{\omega_t^j m(X_t^j, X_{t+1}^i)}{\sum_j \omega_t^j m(X_t^j, X_{t+1}^i)}. \tag{63}$$

At time n, an index J_n is sampled in the set $\{1, \ldots, N\}$ with probability proportional to the weights $\{\omega_n^i\}_{i=1}^N$.

At time $t \leq n-1$, assuming that the indexes $J_{t+1:n}$ have already been sampled, J_t is sampled in $\{1, \ldots, N\}$ with probability $\{\Lambda_t^N(J_{t+1}, j)\}_{j=1}^N$.

Let $\mathcal{F}_n^N = \sigma(\{(X_t^i, \omega_t^i)\}_{i=1}^N, t \in \{0, \ldots, n\})$. Then the conditional joint distribution of $J_{0:n}$ is, therefore, given by

$$\mathrm{p}\left(J_{0:n} = j_{0:n} | \mathcal{F}_n^N\right) = \frac{\omega_n^{j_n}}{\Omega_n^N} \Lambda_{n-1}^N(j_n, j_{n-1}) \ldots \Lambda_0^N(j_1, j_0).$$

Thus, the FFBSm estimator (62) can be written as a conditional expectation with respect to $J_{0:n}$:

$$\phi_{0:n|n}^N(f) = \mathrm{E}\left[f(X_0^{J_0}, \ldots, X_n^{J_n}) | \mathcal{F}_n^N\right].$$

We may construct an unbiased *estimator*, $\tilde{\phi}_{0:n|n}^N$ of the FFBSm *estimator*, $\phi_{0:n}^N$, by drawing N paths of $\{J_{0:n}^i\}_{i=1}^N$ from (63):

$$\tilde{\phi}_{0:n|n}^N = N^{-1} \sum_{i=1}^N f\left(X_0^{J_0^i}, \ldots, X_n^{J_n^i}\right).$$

The computational complexity of sampling a single path of $J_{0:n}$ is $O(Nn)$ (N computations for each step because of (63)). The overall computational effort required to compute the FFBSi estimator $\tilde{\phi}_{0:n|n}^N$ is $O(N^2 n)$.

When the transition kernel m is bounded, i.e. $m(x, x') < \sigma_+$ for all $(x, x') \in \mathsf{X}^2$, the computational efficiency above can be significantly improved using the following acceptance-rejection procedure.

6.3.1 Accept-Reject Algorithm

We wish to sample the target density π, which is known up to a multiplicative constant. Let q be a proposal density. We assume that there is a constant $M < \infty$ such that $\pi(x)/q(x) < M$ for all $x \in \mathsf{X}$. The algorithm goes as follows:

1. We generate $Y \sim q$ and, independently, we generate $U \sim \mathrm{U}[0, 1]$.

2. If $Mq(Y)U \leq \pi(Y)$, then we set $X = Y$. If the inequality is not satisfied, we then discard Y and U and start again.

Then, X is distributed according to $\pi \Big/ \int \pi$.

Proof. The cumulative distribution function of X

$$\begin{aligned} F_X(x) &= \mathrm{p}(X \le x) \\ &= \mathrm{p}(Y \le x | Mq(Y)U \le \pi(Y)) \\ &= \frac{\mathrm{p}(Y \le x, Mq(Y)U \le \pi(Y))}{\mathrm{p}(Mq(Y)U \le \pi(Y))}. \end{aligned}$$

Here,

$$\begin{aligned} \mathrm{p}(Y \le x, Mq(Y)U \le \pi(Y)) &= \int_{-\infty}^{x} q(y) \int_0^{\frac{\pi(y)}{Mq(y)}} 1 \mathrm{d}u \mathrm{d}y \\ &= \int_{-\infty}^{x} \frac{\pi(y)}{M} \mathrm{d}y, \end{aligned}$$

and

$$\begin{aligned} \mathrm{p}(Mq(Y)U \le \pi(Y)) &= \int q(y) \int_0^{\frac{\pi(y)}{Mq(y)}} 1 \mathrm{d}u \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{\pi(y)}{M} \mathrm{d}y. \end{aligned}$$

So the density function

$$\begin{aligned} f_X(x) &= \frac{\mathrm{d}}{\mathrm{d}x} F_X(x) \\ &= \frac{\pi(x)}{\int \pi(y) \mathrm{d}y}. \end{aligned}$$

□

6.3.2 FFBSi Linear Algorithm

As in the standard FFBSi algorithm, we sample J_n multinomially with probability proportional to $\{\omega_n^i\}_{i=1}^N$.

At time $t \le n-1$, assuming that $J_{t+1:n}$ have already been sampled, in order to draw J_t, we propose I_t taking the value $\{1, \ldots, N\}$ with a probability proportional to $\{\omega_t^i\}_{i=1}^N$. And then an independent uniform random variable U_t is drawn on $[0, 1]$. Then we set $J_t = I_t$ if $U_t \sigma_+ \le m(X_t^{I_t}, X_{t+1}^{J_{t+1}})$; otherwise, we reject the proposed index and make another trial.

Algorithm 6.1 (FFBSi: Linear Smoothing Algorithm)**.**

1. Sample $J_n^1, \ldots, J_n^N$ multinomially with probabilities proportional to $\{\omega_n^i\}_{i=1}^N$.

2. To sample $J_s^1, J_s^2, \ldots, J_s^N$ for $s = n-1, \ldots, 0$,

for s from $n-1$ down to 0, ***do***

$L \leftarrow (1, \ldots, N)$

while L is not empty ***do***

$K \leftarrow \text{size}(L)$

sample $I_1, \ldots, I_K$ with multinomial probabilities proportional to $\{\omega_s^i\}_{i=1}^N$

sample $U_1, \ldots, U_K$ independently and uniformly from $[0, 1]$

$nL \leftarrow \emptyset$

for k from 1 to K ***do***

if $U_k \sigma_+ \leq m(X_s^{I_k}, X_{s+1}^{J_{s+1}^{L_k}})$ ***then***

$J_s^{L_k} = I^k$

else

$nL \leftarrow nL \cup \{L_k\}$

end if

end for

$L \leftarrow nL$

end while

end for

6.4 Particle Independent Metropolis-Hastings

An alternate approach that addresses the joint smoothing problem is to target $\phi_{0:n|n}$ with a Metropolis-Hastings sampler.

6.4.1 Proposal

Recall that the poor man's smoother generates a particle system $\{(X_{0:n}^i, \omega_n^i)\}_{i=1}^N$ targeting $\phi_{0:n|n}$, where $X_{0:n}^i$ is the ancestral path of X_n^i defined in Section 6.1. Hence,

Theorem 6.4. *if the random variable J is sampled from $\{1, \ldots, N\}$ with probability given by*

$$\mathrm{p}(J = j|\mathcal{F}_n^N) = \frac{\omega_n^j}{\Omega_n^N}, \quad j = 1, \ldots, n, \tag{64}$$

then the random variable $X_{0:n}^J$, the random ancestral path, is approximately distributed according to $\phi_{0:n|n}$.

Proof. For $A \in \mathcal{X}^{\otimes(n+1)}$,

$$\begin{aligned}\mathrm{p}(X_{0:n}^J \in A) &= \mathrm{E}\left[\mathrm{p}(X_{0:n}^J \in A|\mathcal{F}_n^N)\right] \\ &= \mathrm{E}\left[\sum_{j=1}^N \frac{\omega_n^j}{\Omega_n^N} 1(X_{0:n}^j \in A)\right],\end{aligned} \tag{65}$$

where the expectation is taken with respect to $\mathcal{F}_n^N$, or the particles and weights generated by the particle filtering process. According to Theorem 6.1, as $N \to \infty$,

$$\sum_{j=1}^N \frac{\omega_n^j}{\Omega_n^N} 1(X_{0:n}^j \in A) \xrightarrow{P} \int \phi_{0:n|n}(\mathrm{d}x_{0:n}) 1(x_{0:n} \in A).$$

Therefore,

$$\begin{aligned}\mathrm{p}(X_{0:n}^J \in A) &\xrightarrow{P} \mathrm{E}\left[\int \phi_{0:n|n}(\mathrm{d}x_{0:n}) 1(x_{0:n} \in A)\right] \\ &= \int \phi_{0:n|n}(\mathrm{d}x_{0:n}) 1(x_{0:n} \in A) \\ &= \phi_{0:n|n}(A).\end{aligned}$$

Therefore, when N is large, the ancestral path $X_{0:n}^J$ is approximately distributed according to $\phi_{0:n|n}$. □

The ancestral path $X_{0:n}^J$ can be used as a proposal for an independent Metropolis-Hastings sampler (introduced in Example 3.6). The proposal distribution, according to (65), is given by

$$\mathrm{p}(X_{0:n}^J \in A) = \mathrm{E}\left[\sum_{i=1}^N \frac{\omega_n^i}{\Omega_n^N} 1(X_{0:n}^i \in A)\right], \quad A \in \mathcal{X}^{\otimes(n+1)},$$

However, computing this expectation is not feasible in practice.

