Electronic Journal of Probability

Vol. 14 (2009), Paper no. 2, pages 27–49.

Journal URL
http://www.math.washington.edu/~ejpecp/

Forgetting of the initial condition for the filter in general state-space hidden Markov chain: a coupling approach

Randal Douc
Institut Télécom/Télécom SudParis,
France,
randal.douc@it-sudparis.eu

Eric Moulines
Institut Télécom/Télécom ParisTech, CNRS UMR 8151
46 rue Barrault,
75634 Paris Cédex 13, France,
eric.moulines@telecom-paristech.fr

Ya'acov Ritov
Department of Statistics, The Hebrew University of Jerusalem,
yaacov.ritov@huji.ac.il

Abstract

We give simple conditions that ensure exponential forgetting of the initial conditions of the filter for general state-space hidden Markov chain. The proofs are based on the coupling argument applied to the posterior Markov kernels. These results are useful both for filtering hidden Markov models using approximation methods (e.g., particle filters) and for proving asymptotic properties of estimators. The results are general enough to cover models *like* the Gaussian state space model, without using the special structure that permits the application of the Kalman filter.

Key words: hidden Markov chain, stability, non-linear filtering, coupling.

AMS 2000 Subject Classification: Primary 93E11; Secondary: 60J57.

Submitted to EJP on December 3, 2007, final version accepted December 17, 2008.

1 Introduction and Notation

We consider the filtering problem for a Markov chain $\{X_k, Y_k\}_{k\geq 0}$ with *state* X and *observation* Y. The state process $\{X_k\}_{k\geq 0}$ is an homogeneous Markov chain taking value in a measurable set X equipped with a σ-algebra $\mathscr{B}(\mathsf{X})$. We let Q be the transition kernel of the chain. The observations $\{Y_k\}_{k\geq 0}$ takes values in a measurable set Y ($\mathscr{B}(\mathsf{Y})$ is the associated σ-algebra). For $i \leq j$, denote $Y_{i:j} \triangleq (Y_i, Y_{i+1}, \cdots, Y_j)$. Similar notation will be used for other sequences. We assume furthermore that for each $k \geq 1$ and given X_k, Y_k is independent of $X_{1:k-1}$,$X_{k+1:\infty}$, $Y_{1:k-1}$, and $Y_{k+1:\infty}$. We also assume that for each $x \in \mathsf{X}$, the conditional law has a density $g(x,\cdot)$ with respect to some fixed σ-finite measure on the Borel σ-field $\mathscr{B}(\mathscr{Y})$.

We denote by $\phi_{\xi,n}[y_{0:n}]$ the distribution of the hidden state X_n conditionally on the observations $y_{0:n} \stackrel{\text{def}}{=} [y_0, \ldots, y_n]$, which is given by

$$\phi_{\xi,n}[y_{0:n}](A) = \frac{\int_{\mathsf{X}^{n+1}} \xi(dx_0) g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i) g(x_i, y_i) \mathbb{1}_A(x_n)}{\int_{\mathsf{X}^{n+1}} \xi(dx_0) g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i) g(x_i, y_i)}\,, \tag{1}$$

In practice the model is rarely known exactly and therefore suboptimal filters are computed by replacing the unknown transition kernel, likelihood function and initial distribution by approximations.

The choice of these quantities plays a key role both when studying the convergence of sequential Monte Carlo methods or when analysing the asymptotic behaviour of the maximum likelihood estimator (see *e.g.,* [8] or [5] and the references therein). A key point when analyzing maximum likelihood estimator or the stability of the filter over infinite horizon is to ask whether $\phi_{\xi,n}[y_{0:n}]$ and $\phi_{\xi',n}[y_{0:n}]$ are close (in some sense) for large values of n, and two different choices of the initial distribution ξ and ξ'.