A solution to this problem lies in including those particles and weights as auxiliary variables and thus avoiding having to marginalize over them. Denote by

$$\boldsymbol{X}_t = (X_t^1, \ldots, X_t^N), \quad \boldsymbol{x}_t = (x_t^1, \ldots, x_t^N) \tag{66}$$

$$\boldsymbol{I}_t = (I_t^1, \ldots, I_t^N), \quad \boldsymbol{i}_t = (i_t^1, \ldots, i_t^N), \tag{67}$$

where I_t^i, $i = 1, \ldots, N$, is the one-step ancestor index of X_t^i defined in Section 6.1.

If the model under study is fully dominated, according to (52), the transition density with respect to the dominating measure λ is given by

$$\mathrm{p}_t(i_t, x_t) = \frac{\omega_{t-1}^{i_t} \vartheta_t(X_{t-1}^{i_t})}{\sum_{j=1}^N \omega_{t-1}^j \vartheta_t(X_{t-1}^j)} r_t(X_{t-1}^{i_t}, x_t), \quad t = 1, \ldots, n, \tag{68}$$

where, according to (53),

$$\omega_{t-1}^j = \frac{w_t(X_{t-2}^{i_{t-1}^j}, X_{t-1}^j)}{\vartheta_{t-1}(X_{t-2}^{i_{t-1}^j})}, \quad t = 2, \ldots, n; \tag{69}$$

with w_t defined in (41). Therefore, for $t \geq 2$, the process $(\boldsymbol{X}_t, \boldsymbol{I}_t)$ is a second-order nonhomogeneous Markov chain.

For $t = 0$, $\boldsymbol{X}_0$ is sampled from the initial proposal distribution R_0 defined in (40). The weights, according to (42),

$$\omega_0^j = g_0(X_0^j) w_0(X_0^j), \quad j = 1, \ldots, N. \tag{70}$$

And the ancestor index is not defined.

So the joint density of $\boldsymbol{X}_0$ and $\{(\boldsymbol{X}_t, \boldsymbol{I}_t)\}_{t=1}^n$ is as follows:

$$\psi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}) = \left(\prod_{\ell=1}^N r_0(x_0^\ell) \right) \left(\prod_{t=1}^n \prod_{\ell=1}^N \mathrm{p}_t(i_t^\ell, x_t^\ell) \right),$$

which is defined on the space $\mathsf{X}^{N(n+1)} \times \{1, \ldots, N\}^n$. And the joint density of $\{\boldsymbol{X}_{0:n}, \boldsymbol{I}_{1:n}, J\}$, which is a family of random variables defined on the space $\mathsf{X}^{N(n+1)} \times \{1, \ldots, N\}^{n+1}$, is given by

$$\psi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}) \frac{\omega_n^j}{\Omega_n^N}. \tag{71}$$

So now the situation is that we want to use a proposal distribution with density (71) to target the joint smoothing distribution $\phi_{0:n|n}$.

However, the target and the proposal distributions are not defined on the same space.

6.4.2 Target

The solution lies in extending the target distribution by defining an artificial target.

According to (17), the joint smoothing distribution $\phi_{0:n|n}$ has a density, also denoted by $\phi_{0:n|n}$, defined as

$$\phi_{0:n|n}(x_{0:n}) = (\mathrm{p}(Y_{0:n}))^{-1}\xi(x_0)g_0(x_0)\prod_{t=1}^{n} q_t(x_{t-1}, x_t), \tag{72}$$

where $\mathrm{p}(Y_{0:n})$ is the likelihood the observations $Y_{0:n}$. We factorize the joint density as follows

$$\phi_{0:n|n}(x_{0:n}) = \phi_{0|0}(x_0)\prod_{t=1}^{n}\frac{\phi_{0:t|t}(x_{0:t})}{\phi_{0:t-1|t-1}(x_{0:t-1})}, \tag{73}$$

where, according to (72),

$$\begin{aligned}\frac{\phi_{0:t|t}(x_{0:t})}{\phi_{0:t-1|t-1}(x_{0:t-1})} &= \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}q_t(x_{t-1}, x_t)\\ &= \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}w_t(x_{t-1}, x_t)r_t(x_{t-1}, x_t),\end{aligned}$$

of which the second equation is because of (41).

In Section 6.1, the ancestor indexes, $B^j_{0:n}$, of X^j_n are defined for $j = 1, \ldots, n$. For any $b^j_{0:n}$,

$$\frac{\phi_{0:t|t}(x^{b^j_{0:t}}_{0:t})}{\phi_{0:t-1|t-1}(x^{b^j_{0:t-1}}_{0:t-1})} = \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}w_t(x^{b^j_{t-1}}_{t-1}, x^{b^j_t}_t)r_t(x^{b^j_{t-1}}_{t-1}, x^{b^j_t}_t),$$

which, according to (69), equals

$$\begin{aligned}&\frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}\omega^{b^j_t}_t\vartheta_t(x^{b^j_{t-1}}_{t-1})r_t(x^{b^j_{t-1}}_{t-1}, x^{b^j_t}_t)\\ &= \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}\frac{\omega^{b^j_t}_t}{\omega^{b^j_{t-1}}_{t-1}}\left(\sum_{\ell=1}^{N}\omega^{\ell}_{t-1}\vartheta_t(x^{\ell}_{t-1})\right)\frac{\omega^{b^j_{t-1}}_{t-1}\vartheta_t(x^{b^j_{t-1}}_{t-1})}{\sum_{\ell=1}^{N}\omega^{\ell}_{t-1}\vartheta_t(x^{\ell}_{t-1})}r_t(x^{b^j_{t-1}}_{t-1}, x^{b^j_t}_t),\end{aligned}$$

and, according to (68), in turn equals

$$\begin{aligned}&\frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}\frac{\omega^{b^j_t}_t}{\omega^{b^j_{t-1}}_{t-1}}\left(\sum_{\ell=1}^{N}\omega^{\ell}_{t-1}\vartheta_t(x^{\ell}_{t-1})\right)\mathrm{p}_t(b^j_{t-1}, x^{b^j_t}_t)\\ &= \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})}\frac{\omega^{b^j_t}_t}{\omega^{b^j_{t-1}}_{t-1}}\left(\sum_{\ell=1}^{N}\omega^{\ell}_{t-1}\vartheta_t(x^{\ell}_{t-1})\right)\mathrm{p}_t(i^{b^j_t}_t, x^{b^j_t}_t),\end{aligned}$$

where the last equation is because $B_{t-1}^j = I_t^{B_t^j}$. In summary,

$$\frac{\phi_{0:t|t}(x_{0:t}^{b_{0:t}^j})}{\phi_{0:t-1|t-1}(x_{0:t-1}^{b_{0:t-1}^j})} = \frac{\mathrm{p}(Y_{0:t-1})}{\mathrm{p}(Y_{0:t})} \frac{\omega_t^{b_t^j}}{\omega_{t-1}^{b_{t-1}^j}} \left(\sum_{\ell=1}^N \omega_{t-1}^\ell \vartheta_t(x_{t-1}^\ell) \right) \mathrm{p}_t(i_t^{b_t^j}, x_t^{b_t^j}).$$

Therefore, from (73),

$$\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}) = \phi_{0|0}(x_0^{b_0^j}) \frac{\mathrm{p}(Y_0)}{\mathrm{p}(Y_{0:n})} \frac{\omega_n^j}{\omega_0^{b_0^j}} \left(\prod_{t=1}^n \sum_{\ell=1}^N \omega_{t-1}^\ell \vartheta_t(x_{t-1}^\ell) \right) \prod_{t=1}^n \mathrm{p}_t(i_t^{b_t^j}, x_t^{b_t^j})$$

where the filtering density, according to (38),

$$\begin{aligned} \phi_{0|0}(x_0^{b_0^j}) &= \frac{\xi(x_0^{b_0^j}) g_0(x_0^{b_0^j})}{\int \xi(x_0) g_0(x_0) \mathrm{d}x_0} \\ &= \frac{\xi(x_0^{b_0^j}) g_0(x_0^{b_0^j})}{\mathrm{p}(Y_0)}, \end{aligned}$$

and the weight

$$\omega_0^{b_0^j} = g_0(x_0^{b_0^j}) w_0(x_0^{b_0^j}).$$

And we know from (40),

$$\xi(x_0^{b_0^j}) = w_0(x_0^{b_0^j}) r_0(x_0^{b_0^j}),$$

therefore,

$$\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}) = \frac{\omega_n^j}{\mathrm{p}(Y_{0:n})} \left(\prod_{t=1}^n \sum_{\ell=1}^N \omega_{t-1}^\ell \vartheta_t(x_{t-1}^\ell) \right) r_0(x_0^{b_0^j}) \prod_{t=1}^n \mathrm{p}_t(i_t^{b_t^j}, x_t^{b_t^j}).$$