The forgetting property of the initial condition of the optimal filter in nonlinear state space models has attracted many research efforts and it is impossible to give credit to every contributors. The purpose of the short presentation of the existing results below is mainly to allow comparison of assumptions and results presented in this contributions with respect to those previously reported in the literature. The first result in this direction has been obtained by [15], who established L_p-type convergence of the optimal filter initialised with the wrong initial condition to the filter initialised with the true initial distribution; their proof does not provide rate of convergence. A new approach based on the Hilbert projective metric has later been introduced in [1] to establish the exponential stability of the optimal filter with respect to its initial condition. However their results are based on stringent *mixing* conditions for the transition kernels; these conditions state that there exist positive constants ε_- and ε_+ and a probability measure λ on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$ such that for $f \in \mathbb{B}_+(\mathsf{X})$,

$$\varepsilon_- \lambda(f) \leq Q(x, f) \leq \varepsilon_+ \lambda(f)\,, \quad \text{for any } x \in \mathsf{X}\,. \tag{2}$$

This condition implies in particular that the chain is uniformly geometrically ergodic. Similar results were obtained independently by [9] using the Dobrushin ergodicity coefficient (see [10] for further refinements of this result). The mixing condition has later been weakened by [6], under the assumption that the kernel Q is positive recurrent and is dominated by some reference measure λ:

$$\sup_{(x,x') \in \mathsf{X}\times\mathsf{X}} q(x,x') < \infty \quad \text{and} \quad \int \operatorname{essinf} q(x,x') \pi(x) \lambda(dx) > 0\,,$$

where $q(x,\cdot) = \frac{dQ(x,\cdot)}{d\lambda}$, essinf is the essential infimum with respect to λ and $\pi d\lambda$ is the stationary distribution of the chain Q . Although the upper bound is reasonable, the lower bound is restrictive in many applications and fails to be satisfied *e.g.,* for the linear state space Gaussian model.

In [13], the stability of the optimal filter is studied for a class of kernels referred to as *pseudo-mixing*. The definition of pseudo-mixing kernel is adapted to the case where the state space is $\mathsf{X} = \mathbb{R}^d$, equipped with the Borel sigma-field $\mathscr{B}(\mathsf{X})$. A kernel Q on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$ is *pseudo-mixing* if for any compact set C with a diameter d large enough, there exist positive constants $\varepsilon_-(d) > 0$ and $\varepsilon_+(d) > 0$ and a measure λ_C (which may be chosen to be finite without loss of generality) such that

$$\varepsilon_-(d)\lambda_C(A) \le Q(x,A) \le \varepsilon_+(d)\lambda_C(A)\,, \quad \text{for any } x \in C, A \in \mathscr{B}(\mathsf{X}) \tag{3}$$

This condition implies that for any $(x', x'') \in C \times C$,

$$\frac{\varepsilon_-(d)}{\varepsilon_+(d)} < \mathrm{essinf}_{x\in\mathsf{X}} q(x',x)/q(x'',x) \le \mathrm{esssup}_{x\in\mathsf{X}} q(x',x)/q(x'',x) \le \frac{\varepsilon_+(d)}{\varepsilon_-(d)}\,,$$

where $q(x,\cdot) \stackrel{\text{def}}{=} dQ(x,\cdot)/d\lambda_C$, and esssup and essinf denote the essential supremum and infimum with respect to λ_C. This condition is obviously more general than (2): in particular, [13] gives non-trivial examples of pseudo-mixing Markov chains which are not uniformly ergodic. Nevertheless, this assumption is not satisfied in the linear Gaussian case (see [13, Example 4.3]).

Several attempts have been made to establish the stability conditions under the so-called *small* noise condition. The first result in this direction has been obtained by [1] (in continuous time) who considered an ergodic diffusion process with constant diffusion coefficient and linear observations: when the variance of the observation noise is sufficiently small, [1] established that the filter is exponentially stable. Small noise conditions also appeared (in a discrete time setting) in [4] and [16]. These results do not allow to consider the linear Gaussian state space model with arbitrary noise variance.

More recently, [7] prove that the nonlinear filter forgets its initial condition in mean over the observations for functions satisfying some integrability conditions. The main result presented in this paper relies on the martingale convergence theorem rather than direct analysis of filtering equations. Unfortunately, this method of proof cannot provide any rate of convergence.

It is tempting to assume that forgetting of the initial condition should be true in general, and that the lack of proofs for the general state-space case is only a matter of technicalities. The heuristic argument says that either

- the observations Y's are informative, and we learn about the hidden state X from the Ys around it, and forget the initial starting point.
- the observations Ys are non-informative, and then the X chain is moving by itself, and by itself it forgets its initial condition, for example if it is positive recurrent.

Since we expect that the forgetting of the initial condition is retained in these two extreme cases, it is probably so under any condition. However, this argument is false, as is shown by the following examples where the conditional chain does not forget its initial condition whereas the unconditional chain does. On the other hand, it can be that observed process, $\{Y_k\}_{k\ge0}$ is not ergodic, while the conditional chain uniformly forgets the initial condition.

Example 1. Suppose that $\{X_k\}_{k\geq 0}$ are i.i.d. $B(1,1/2)$. Suppose $Y_i = \mathbb{1}(X_i = X_{i-1})$. Then $\mathrm{P}\left(X_i = 1 \,\middle|\, X_0 = 0, Y_{0:n}\right) = 1 - \mathrm{P}\left(X_i = 1 \,\middle|\, X_1 = 1, Y_{0:n}\right) \in \{0,1\}$.

Here is a slightly less extreme example. Consider a Markov chain on the unit circle. All values below are considered modulus 2π. We assume that $X_i = X_{i-1} + U_i$, where the state noise $\{U_k\}_{k\geq 0}$ are i.i.d. . The chain is hidden by additive white noise: $Y_i = X_i + \varepsilon_i$, $\varepsilon_i = \pi W_i + V_i$, where W_i is Bernoulli random variable independent of V_i. Suppose now that U_i and V_i are symmetric and supported on some small interval. The hidden chain does not forget its initial distribution under this model. In fact the support of the distribution of X_i given $Y_{0:n}$ and $X_0 = x_0$ is disjoint from the support of its distribution given $Y_{0:n}$ and $X_0 = x_0 + \pi$.

On the other hand, let $\{Y_k\}_{k\geq 0}$ be an arbitrary process. Suppose it is modeled (incorrectly!) by a autoregressive process observed in additive noise. We will show that under different assumptions on the distribution of the state and the observation noise, the conditional chain (given the observations Ys which are not necessarily generated by the model) forgets its initial condition geometrically fast.

The proofs presented in this paper are based on generalization of the notion of small sets and coupling of the two (non-homogenous) Markov chains sampled from the distribution of $X_{0:n}$ given $Y_{0:n}$. The coupling argument is based on presenting two chains $\{X_k\}$ and $\{X'_k\}$, which marginally follow the same sequence of transition kernels, but have different initial distributions of the starting state. The chains move independently, until they *coupled* at a random time T, and from that time on they remain equal.

Roughly speaking, the two copies of the chain may couple at a time k if they stand close one to the other. Formally, we mean by that, that the the pair of states of the two chains at time k belong to some set, which may depend on the current, but also past and future observations. The novelty of the current paper is by considering sets which are in fact genuinely defined by the pair of states. For example, the set can be defined as $\{(x,x') : \|x - x'\| < c\}$. That is, close in the usual sense of the word.

The prototypical example we use is the non-linear state space model:

$$\begin{aligned} X_i &= a(X_{i-1}) + U_i \\ Y_i &= b(X_i) + V_i, \end{aligned} \tag{4}$$

where $\{U_k\}_{k\geq 0}$ is the *state noise* and $\{V_k\}_{k\geq 0}$ is the *measurement noise*. Both $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are assumed to be i.i.d. and mutually independent. Of course, the filtering problem for the linear version of this model with independent Gaussian noise is solved explicitly by the Kalman filter. But this is one of the few non-trivial models which admits a simple solution. Under the Gaussian linear model, we argue that whatever are $Y_{0:n}$, two independent chains drawn from the conditional distribution will be remain close to each other even if the Ys are drifting away. Any time they will be close, they will be able to couple, and this will happen quite frequently.

Our approach for proving that a chain forgets its initial conditions can be decomposed in two stages. We first argue that there are *coupling sets* (which may depend on the observations, and may also vary according to the iteration index) where we can couple two copies of the chains, drawn independently from the conditional distribution given the observations and started from two different initial conditions, with a probability which is an explicit function of the observations. We then argue that a pair of chains are likely to drift frequently towards these coupling sets.