For any $\ell \in \{1, \ldots, N\}$, denote

$$\boldsymbol{x}_t^{-\ell} = (x_t^1, \ldots, x_t^{\ell-1}, x_t^{\ell+1}, \ldots, x_t^N),$$

and for any $\ell_{0:n} \in \{1, \ldots, N\}^{n+1}$, define

$$\boldsymbol{x}_{0:n}^{-\ell_{0:n}} = (\boldsymbol{x}_0^{-\ell_0}, \ldots, \boldsymbol{x}_n^{-\ell_n}).$$

Now define a density, π_n^N, on $\mathsf{X}^{N(n+1)} \times \{1, \ldots, N\}^{n+1}$. We factorize it before defining it

$$\begin{aligned} \pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j) &= \pi_n^N(x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) \pi_n^N(\boldsymbol{x}_{0:n}^{-b_{0:n}^j}, \boldsymbol{i}_{0:n}^{-b_{0:n}^j} \big| x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) \\ &= \pi_n^N(b_{0:n}^j) \pi_n^N(x_{0:n}^{b_{0:n}^j} \big| b_{0:n}^j) \pi_n^N(\boldsymbol{x}_{0:n}^{-b_{0:n}^j}, \boldsymbol{i}_{0:n}^{-b_{0:n}^j} \big| x_{0:n}^{b_{0:n}^j}, b_{0:n}^j). \end{aligned}$$

We define, by abuse of notation,

$$\pi_n^N(b_{0:n}^j) \equiv \frac{1}{N^{n+1}}, \tag{74}$$

$$\pi_n^N(x_{0:n}^{b_{0:n}^j} \big| b_{0:n}^j) \equiv \phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}), \tag{75}$$

$$\pi_n^N(\boldsymbol{x}_{0:n}^{-b_{0:n}^j}, \boldsymbol{i}_{1:n}^{-b_{0:n}^j} \big| x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) \equiv \left(\prod_{\ell \neq b_0^j, \ell=1}^{N} r_0(x_0^\ell) \right) \prod_{t=1}^{n} \left(\prod_{\ell \neq b_t^j, \ell=1}^{N} p_t(i_t^\ell, x_t^\ell) \right), \tag{76}$$

where p_t is as defined in (68). That is, the artificial density

$$\begin{aligned} \pi_n^N(\boldsymbol{x}_{0:n}^{-b_{0:n}^j}, \boldsymbol{i}_{1:n}^{-b_{1:n}^j}, x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) &= \pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}^{b_{1:n}^j}, b_{0:n-1}^j, j) \\ &= \pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}, j) \end{aligned}$$

is defined as

$$\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}, j) \equiv \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \left(\prod_{\ell \neq b_0^j, \ell=1}^{N} r_0(x_0^\ell) \right) \prod_{t=1}^{n} \left(\prod_{\ell \neq b_t^j, \ell=1}^{N} p_t(i_t^\ell, x_t^\ell) \right). \tag{77}$$

Theorem 6.5. *By construction, $X_{0:n}^{B_{0:n}^J}$ under π_n^N is distributed according to $\phi_{0:n|n}$.*

Proof. For any $A \in \mathcal{X}^{\otimes(n+1)}$, the probability under π_n^N

$$\mathrm{p}(X_{0:n}^{B_{0:n}^J} \in A) = \mathrm{E}\left[\pi_n^N(X_{0:n}^{B_{0:n}^J} \in A \big| B_{0:n}^J) \right],$$

where the expectation is taken with respect $B_{0:n}^J$, and, according to (75),

$$\pi_n^N(X_{0:n}^{B_{0:n}^J} \in A \big| B_{0:n}^J) = \phi_{0:n|n}(A).$$

So

$$\begin{aligned} \mathrm{p}(X_{0:n}^{B_{0:n}^J} \in A) &= \mathrm{E}\left[\phi_{0:n|n}(A) \right] \\ &= \phi_{0:n|n}(A). \end{aligned}$$

□

This has the important consequence that (77) can be used as a surrogate for the target distribution $\phi_{0:n|n}$.

6.4.3 PIMH

Consider now the IMH sampler with proposal density (71) and target density (77). The acceptance probability from a proposed move from $\{\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j\}$ to $\{\tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j}\}$ is given by

$$\alpha(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j; \tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j}) = \min\left\{\frac{\pi_n^N(\tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j})}{\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j)} \cdot \frac{\psi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})\omega_n^j \tilde{\Omega}_n^N}{\psi_n^N(\tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j})\tilde{\omega}^{\tilde{j}}\Omega_n^N}, 1\right\}.$$

From (71) and (77) we know that for any $\{\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j\}$,

$$\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j) = \frac{\Omega_n^N}{N^{n+1}\mathrm{p}(Y_{0:n})}\left[\prod_{t=1}^{n}\sum_{\ell=1}^{N}\omega_{t-1}^{\ell}\vartheta_t(x_{t-1}^{\ell})\right]\left[\psi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n})\frac{\omega_n^j}{\Omega_n^N}\right].$$

Therefore,

$$\alpha(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j; \tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j}) = \min\left\{\frac{\tilde{\Omega}_n^N \prod_{t=1}^{n}\sum_{\ell=1}^{N}\tilde{\omega}_{t-1}^{\ell}\vartheta_t(\tilde{x}_{t-1}^{\ell})}{\Omega_n^N \prod_{t=1}^{n}\sum_{\ell=1}^{N}\omega_{t-1}^{\ell}\vartheta_t(x_{t-1}^{\ell})}, 1\right\}.$$

Denote

$$\hat{L}_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) = \left(\frac{1}{N}\sum_{\ell=1}^{N}\omega_n^{\ell}\right)\left[\prod_{t=1}^{n}\frac{1}{N}\sum_{\ell=1}^{N}\omega_{t-1}^{\ell}\vartheta_t(x_{t-1}^{\ell})\right]. \tag{78}$$

On the textbook (Douc et al., 2014, P375), this $\hat{L}_n^N$ is said to be a consistent estimator for the likelihood. I do not quite understand. My guess is that $N^{n+1}\hat{L}_n^N$ is a consistent estimator for the likelihood. But again, not sure. Then,

$$\alpha(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, j; \tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n}, \tilde{j}) = \min\left\{\frac{\hat{L}_n^N(\tilde{\boldsymbol{x}}_{0:n}, \tilde{\boldsymbol{i}}_{0:n})}{\hat{L}_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})}, 1\right\}. \tag{79}$$

Algorithm 6.2. [Particle Independent Metropolis-Hastings]

Initialize:

- Run an APF targeting $\phi_{0|0} \ldots \phi_{n|n}$ and compute $\ell_n^N[0] = \hat{L}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.
- Sample J with $\mathrm{p}(J = j|\mathcal{F}_n^N) = \omega_n^j/\Omega_n^N$, and set

$$X_{0:n}[0] = X_{0:n}^J.$$

Iterate for $k \geq 1$:

- Run an APF targeting $\phi_{0|0} \ldots \phi_{n|n}$ and compute $\tilde{\ell}_n^N = \hat{L}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.

- Sample J with $\mathrm{p}(J = j|\mathcal{F}_n^N) = \omega_n^j/\Omega_n^N$.
- Draw U uniformly on $[0, 1]$.
- If $U \leq \min\{1, \tilde{\ell}_n^N/\ell_n^N[k-1]\}$, set

$$X_{0:n}[k] = X_{0:n}^J,$$
$$\ell_n^N[k] = \tilde{\ell}_n^N.$$

  Otherwise set

$$X_{0:n}[k] = X_{0:n}[k-1],$$
$$\ell_n^N[k] = \ell_n^N[k-1].$$

Let $f \in \mathbb{F}_b(\mathsf{X}^{n+1}, \mathcal{X}^{\otimes(n+1)})$ and assume that we run Algorithm 6.2 for K iterations (possibly with some burn-in). PIMH then provides the following estimator of $\phi_{0:n|n}(f)$,

$$\hat{\phi}_{0:n|n}^{\mathrm{PIMH}}(f) = \frac{1}{K}\sum_{k=1}^{K} f(X_{0:n}[k]). \tag{80}$$

It seems wasteful to generate N particle trajectories at each iteration of the PIMH sampler, but keep only a single one.

Note that the acceptance probability α in (79) does not depend on J. That is, the PIMH inference about the target distribution does not depend on J. Let

$$\mathcal{F}_n^N[k] = \sigma\{\{U[m], (\boldsymbol{X}_{0:n}[m], \boldsymbol{I}_{0:n}[m])\}, m \leq k\}.$$

Then the PIMH inference about the target distribution depends on $(\mathcal{F}_n^N[K], \{J[k]\}_{k=1}^K)$ only through $\mathcal{F}_n^N[K]$. In other words, according to Definition 1.10, $\mathcal{F}_n^N[K]$ is a sufficient statistic for the target distribution.

We then, according to Theorem 1.13 and Definition 1.14, compute the Rao-Blackwell

estimator,

$$\begin{aligned}\hat{\phi}_{0:n|n}^{\text{PIMH-RB}}(f) &= \mathrm{E}\left[\hat{\phi}_{0:n|n}^{\text{PIMH}}\middle|\mathcal{F}_n^N[K]\right] \\ &= \frac{1}{K}\sum_{k=1}^{K}\mathrm{E}\left[f(X_{0:N}[k])|\mathcal{F}_n^N[K]\right] \\ &= \frac{1}{K}\sum_{k=1}^{K}\mathrm{E}\left[f(X_{0:N}[k])|\mathcal{F}_n^N[k]\right] \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{N}\frac{\omega_n^j[k]}{\Omega_n^N[k]}f(X_{0:n}^j[k]). \end{aligned} \tag{81}$$

The possibility to make use of all the generated particles to reduce the variance seems promising. However, a problem with the Rao-Blackwell estimator is that the particle systems $\{(X_{0:n}^j[k], \omega_{0:n}^j[k])\}_{j=1}^N$ suffer from *path* degeneracy. Hence, the possible benefit of Rao-Blackwellization is limited.

The problem lies in the fact that PIMH relies on the poor man's smoother as its basic building block.

To obtain a larger variance reduction, we can use the same idea to Metropolise a more advanced particle filter, such as the FFBSi. Olsson and Ryden (2011) have proposed to modify the PIMH by replacing (64) by the run of a backward simulator (63). Interestingly, the acceptance probability for such an algorithm is the same as (79).

Algorithm 6.3 (Particle Independent Metropolis-Hastings - FFBSi)**.**
Initialize:

- Run an APF targeting $\phi_{0|0}, \cdots, \phi_{n|n}$ and compute $\ell_n^N[0] = \hat{L}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.
- Sample $(J_0, \ldots, J_n)$ according to (63), and set

$$X_{0:n}[0] = (X_0^{J_0}, \ldots, X_n^{J_n}).$$

Iterate for $k \geq 1$**:**

- Run an APF targeting $\phi_{0|0}, \ldots, \phi_{n|n}$ and compute $\tilde{\ell}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.
- Draw U uniformly on $[0:1]$.

- If $U \leq \min\{1, \tilde{\ell}_n^N / \ell_n^N[k-1]\}$, sample $(J_0, \ldots, J_n)$ according to (63), and set

$$X_{0:n}[k] = (X_0^{J_0}, \ldots, X_n^{J_n});$$
$$\ell_n^N[k] = \tilde{\ell}_n^N.$$

  Otherwise, set

$$X_{0:n}[k] = X_{0:n}[k-1],$$
$$\ell_n^N[k] = \ell_n^N[k-1].$$

Again, by a Rao-Blackwellization type of argument, denote

$$\mathcal{F}_n^{N,\mathrm{BSi}}[k] = \sigma\{(U[m], \boldsymbol{X}_{0:n}[m]), m \leq k\}.$$

Then, the FFBSi-based Rao-Blackwell estimator, for any $f \in \mathbb{F}_b(\mathsf{X}^{n+1}, \mathcal{X}^{\otimes(n+1)})$,

$$\begin{aligned}\hat{\phi}_{0:n|n}^{\text{PIMH - BSi - RB}}(f) &= \mathrm{E}\left[\frac{1}{K}\sum_{k=1}^{K} f(X_{0:n}[k]) \middle| \mathcal{F}_n^{N,\mathrm{BSi}}[K]\right] \\ &= \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}\left[f(X_{0:n}[k]) \middle| \mathcal{F}_n^{N,\mathrm{BSi}}[k]\right]\end{aligned}$$

where, at iteration k of the PIMH sampler, the expectation is taken with respect to $J_{0:n}$. By drawing M paths from (63) at iteration k, the expectation could be estimated unbiasedly. Therefore,

$$\hat{\phi}_{0:n|n}^{\text{PIMH - BSi - RB}}(f) \approx \frac{1}{KM}\sum_{k=1}^{K}\sum_{m=1}^{M} f(X_{0:n}^m[k]).$$

Algorithm 6.4 (Particle Independent Metropolis-Hastings - FFBSi - RB)**.**
Initialize:

- Run an APF targeting $\phi_{0|0}, \cdots, \phi_{n|n}$ and compute $\ell_n^N[0] = \hat{L}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.

- For $1 \leq m \leq M$, sample $(J_0^m, \ldots, J_n^m)$ according to (63), and set

$$X_{0:n}^m[0] = (X_0^{J_0^m}, X_1^{J_1^m}, \ldots, X_n^{J_n^m}).$$

Iterate for $k \geq 1$**:**

- Run an APF targeting $\phi_{0|0}, \ldots, \phi_{n|n}$ and compute $\tilde{\ell}_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n})$.
- Draw U uniformly on $[0:1]$.
- If $U \leq \min\{1, \tilde{\ell}_n^N / \ell_n^N[k-1]\}$, for $1 \leq m \leq M$, sample $(J_0^m, \ldots, J_n^m)$ according to (63), and set

$$X_{0:n}^m[k] = (X_0^{J_0^m}, \ldots, X_n^{J_n^m});$$
$$\ell_n^N[k] = \tilde{\ell}_n^N.$$

  Otherwise, set

$$X_{0:n}^m[k] = X_{0:n}^m[k-1],$$
$$\ell_n^N[k] = \ell_n^N[k-1].$$

6.5 Particle Gibbs Sampling

6.5.1 Particle Gibbs

Introduced in Andrieu, Doucet, and Holenstein (2010).

It is possible to construct a Gibbs sampler for the extended target distribution π_n^N defined in (77),

$$\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}, j) \equiv \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \left(\prod_{\ell \neq b_0^j, \ell=1}^{N} r_0(x_0^\ell) \right) \prod_{t=1}^{n} \left(\prod_{\ell \neq b_t^j, \ell=1}^{N} p_t(i_t^\ell, x_t^\ell) \right),$$

where the bold symbols are as defined in (66) and (67). Note that,

$$\begin{aligned} \{\boldsymbol{I}_{1:n}, J\} &= \{\boldsymbol{I}_{1:n}^{-B_{0:n}^j}, I_{1:n}^{B_{0:n}^J}, J\} \\ &= \{\boldsymbol{I}_{1:n}^{-B_{0:n}^j}, B_{0:n-1}^J, J\} \\ &= \{\boldsymbol{I}_{1:n}^{-B_{0:n}^j}, B_{0:n}^J\}, \end{aligned}$$

and

$$\boldsymbol{X}_{0:n} = \{\boldsymbol{X}_{0:n}^{-B_{0:n}^J}, X_{0:n}^{B_{0:n}^J}, \}.$$

So π_n^N can be viewed as a distribution of compound variables $\{\boldsymbol{X}_{0:n}^{-B_{0:n}^J}, \boldsymbol{I}_{0:n}^{-B_{0:n}^J}, X_{0:n}^{B_{0:n}^J}, B_{0:n}^J\}$, or equivalently a distribution of $\{\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n}, X_{0:n}^{B_{0:n}^J}, B_{0:n}^J\}$, where compound variables $\{X_{0:n}^{B_{0:n}^J}, B_{0:n}^J\}$ are viewed as new simple variables,

$$\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}, x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) = \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \frac{\left(\prod_{\ell=1}^N r_0(x_0^\ell)\right) \prod_{t=1}^n \left(\prod_{\ell=1}^N p_t(i_t^\ell, x_t^\ell)\right)}{r_0(x_0^{b_0^j}) \prod_{t=1}^n \mathrm{p}_t(i_t^{b_t^j}, x_t^{b_t^j})}.$$

The Gibbs sampler to sample from $\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}, x_{0:n}^{b_{0:n}^j}, b_{0:n}^j)$ consists of sampling iteratively from $\pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}|x_{0:n}^{b_{0:n}^j}, b_{0:n}^j)$ and $\pi_n^N(x_{0:n}^{b_{0:n}^j}, b_{0:n}^j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})$.