The first group of results identify situations in which the coupling set is given in a product form, and in particular in situations where $\mathsf{X} \times \mathsf{X}$ is a coupling set. In the typical situation, many values of

Y_i entail that X_i is in some set with high probability, and hence the two conditionally independent copies are likely to be in this set and close to each other. In particular, this enables us to prove the convergence of (nonlinear) state space processes with bounded noise and, more generally, in situations where the tails of the observations error is thinner than those of dynamics innovations.

The second argument generalizes the standard drift condition to the coupling set. The general argument specialized to the linear-Gaussian state model is surprisingly simple. We generalize this argument to the linear model where both the dynamics innovations and the measurement errors have strongly unimodal density.

2 Notations and definitions

Let n be a given positive index and consider the finite-dimensional distributions of $\{X_k\}_{k\geq 0}$ given $Y_{0:n}$. It is well known (see [5, Chapter 3]) that, for any positive index k, the distribution of X_k given $X_{0:k-1}$ and $Y_{0:n}$ reduces to that of X_k given X_{k-1} only and $Y_{0:n}$. The following definitions will be instrumental in decomposing the joint posterior distributions.

Definition 1 (Backward functions)**.** *For $k \in \{0, \dots, n\}$, the backward function $\beta_{k|n}$ is the non-negative measurable function on $\mathsf{Y}^{n-k} \times \mathsf{X}$ defined by*

$$\beta_{k|n}(x) = \int \cdots \int Q(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \prod_{l=k+2}^{n} Q(x_{l-1}, dx_l) g(x_l, y_l) \,, \tag{5}$$

for $k \leq n-1$ (with the same convention that the rightmost product is empty for $k = n-1$); $\beta_{n|n}(\cdot)$ is set to the constant function equal to 1 on X.

For notational simplicity, the dependence of the backward function in the observations y's is implicit. The term "backward variables" is part of the HMM credo and dates back to the seminal work of Baum and his colleagues [2, p. 168]. The backward functions may be obtained, for all $x \in \mathsf{X}$ by the recursion

$$\beta_{k|n}(x) = \int Q(x, dx') g(x', y_{k+1}) \beta_{k+1|n}(x') \tag{6}$$

operating on decreasing indices $k = n-1$ down to 0 from the initial condition

$$\beta_{n|n}(x) = 1 \,. \tag{7}$$

Definition 2 (Forward Smoothing Kernels)**.** *Given $n \geq 0$, define for indices $k \in \{0, \dots, n-1\}$ the transition kernels*

$$\mathrm{F}_{k|n}(x, A) \stackrel{\text{def}}{=} \begin{cases} [\beta_{k|n}(x)]^{-1} \int_A Q(x, dx') g(x', y_{k+1}) \beta_{k+1|n}(x') & \textit{if } \beta_{k|n}(x) \neq 0 \\ 0 & \textit{otherwise} \,, \end{cases} \tag{8}$$

for any point $x \in \mathsf{X}$ and set $A \in \mathscr{B}(\mathsf{X})$. For indices $k \geq n$, simply set

$$\mathrm{F}_{k|n} \stackrel{\text{def}}{=} Q \,, \tag{9}$$

where Q is the transition kernel of the unobservable chain $\{X_k\}_{k\geq 0}$.

Note that for indices $k \le n-1$, $\mathrm{F}_{k|n}$ depends on the future observations $Y_{k+1:n}$ through the backward variables $\beta_{k|n}$ and $\beta_{k+1|n}$ only. The subscript n in the $\mathrm{F}_{k|n}$ notation is meant to underline the fact that, like the backward functions $\beta_{k|n}$, the forward smoothing kernels $\mathrm{F}_{k|n}$ depend on the final index n where the observation sequence ends. Thus, for any $x \in \mathsf{X}$, $A \mapsto \mathrm{F}_{k|n}(x,A)$ is a probability measure on $\mathscr{B}(\mathsf{X})$. Because the functions $x \mapsto \beta_{k|n}(x)$ are measurable on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$, for any set $A \in \mathscr{B}(\mathsf{X})$, $x \mapsto \mathrm{F}_{k|n}(x,A)$ is $\mathscr{B}(\mathsf{X})/\mathscr{B}(\mathbb{R})$-measurable. Therefore, $\mathrm{F}_{k|n}$ is indeed a Markov transition kernel on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$.