To sample $\{\boldsymbol{X}_{0:n}, \boldsymbol{I}_{0:n}\}$ from $\pi_n^N(\cdot, \cdot|x_{0:n}^{b_{0:n}^j}, b_{0:n}^j)$,

Algorithm 6.5. [Particle Gibbs Part 1]

For $t = 0$**:**

- For $\ell \neq b_0^j$, draw $X_0^\ell \sim r_0(\cdot)$;
- set $X_0^{b_0^j} = x_0^{b_0^j}$.
- For $\ell = 1, \ldots, N$, compute $\omega_0^\ell = g_0(X_0^\ell) w_0(X_0^\ell)$.

For $t = 1, \ldots, n$**:** Given $\{\boldsymbol{X}_{t-1}, \boldsymbol{\omega}_{t-1}\}$,

- for $\ell \neq b_t^j$, draw $(I_t^\ell, X_t^\ell) \sim \mathrm{p}_t(\cdot, \cdot)$, which is the transition density given by (68);
- set $X_t^{b_t^j} = x_t^{b_t^j}$. Set $I_t^{b_t^j} = b_{t-1}^j$.
- For $\ell = 1, \ldots, N$, compute $\omega_t^\ell = w_t(X_{t-1}^{I_t^\ell}, X_t^\ell)/\vartheta_t(X_{t-1}^{I_t^\ell})$.

To sample $\{X_{0:n}^{B_{0:n}^J}, B_{0:n}^J\}$ from $\pi_n^N(\cdot, \cdot|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})$, note that

$$\begin{aligned}
\pi_n^N(x_{0:n}^{b_{0:n}^j}, b_{0:n}^j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) &= \pi_n^N(x_{0:n}^{b_{0:n}^j}, b_{0:n-1}^j, j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \pi_n^N(J = j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \\
&= \pi_n^N(x_{0:n}^{b_{0:n}^j}, b_{0:n-1}^j|j, \boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \pi_n^N(J = j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \\
&= \pi_n^N(J = j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})
\end{aligned}$$

Property 6.6. *For* $j = 1, \ldots, N$,

$$\pi(j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}) = \frac{\omega_n^j}{\Omega_n^N}.$$

Proof.

$$\begin{aligned}\pi_n^N(j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}) &\propto \pi_n^N(\boldsymbol{x}_{0:n}, \boldsymbol{i}_{1:n}, j)\\ &= \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}}\left(\prod_{\ell\neq b_0^j,\ell=1}^{N} r_0(x_0^\ell)\right)\prod_{s=1}^{n}\left(\prod_{\ell\neq b_s^j,\ell=1}^{N} \mathrm{p}_s(i_s^\ell, x_s^\ell)\right)\\ &= \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}}\frac{\left(\prod_{\ell=1}^{N} r_0(x_0^\ell)\right)\prod_{s=1}^{n}\left(\prod_{\ell=1}^{N} \mathrm{p}_s(i_s^\ell, x_s^\ell)\right)}{r_0(x_0^{b_0^j})\prod_{s=1}^{n}\mathrm{p}_s(i_s^{b_s^j}, x_s^{b_s^j})}.\end{aligned}$$

Again, given $\{\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}\}$, the value of $\left(\prod_{\ell=1}^{N} r_0(x_0^\ell)\right)\prod_{s=1}^{n}\left(\prod_{\ell=1}^{N} \mathrm{p}_s(i_s^\ell, x_s^\ell)\right)$ is known. Hence,

$$\pi_n^N(j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \propto \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{r_0(x_0^{b_0^j})\prod_{s=1}^{n}\mathrm{p}_s(i_s^{b_s^j}, x_s^{b_s^j})}, \tag{82}$$

where, according to (72),

$$\begin{aligned}\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}) &\propto \xi(x_0^{b_0^j})g_0(x_0^{b_0^j})\prod_{s=1}^{n} q_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})\\ &= w_0(x_0^{b_0^j})r_0(x_0^{b_0^j})g_0(x_0^{b_0^j})\prod_{s=1}^{n} w_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})r_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j}).\end{aligned}$$

where the last equality follows from (40) and (41). Then, according to (70), (69), and (68),

$$\begin{aligned}&\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})\\ &\propto \omega_0^{b_0^j} r_0(x_0^{b_0^j})\prod_{s=1}^{n}\frac{w_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})}{\vartheta_s(x_{s-1}^{b_{s-1}^j})}\vartheta_s(x_{s-1}^{b_{s-1}^j})r_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})\\ &= \omega_0^{b_0^j} r_0(x_0^{b_0^j})\frac{w_1(x_0^{b_0^j}, x_1^{b_1^j})}{\vartheta_1(x_0^{b_0^j})}\vartheta_1(x_0^{b_0^j})r_1(x_0^{b_0^j}, x_1^{b_1^j})\prod_{s=2}^{n}\frac{w_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})}{\vartheta_s(x_{s-1}^{b_{s-1}^j})}\vartheta_s(x_{s-1}^{b_{s-1}^j})r_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})\\ &= r_0(x_0^{b_0^j})\frac{w_1(x_0^{b_0^j}, x_1^{b_1^j})}{\vartheta_1(x_0^{b_0^j})}\left[\omega_0^{b_0^j}\vartheta_1(x_0^{b_0^j})r_1(x_0^{b_0^j}, x_1^{b_1^j})\right]\prod_{s=2}^{n}\frac{w_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})}{\vartheta_s(x_{s-1}^{b_{s-1}^j})}\vartheta_s(x_{s-1}^{b_{s-1}^j})r_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})\\ &\propto r_0(x_0^{b_0^j})\omega_1^{b_1^j}\mathrm{p}_1(i_1^{b_1^j}, x_1^{b_1^j})\prod_{s=2}^{n}\frac{w_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})}{\vartheta_s(x_{s-1}^{b_{s-1}^j})}\vartheta_s(x_{s-1}^{b_{s-1}^j})r_s(x_{s-1}^{b_{s-1}^j}, x_s^{b_s^j})\\ &\propto r_0(x_0^{b_0^j})\prod_{s=1}^{n}\mathrm{p}_s(i_s^{b_s^j}, x_s^{b_s^j})\omega_n^{b_n^j},\end{aligned}$$

which, combined with (82), gives

$$\pi_n^N(j|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n}) \propto \omega_n^{b_n^j}.$$

□

Therefore, to sample $\{X_{0:n}^{B_{0:n}^J}, B_{0:n}^J\}$ from $\pi_n^N(\cdot, \cdot|\boldsymbol{x}_{0:n}, \boldsymbol{i}_{0:n})$,

Algorithm 6.6. [Particle Gibbs Part 2]

For $t = n$**,**

- sample
$$J \sim \left\{ \frac{\omega_n^1}{\Omega_n^N}, \ldots, \frac{\omega_n^N}{\Omega_n^N} \right\}.$$
- Set $B_n^J = J$.
- Set $X_n^J = x_n^J$.

For $t = n-1, \ldots, 0$**,**

- Set $B_t^J = i_{t+1}^{B_{t+1}^J}$.
- Set $X_t^J = x_t^{B_t^J}$.

As we can see, Algorithm 6.5 is very similar to auxiliary particle filtering. However, an important difference is that in Algorithm 6.5, one particle at each time point is specified/conditioned *a priori*. This particle-filtering-like procedure is referred to as the *conditional particle filtering (CPF)*.

Algorithm 6.7. [Particle Gibbs]

1. Initialize. Set $\{X_{0:n}[0], B_{0:n}[0]\}$ arbitrarily.
2. For Iteration $k \geq 1$,
   (a) run a CPF algorithm 6.5 conditional on $\{X_{0:n}[k-1], B_{0:n}[k-1]\}$, and
   (b) run Algorithm 6.6 and set $\{X_{0:n}[k], B_{0:n}[k]\} = \{X_{0:n}^{B_{0:N}^J}, B_{0:n}^J\}$.