Given n, for any index $k \ge 0$ and function $f \in \mathscr{F}_{\mathrm{b}}(\mathsf{X})$,

$$\mathrm{E}_{\xi}[f(X_{k+1}) \mid X_{0:k}, Y_{0:n}] = \mathrm{F}_{k|n}(X_k, f) .$$

More generally, for any integers n and m, function $f \in \mathscr{F}_{\mathrm{b}}\left(\mathsf{X}^{m+1}\right)$ and initial probability ξ on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$,

$$\mathrm{E}_{\xi}[f(X_{0:m}) \mid Y_{0:n}] = \int \cdots \int f(x_{0:m})\, \phi_{\xi,0|n}(dx_0) \prod_{i=1}^{m} \mathrm{F}_{i-1|n}(x_{i-1}, dx_i) , \tag{10}$$

where $\{\mathrm{F}_{k|n}\}_{k\ge 0}$ are defined by (8) and (9), and $\phi_{\xi,k|n}$ is the marginal smoothing distribution of the state X_k given the observations $Y_{0:n}$. Note that $\phi_{\xi,k|n}$ may be expressed, for any $A \in \mathscr{B}(\mathsf{X})$, as

$$\phi_{\xi,k|n}(A) = \left[\int \phi_{\xi,k}(dx)\beta_{k|n}(x)\right]^{-1} \int_A \phi_{\xi,k}(dx)\beta_{k|n}(x) , \tag{11}$$

where $\phi_{\xi,k}$ is the filtering distribution defined in (1) and $\beta_{k|n}$ is the backward function.

3 Coupling constants, coupling sets and the coupling construction

3.1 Coupling constant and coupling sets

As outlined in the introduction, our proofs are based on coupling two copies of the conditional chain started from two different initial conditions. For any two probability measures μ_1 and μ_2 we define the total variation distance $\left\|\mu_1 - \mu_2\right\|_{\mathrm{TV}} = 2\sup_A |\mu_1(A) - \mu_2(A)|$ and we also recall the identities $\sup_{|f|\le 1} |\mu_1(f) - \mu_2(f)| = \left\|\mu_1 - \mu_2\right\|_{\mathrm{TV}}$ and $\sup_{0\le f\le 1} |\mu_1(f) - \mu_2(f)| = (1/2)\left\|\mu_1 - \mu_2\right\|_{\mathrm{TV}}$. Let n and m be integers, and $k \in \{0, \dots, n-m\}$. Define the m-skeleton of the forward smoothing kernel as follows:

$$\mathrm{F}_{k,m|n} \stackrel{\text{def}}{=} \mathrm{F}_{km|n} \dots \mathrm{F}_{km+m-1|n} , \tag{12}$$

Definition 3 (Coupling constant of a set). *Let n and m be integers, and let $k \in \{0, \dots, n-m\}$. The* coupling constant *of the set $C \subset \mathsf{X} \times \mathsf{X}$ is defined as*

$$\varepsilon_{k,m|n}(C) \stackrel{\text{def}}{=} 1 - \frac{1}{2} \sup_{(x,x')\in C} \left\|\mathrm{F}_{k,m|n}(x,\cdot) - \mathrm{F}_{k,m|n}(x',\cdot)\right\|_{\mathrm{TV}} . \tag{13}$$

This definition implies that the coupling constant is the largest $\varepsilon \ge 0$ such that there exists a probability kernel ν on $\mathsf{X} \times \mathsf{X}$, satisfying for any $(x, x') \in C$, and $A \in \mathscr{B}(\mathsf{X})$,

$$\mathrm{F}_{k,m|n}(x,A) \wedge \mathrm{F}_{k,m|n}(x',A) \ge \varepsilon \nu(x, x'; A) . \tag{14}$$

The coupling construction is of interest only if we may find set-valued functions $\bar{C}_{k|n}$ whose coupling constants $\varepsilon_{k,m|n}(\bar{C}_{k|n})$ are 'most often' non-zero (recall that these quantities are typically functions of the whole trajectory $y_{0:n}$). It is not always easy to find such sets because the definition of the coupling constant involves the product $\mathrm{F}_{k|n}$ forward smoothing kernels, which is not easy to handle. In some situations (but not always), it is possible to identify appropriate sets from the properties of the unconditional transition kernel Q.