The particle Gibbs sampling can be designed in various ways, leading to a variety of algorithm. The basic one 6.7, introduced in the seminal paper by Andrieu et al. (2010), is the most straightforward. However, this method is known to suffer from path degeneracy.

Hence, for this method to worker properly, the number of particles N needs to be large enough. For many problems, this is unrealistic from a computational point of view (Douc et al., 2014, P396). This issue can be addressed by modifying the basic Gibbs sweep, for instance by adding an ancestor sampling (Lindsten, Jordan, & Schön, 2012, 2014) or by adding a backward simulation (Lindsten & Schön, 2012).

The Particle Gibbs with Ancestor Sampling method, introduced by Lindsten et al. (2012, 2014), corresponds to a partially collapsed Gibbs sampler.

6.5.2 Partially Collapsed Gibbs Sampler

The partially collapsed Gibbs (PCG) sampler was introduced in Van Dyk and Park (2008).

Consider a four-step prototype Gibbs sampler with target distribution $\mathrm{p}(\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. We begin with the Gibbs sampler:

Sampler 1

Step 1 Draw $\boldsymbol{W}$ from $\mathrm{p}(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$.

Step 2 Draw $\boldsymbol{X}$ from $\mathrm{p}(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z})$.

Step 3 Draw $\boldsymbol{Y}$ from $\mathrm{p}(\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z})$.

Step 4 Draw $\boldsymbol{Z}$ from $\mathrm{p}(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Y})$.

Suppose it is possible to directly sample from $\mathrm{p}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and $\mathrm{p}(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})$.

Moving components in a step of an (ordinary) Gibbs sampler from being conditioned on to being sampled can improve the convergence characteristics of the sampler. This does not alter the stationary distribution of the sampled chain, nor does it destroy the compatibility of the conditional distributions. For example, we can sample $\boldsymbol{W}$ together with $\boldsymbol{Y}$ in step 3 and with $\boldsymbol{Z}$ in step 4. In step 3, we first sample from $\mathrm{p}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and then from $\mathrm{p}(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. The resulting Gibbs sampler iterates among the following steps:

Sampler 2

Step 1 Draw $\boldsymbol{W}^*$ from $\mathrm{p}(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$.

Step 2 Draw $\boldsymbol{X}$ from p($\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z}$).

Step 3 Draw ($\boldsymbol{W}^*, \boldsymbol{Y}$) from p($\boldsymbol{W}, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$).

Step 4 Draw ($\boldsymbol{W}, \boldsymbol{Z}$) from p($\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}$).

Here the superscript "$*$" designates an intermediate quantity that is sampled but is not part of the output of an iteration. In each step we condition on the most recently sampled value of each variable. This sampler is inefficient, in that it draws $\boldsymbol{W}$ three times in each iteration.

Permuting the steps of an (ordinary) Gibbs sampler does not alter its stationary distribution:

Sampler 3

Step 1 Draw ($\boldsymbol{W}^*, \boldsymbol{Y}$) from p($\boldsymbol{W}, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$).

Step 2 Draw ($\boldsymbol{W}^*, \boldsymbol{Z}$) from p($\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}$).

Step 3 Draw $\boldsymbol{W}$ from p($\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$).

Step 4 Draw $\boldsymbol{X}$ from p($\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z}$).

This permutation alters the transition kernel while maintaining the stationary distribution.

Note that the first two draws of $\boldsymbol{W}$ are not conditioned on and are not part of the output. Therefore, *removing the two redundant draws* will not change the transition kernel, or the stationary distribution. And that yields a partially collapsed Gibbs sampler that iterates among the following:

Sampler 4

Step 1 Draw $\boldsymbol{Y}$ from p($\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$).

Step 2 Draw $\boldsymbol{Z}$ from p($\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}$).

Step 3 Draw $\boldsymbol{W}$ from p($\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$).

Step 4 Draw $\boldsymbol{X}$ from p($\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Y}, \boldsymbol{Z}$).

The removal of an intermediate quantity must be done with great care. A rule of thumb is that it can be removed if, firstly, it is not conditioned on within the iteration, and, secondly, is not part of the output of the iteration.

6.5.3 Particle Gibbs with Ancestor Sampling

This method, which aims to address the degeneracy problem of particle Gibbs sampler, was introduced by Lindsten et al. (2012, 2014) and Lindsten and Schön (2013). The idea is to sample new values of the ancestor indexes $B^J_{0:n-1}$ as part of the CPF procedure 6.5. For $t = 1, \ldots, n$, after having sampled $(I_t^{-b_t^j}, X_t^{-b_t^j})$, we add a step in which we sample a new value for $I_t^{b_t^j}$, resulting in the following sweep:

Algorithm 6.8. [Particle Gibbs with Ancestor Sampling Part 1]
Input:

- Conditioned particles $(x_{0:n}^{b_{0:n}^j}, j)$.

For $t = 0$**:**

- For $\ell \neq b_0^j$, draw $X_0^\ell \sim r_0(\cdot)$;
- set $X_0^{b_0^j} = x_0^{b_0^j}$.
- For $\ell = 1, \ldots, N$, compute $\omega_0^\ell = g_0(X_0^\ell) w_0(X_0^\ell)$.

For $t = 1, \ldots, n$**:** Given $\{\boldsymbol{X}_{t-1}, \boldsymbol{\omega}_{t-1}\}$,

- for $\ell \neq b_t^j$, draw $(I_t^\ell, X_t^\ell) \sim \mathrm{p}_t(\cdot, \cdot)$, which is the transition density given by (68);
- set $X_t^{b_t^j} = x_t^{b_t^j}$.
- Draw $I_t^{b_t^j}$ with conditional probability of $I_t^{b_t^j} = \ell$ given by

$$\frac{\omega_{t-1}^\ell m(X_{t-1}^\ell, x_t^{b_t^j})}{\sum_{\ell=1}^N \omega_{t-1}^\ell m(X_{t-1}^\ell, x_t^{b_t^j})}.$$

- For $\ell = 1, \ldots, N$, compute $\omega_t^\ell = w_t(X_{t-1}^{I_t^\ell}, X_t^\ell) / \vartheta_t(X_{t-1}^{I_t^\ell})$.

Output:

- $(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{1:n-1}, \boldsymbol{I}_n^{-j})$.

This is referred to as the *conditional particle filter with ancestor sampling* (CPF-AS). The expression *ancestor sampling* refers to the fact that the ancestor indexes for the conditioned particles are sampled using one-step backward simulations. In practice, this step mitigates the effect of path degeneracy, and improves the mixing of the sampler. This procedure corresponds to sample from $\pi_n^N(\boldsymbol{X}_{0:n}, \boldsymbol{I}_{1:n-1}, \boldsymbol{I}_n^{-j} | X_{0:n}^{B_{0:n}^J}, J)$. To see this, note that:

Property 6.7. *For $t = 1, \ldots, n$,*

$$\pi_n^N(i_t^{b_t^j} | \boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) = \frac{\omega_{t-1}^\ell m(X_{t-1}^\ell, x_t^{b_t^j})}{\sum_{\ell=1}^N \omega_{t-1}^\ell m(X_{t-1}^\ell, x_t^{b_t^j})}.$$

Proof.

$$\pi_n^N(i_t^{b_t^j} | \boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) \propto \pi_n^N(\boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j).$$

Because $i_t^{b_t^j} = b_{t-1}^j$, and according to (77),

$$\begin{aligned}
\pi_n^N(\boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) &= \pi_n^N(\boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t}^{-b_{1:t}^j}, x_{0:n}^{b_{0:n}^j}, i_{1:t}^{b_{1:t}^j}, b_{t:n}^j) \\
&= \pi_n^N(\boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t}^{-b_{1:t}^j}, x_{0:n}^{b_{0:n}^j}, b_{0:n}^j) \\
&= \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \left(\prod_{\ell \neq b_0^j, \ell=1}^N r_0(x_0^\ell) \right) \prod_{s=1}^t \left(\prod_{\ell \neq b_s^j, \ell=1}^N \mathrm{p}_t(i_s^\ell, x_s^\ell) \right) \\
&= \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \frac{\left(\prod_{\ell=1}^N r_0(x_0^\ell) \right) \prod_{s=1}^t \left(\prod_{\ell=1}^N \mathrm{p}_t(i_s^\ell, x_s^\ell) \right)}{r_0(x_0^{b_0^j}) \prod_{s=1}^t \mathrm{p}_s(i_s^{b_s^j}, x_s^{b_s^j})}.
\end{aligned}$$

That is

$$\pi_n^N(i_t^{b_t^j} | \boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) \propto \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{N^{n+1}} \frac{\left(\prod_{\ell=1}^N r_0(x_0^\ell) \right) \prod_{s=1}^t \left(\prod_{\ell=1}^N \mathrm{p}_t(i_s^\ell, x_s^\ell) \right)}{r_0(x_0^{b_0^j}) \prod_{s=1}^t \mathrm{p}_s(i_s^{b_s^j}, x_s^{b_s^j})},$$

where, given $\{\boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j\}$, $\left(\prod_{\ell=1}^N r_0(x_0^\ell) \right) \prod_{s=1}^t \left(\prod_{\ell=1}^N \mathrm{p}_t(i_s^\ell, x_s^\ell) \right)$ does not depend on $i_t^{b_t^j}$, or equivalently b_{t-1}^j, in the sense that its value is constant no matter what

the value of $i_t^{b_t^j}$ is. Therefore, keeping only the terms that depend on $i_t^{b_t^j}$, or equivalently b_{t-1}^j,

$$\pi_n^N(i_t^{b_t^j} | \boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) \propto \frac{\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})}{\mathrm{p}_{t-1}(i_{t-1}^{b_{t-1}^j}, x_{t-1}^{b_{t-1}^j})}. \tag{83}$$

To expand $\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j})$, note that (72) implies

$$\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}) \propto q_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j}) m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}).$$

And (41) implies that

$$q_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j}) = w_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j}) r_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j}).$$

So

$$\begin{aligned}
\phi_{0:n|n}(x_{0:n}^{b_{0:n}^j}) &\propto w_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j}) r_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-2}^{b_{t-2}^j}) m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}) \\
&= \frac{w_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-1}^{b_{t-1}^j})}{\vartheta_{t-1}(x_{t-2}^{b_{t-2}^j})} \vartheta_{t-1}(x_{t-2}^{b_{t-2}^j}) r_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-2}^{b_{t-2}^j}) m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}) \\
&= \omega_{t-1}^{b_{t-1}^j} \frac{\omega_{t-2}^{b_{t-2}^j} \vartheta_{t-1}(x_{t-2}^{b_{t-2}^j}) r_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-2}^{b_{t-2}^j})}{\omega_{t-2}^{b_{t-2}^j}} m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}),
\end{aligned}$$

where the last equality is because of (69). And from (68),

$$\frac{\omega_{t-2}^{b_{t-2}^j} \vartheta_{t-1}(x_{t-2}^{b_{t-2}^j}) r_{t-1}(x_{t-2}^{b_{t-2}^j}, x_{t-2}^{b_{t-2}^j})}{\omega_{t-2}^{b_{t-2}^j}} \propto \mathrm{p}_{t-1}(i_{t-1}^{b_{t-1}^j}, x_{t-1}^{b_{t-1}^j}).$$

Therefore,

$$\phi_{0:n|n}(x^{b_{0:n}^j}) \propto \omega_{t-1}^{b_{t-1}^j} \mathrm{p}_{t-1}(i_{t-1}^{b_{t-1}^j}, x_{t-1}^{b_{t-1}^j}) m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}),$$

which, combined with (83), gives

$$\pi_n^N(i_t^{b_t^j} | \boldsymbol{x}_{0:t}^{-b_{0:t}^j}, \boldsymbol{i}_{1:t-1}, \boldsymbol{i}_t^{-b_t^j}, x_{0:n}^{b_{0:n}^j}, b_{t:n}^j) \propto \omega_{t-1}^{b_{t-1}^j} m(x_{t-1}^{b_{t-1}^j}, x_t^{b_t^j}).$$

□

To sample from $\pi_n^N(X_{0:n}^{B_{0:n}^J}, J | \boldsymbol{X}_{0:n}, \boldsymbol{I}_{1:n-1}, \boldsymbol{I}_n^{-j})$, apply Algorithm 6.6.

Algorithm 6.9. [Particle Gibbs with Ancestor Sampling]

1. Initialize. Set $\{X_{0:n}[0], B_{0:n}[0]\}$ arbitrarily.

2. For Iteration $k \geq 1$,

   (a) run a CPF-AS algorithm 6.8 conditional on $\{X_{0:n}[k-1], B_{0:n}[k-1]\}$, and

   (b) run Algorithm 6.6 and set $\{X_{0:n}[k], B_{0:n}[k]\} = \{X_{0:n}^{B_{0:N}^J}, B_{0:n}^J\}$.

It can be verified that Algorithm 6.9 corresponds to a partially collapsed Gibbs sampler.

Theorem 6.8. *For any $N \geq 2$, the PGAS sampler generates a process $\{X_{0:n}[k]\}_{k\geq 0}$ whose distribution converge in total variation to the posterior distribution $\phi_{0:n|n}$ for $\phi_{0:n|n}$-almost all starting points.*

For many models in practice, a moderate N, e.g. in the range 5 - 20, is enough to obtain a rapid mixing Gibbs kernel.

6.5.4 Backward Simulation in the particle Gibbs

Another option to address the problem of path degeneracy is Lindsten and Schön (2012).

7 Inference for Nonlinear State Space Models

Example 7.1 (NGM Model)**.** Consider the univariate model introduced in Netto, Gimeno, and Mendes (1978) discussed by Kitagawa (1987) and Carlin, Polson, and Stoffer (1992).

$$X_t = F_t^\theta(X_{t-1}) + W_t,$$
$$Y_t = H_t(X_t) + V_t,$$

with

$$F_t^\theta(X_{t-1}) = \alpha X_{t-1} + \beta \frac{X_{t-1}}{1 + X_{t-1}^2} + \gamma \cos[1.2(t-1)],$$
$$H_t(X_t) = \frac{X_t^2}{20},$$

where $X_0 \sim \mathrm{N}(\mu_0, \sigma_0^2)$, with $W_t \sim$ i.i.d. $N(0, \sigma_w^2)$ independent of $V_t \sim$ i.i.d. $\mathrm{N}(0, \sigma_v^2)$.

This model has become a standard model for testing numerical procedures.

We now consider applying some Bayesian approaches to fitting NLSS models via MCMC methods.

7.1 Particle Marginal Metropolis-Hastings

Revised according to Andrieu et al. (2010), which introduced the Particle MCMC (both Particle Marginal Metropolis-Hastings and Particle Gibbs) for the first time.

Consider now the scenario where we are interested in sampling from

$$\mathrm{p}(\theta, x_{0:n}|y_{0:n}) \propto \mathrm{p}(x_{0:n}|\theta, y_{0:n})\mathrm{L}(\theta; y_{0:n})\mathrm{p}(\theta),$$

where $\mathrm{p}(\theta)$ stands here for the prior parameter distribution; $\mathrm{L}(\theta; Y_{0:n})$ is the likelihood of the data under the model parameterized by θ.

It is natural to suggest the following form of proposal density for an MH update:

$$\mathrm{q}(\theta^*|\theta)\mathrm{p}(x_{0:n}^*|\theta^*, y_{0:n}).$$

The resulting acceptance probability is given by

$$1 \wedge \frac{\mathrm{L}(\theta^*; y_{0:n})\mathrm{p}(\theta^*)}{\mathrm{L}(\theta; y_{0:n})\mathrm{p}(\theta)} \frac{\mathrm{q}(\theta|\theta^*)}{\mathrm{q}(\theta^*|\theta)}.$$

For a general NLSS, it is not possible to compute the probability directly since the likelihood function is not available in closed form. However, by running an APF, we can construct a consistent estimate of the likelihood according to (78).

Algorithm 7.1 (Particle Marginal Metropolis-Hastings)**.**

1. Initialization, i=0.

   (a) set $\theta(0)$ arbitrarily;

   (b) run an APF algorithm targeting $\mathrm{p}(x_{0:n}|\theta(0), y_{0:n})$, sample $X_{0:n}(0) \sim \mathrm{p}(\cdot|\theta(0), y_{0:n})$, and let $\hat{\mathrm{L}}(\theta(0); y_{0:n})$ denote the likelihood estimate.