Definition 4 (Strong small set)**.** *A set $C \in \mathscr{B}(\mathsf{X})$ is a* strong small set *for the transition kernel Q, if there exists a measure ν_C and constants $\sigma_-(C) > 0$ and $\sigma_+(C) < \infty$ such that, for all $x \in C$ and $A \in \mathscr{B}(\mathsf{X})$,*

$$\sigma_-(C)\nu_C(A) \leq Q(x,A) \leq \sigma_+(C)\nu_C(A)\,. \tag{15}$$

The following Lemma helps to characterize appropriate sets where coupling may occur with a positive probability from products of strong small sets.

Proposition 5. *Assume that C is a strong small set. Then, for any n and any $k \in \{0,\ldots,n\}$, the coupling constant of the set $C \times C$ is uniformly lower bounded by the ratio $\sigma_-(C)/\sigma_+(C)$.*

Proof. The proof is postponed to the appendix. □

Assume that $\mathsf{X} = \mathbb{R}^d$, and that the kernel satisfies the *pseudo-mixing* condition (3). We may choose a compact set C with diameter $d = \mathrm{diam}(C)$ large enough so that C is a strong small set (*i.e.,* (15) is satisfied). The coupling constant of $\bar{C} = C \times C$ is lower bounded by $\varepsilon_-(d)/\varepsilon_+(d)$ uniformly over the observations, where the constant $\varepsilon_-(d)$ and $\varepsilon_+(d)$ are defined in (3).

Nevertheless, though the existence of small sets is automatically guaranteed for phi-irreducible Markov chains, the conditions imposed for the existence of a strong small set are much more stringent. As shown below, it is sometimes worthwhile to consider coupling set which are much larger than products of strong small sets.

3.2 The coupling construction

We may now proceed to the coupling construction. The construction introduced here is a straightforward adaptation of the coupling construction for Markov Chain on general state-spaces (see for example [12], [14] and [17]). Let n be an integer, and for any $k \in \{0,\ldots,\lfloor n/m \rfloor\}$, let $\bar{C}_{k|n}$ be a set-valued function, $\bar{C}_{k|n} : \mathsf{Y}^n \to \mathscr{B}(\mathsf{X}\times\mathsf{X})$. We define $\bar{\mathrm{R}}_{k,m|n}$ as the Markov transition kernel satisfying, for all $(x,x') \in \bar{C}_{k|n}$ and for all $A,A' \in \mathscr{B}(\mathsf{X})$ and $(x,x') \in \bar{C}_{k|n}$,

$$\begin{aligned}\bar{\mathrm{R}}_{k,m|n}(x,x';A\times A') = &\left\{(1-\varepsilon_{k,m|n})^{-1}(\mathrm{F}_{k,m|n}(x,A) - \varepsilon_{k,m|n}\nu_{k,m|n}(x,x';A))\right\} \\ &\times \left\{(1-\varepsilon_{k,m|n})^{-1}(\mathrm{F}_{k,m|n}(x',A') - \varepsilon_{k,m|n}\nu_{k,m|n}(x,x';A'))\right\}\,,\end{aligned} \tag{16}$$

where the dependence on $\bar{C}_{k|n}$ of the coupling constant $\varepsilon_{k,m|n}$ and of the minorizing probability $\nu_{k,m|n}$ is omitted for simplicity. For all $(x,x') \in \mathsf{X}\times\mathsf{X}$, we define

$$\bar{\mathrm{F}}_{k,m|n}(x,x';\cdot) = \mathrm{F}_{k,m|n} \otimes \mathrm{F}_{k,m|n}(x,x';\cdot)\,, \tag{17}$$

where, for two kernels K and L on X, $K \otimes L$ is the tensor product of the kernels K and L, *i.e.,* for all $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $A, A' \in \mathscr{B}(\mathsf{X})$