2. For iteration $i \geq 1$,

   (a) sample $\theta^* \sim \mathrm{q}(\cdot|\theta(i-1))$;

   (b) run an APF targeting $\mathrm{p}(x_{0:n}|\theta^*, y_{0:n})$, sample $X^*_{0:n} \sim \mathrm{p}(\cdot|\theta^*, y_{0:n})$, and let $\hat{\mathrm{L}}(\theta^*; y_{0:n})$ denote the likelihood estimate;

   (c) with probability

   $$1 \wedge \frac{\hat{\mathrm{L}}(\theta^*; y_{0:n})\mathrm{p}(\theta^*)}{\hat{\mathrm{L}}(\theta(i-1); y_{0:n})\mathrm{p}(\theta(i-1))} \frac{\mathrm{q}(\theta(i-1)|\theta^*)}{\mathrm{q}(\theta^*|\theta(i-1))}$$

   set

   $$\theta(i) = \theta^*, \quad X_{0:n}(i) = X^*_{0:n};$$

   otherwise set

   $$\theta(i) = \theta(i-1), \quad X_{0:n}(i) = X_{0:n}(i-1).$$

7.2 Gibbs Sampling for NLSS Model

Algorithm 7.2. [Idealized Gibbs Algorithm]

1. Draw $\theta' \sim \mathrm{p}(\theta|x_{0:n}, y_{0:n})$.

2. Draw $X'_{0:n} \sim \mathrm{p}(x_{0:n}|\theta', y_{0:n})$.

Sampling θ in Algorithm 7.2 - (1) is generally easier. The parameters are conditionally independent with distributions from standard parametric families, as long as the prior distribution is conjugate relative to the model specification. For non-conjugate models, one option is to replace Algorithm 7.2 - (1) with a Metropolis-Hastings step (Douc et al., 2014, P419). And another option is to sample θ' using an implicit importance sampler as introduced in Section 5.6 or Morzfeld, Tu, Wilkening, and Chorin (2015).

A common approach for Algorithm 7.2 - (2) is to sample the state variables one-at-a-time in separate Metropolis within Gibbs steps.

7.2.1 Metropolis Within Gibbs

The particular structure of the NLSS implies a simple decomposition of the full marginal distribution of the states. According to (72),

$$\begin{aligned} \mathrm{p}(x_{0:n}|\theta, y_{0:n}) &\propto \mathrm{p}(x_{0:n}, y_{0:n}|\theta) \\ &= \xi^\theta(x_0) g^\theta(x_0, y_0) \prod_{t=1}^{n} m_t^\theta(x_{t-1}, x_t) g^\theta(x_t, y_t). \end{aligned}$$

So the conditional probability density function of a single state variable

$$\mathrm{p}(x_t|x_{-t}, y_{0:n}) \propto m_t^\theta(x_{t-1}, x_t) m_{t+1}^\theta(x_t, x_{t+1}) g^\theta(x_t, y_t), \quad t = 1, \ldots, n-1. \tag{84}$$

At the two endpoints,

$$\begin{aligned} \mathrm{p}(x_0|x_{-0}, y_{0:n}, \theta) &\propto \xi(x_0) m_1^\theta(x_0, x_1) g^\theta(x_0, y_0), \\ \mathrm{p}(x_n|x_{-n}, y_{0:n}, \theta) &\propto m_n^\theta(x_{n-1}, x_n) g^\theta(x_n, y_n). \end{aligned}$$

For $t = 0, \ldots, n$, to sample the state X_t, a random walk Metropolis algorithm, introduced in Example 3.5, works as follows.

(a) At iteration j, denote the current value by $X_t^{(j)}$.

(b) Sample X_t^* from $\mathrm{N}(X_t^{(j)}, v_x^2)$, where v_x^2 is the tuning variance. Refer to Example 3.5.

(c) Determine the acceptance probability

$$\alpha = 1 \wedge \frac{\mathrm{p}(X_t^*|X_{-t}, Y_t, \theta)}{\mathrm{p}(X_t^{(j)}|X_{-t}, Y_t, \theta)},$$

where $\mathrm{p}(X_t^*|X_{-t}, Y_t, \theta)$ is given by (84).

(d) Select the new value as

$$X_t^{(j+1)} = \begin{cases} X_t^* & \text{with prob. } \alpha \\ X_t^{(j)} & \text{with prob. } 1-\alpha \end{cases}$$

Example 7.2 (NGM Model)**.** For Example 7.1, ν_x is set to be 0.05. The burn-in size was 500, the step size was 25(i.e. every 25th sampled values are retained), and they obtained 1000 samples *(why only the 25th? I would like to keep all samples)*. The entire procedure consisted of 25,500 draws. The time to complete the run was 10.2 minutes (for the authors).

7.2.2 Particle Gibbs With Ancestor Sampling

Another approach for Algorithm 7.2 - (2) is to replace it with a run of the PGAS algorithm.

References

Akashi, H., & Kumamoto, H. (1977). Random sampling approach to state estimation in switching environments. *Automatica*, *13*(4), 429–434.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.

Carlin, B. P., Polson, N. G., & Stoffer, D. S. (1992). A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, *87*(418), 493–500.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Chorin, A. J., Morzfeld, M., & Tu, X. (2010). Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, *5*(2), 221–240.

Chorin, A. J., & Tu, X. (2009). Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, *106*(41), 17249–17254.

Chorin, A. J., & Tu, X. (2012). An iterative implementation of the implicit nonlinear filter. *ESAIM: Mathematical Modelling and Numerical Analysis*, *46*(3), 535–543.

Corbae, D., Stinchcombe, M. B., & Zeman, J. (2009). *An introduction to mathematical analysis for economic theory and econometrics*. Princeton University Press.

Douc, R., Garivier, A., Moulines, E., Olsson, J., et al. (2011). Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, *21*(6), 2109–2145.

Douc, R., Moulines, E., & Olsson, J. (2009). Optimality of the auxiliary particle filter. *Probability and Mathematical Statistics*, *29*(1), 1–28.

Douc, R., Moulines, E., & Stoffer, D. (2014). *Nonlinear time series: theory, methods and*

applications with r examples. CRC Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In *Iee proceedings f (radar and signal processing)* (Vol. 140, pp. 107–113).

Handschin, J., & Mayne, D. Q. (1969). Monte carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International journal of control*, *9*(5), 547–559.

Johansen, A. M., & Doucet, A. (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, *78*(12), 1498–1504.

Kitagawa, G. (1987). Non-gaussian statespace modeling of nonstationary time series. *Journal of the American statistical association*, *82*(400), 1032–1041.

Kong, A., Liu, J. S., & Wong, W. H. (1994). Sequential imputations and bayesian missing data problems. *Journal of the American statistical association*, *89*(425), 278–288.

Lindsten, F., Jordan, M. I., & Schön, T. B. (2012). Ancestor sampling for particle gibbs. In *Advances in neural information processing systems* (pp. 2591–2599).

Lindsten, F., Jordan, M. I., & Schön, T. B. (2014). Particle gibbs with ancestor sampling. *Journal of Machine Learning Research*, *15*(1), 2145–2184.

Lindsten, F., & Schön, T. B. (2012). On the use of backward simulation in the particle gibbs sampler. In *Acoustics, speech and signal processing (icassp), 2012 ieee international conference on* (pp. 3845–3848).

Lindsten, F., & Schön, T. B. (2013). Backward simulation methods for monte carlo statistical inference. *Foundations and Trends® in Machine Learning*, *6*(1), 1–143.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, *6*(2), 113–119.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Morzfeld, M., Tu, X., Atkins, E., & Chorin, A. J. (2012). A random map implementation of implicit filters. *Journal of Computational Physics*, *231*(4), 2049–2066.

Morzfeld, M., Tu, X., Wilkening, J., & Chorin, A. J. (2015). Parameter estimation by implicit sampling. *Communications in Applied Mathematics and Computational Science*, *10*(2), 205–225.

Netto, M., Gimeno, L., & Mendes, M. (1978). On the optimal and suboptimal nonlinear filtering problem for discrete-time systems. *IEEE Transactions on Automatic Control*, *23*(6), 1062–1067.

Olsson, J., & Ryden, T. (2011). Rao-blackwellization of particle markov chain monte carlo methods using forward filtering backward sampling. *IEEE Transactions on Signal Processing*, *59*(10), 4606–4619.

Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, *94*(446), 590–599.

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, *82*(398), 543–546.

Rubin, D. B. (1988). Using the sir algorithm to simulate posterior distributions. *Bayesian statistics*, *3*(1), 395–402.

Rudin, W. (1976). Principles of mathematical analysis (international series in pure & applied mathematics).

Van Dyk, D. A., & Park, T. (2008). Partially collapsed gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, *103*(482), 790–796.

Walkden, C. (2017). Lectures notes on ergodic theory in the university of manchester.

Zaritskii, V., Svetnik, V., & Shimelevich, L. (1976). Monte-carlo technique in problems of optimal information processing. *Automation and Remote Control*, *36*(12), 2015–2022.