$$K \otimes L(x, x'; A \times A') = K(x, A) L(x', A') . \tag{18}$$

Define the product space $\mathsf{Z} = \mathsf{X} \times \mathsf{X} \times \{0, 1\}$, and the associated product sigma-algebra $\mathscr{B}(\mathsf{Z})$. Define on the space $(\mathsf{Z}^{\mathbb{N}}, \mathscr{B}(\mathsf{Z})^{\otimes \mathbb{N}})$ a Markov chain $Z_i \stackrel{\text{def}}{=} (\tilde{X}_i, \tilde{X}'_i, d_i)$, $i \in \{0, \ldots, n\}$ as follows. If $d_i = 1$, then draw $\tilde{X}_{i+1} \sim \mathrm{F}_{i,m|n}(\tilde{X}_i, \cdot)$, and set $\tilde{X}'_{i+1} = \tilde{X}_{i+1}$ and $d_{i+1} = 1$. Otherwise, if $(\tilde{X}_i, \tilde{X}'_i) \in \bar{C}_{i|n}$, flip a coin with probability of heads $\varepsilon_{i,m|n}$. If the coin comes up heads, then draw $\tilde{X}_{i+1}$ from $\nu_{i,m|n}(\tilde{X}_i, \tilde{X}'_i; \cdot)$, and set $\tilde{X}'_{i+1} = \tilde{X}_{i+1}$ and $d_{i+1} = 1$. If the coin comes up tails, then draw $(\tilde{X}_{i+1}, \tilde{X}'_{i+1})$ from the residual kernel $\bar{\mathrm{R}}_{i,m|n}(\tilde{X}_i, \tilde{X}'_i; \cdot)$ and set $d_{i+1} = 0$. If $(\tilde{X}_i, \tilde{X}'_i) \notin \bar{C}_{i|n}$, then draw $(\tilde{X}_{i+1}, \tilde{X}'_{i+1})$ according to the kernel $\bar{\mathrm{F}}_{i,m|n}(\tilde{X}_i, \tilde{X}'_i; \cdot)$ and set $d_{i+1} = 0$. For μ a probability measure on $\mathscr{B}(\mathsf{Z})$, denote P^Y_μ the probability measure induced by the Markov chain Z_i, $i \in \{0, \ldots, n\}$ with initial distribution μ. It is then easily checked that for any $i \in \{0, \ldots, \lfloor n/m \rfloor\}$ and any initial distributions ξ and ξ', and any $A, A' \in \mathscr{B}(\mathsf{X})$,

$$\mathrm{P}^Y_{\xi \otimes \xi' \otimes \delta_0} \left(Z_i \in A \times \mathsf{X} \times \{0, 1\} \right) = \phi_{\xi, im|n}(A) ,$$
$$\mathrm{P}^Y_{\xi \otimes \xi' \otimes \delta_0} \left(Z_i \in \mathsf{X} \times A' \times \{0, 1\} \right) = \phi_{\xi', im|n}(A) ,$$

where δ_x is the Dirac measure and $\otimes$ is the tensor product of measures and $\phi_{\xi,k|n}$ is the marginal posterior distribution given by (11)

Note that d_i is the *bell variable*, which shall indicate whether the chains have coupled ($d_i = 1$) or not ($d_i = 0$) by time i. Define the *coupling time*

$$T = \inf\{k \geq 1, d_k = 1\} , \tag{19}$$

with the convention $\inf \emptyset = \infty$. By the Lindvall inequality, the total variation distance between the filtering distribution associated to two different initial distribution ξ and ξ' is bounded by the tail distribution of the coupling time,

$$\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \leq 2\, \mathrm{P}^Y_{\xi \otimes \xi' \otimes \delta_0} (T \geq \lfloor n/m \rfloor) . \tag{20}$$

In the following section, we consider several conditions allowing to bound the tail distribution of the coupling time. Such bounds depend crucially on the coupling constant of such sets and also on probability bounds of the return time to these coupling sets.

4 Coupling over the whole state-space

The easiest situation is when the coupling constant of the whole state space $\varepsilon_{k,m|n}(\mathsf{X} \times \mathsf{X})$ is away from zero for sufficiently many trajectories $y_{0:n}$; for unconditional Markov chains, this property occurs when the chain is uniformly ergodic (*i.e.,* satisfies the Doeblin condition). This is still the case here, through now the constants may depend on the observations Y's. As stressed in the discussion, perhaps surprisingly, we will find non trivial examples where the coupling constant $\varepsilon_{k,m|n}(\mathsf{X} \times \mathsf{X})$ is bounded away from zero for all $y_{0:n}$, whereas the underlying unconditional Markov chain is *not* uniformly geometrically ergodic. We state without proof the following elementary result.

Theorem 6. *Let n be an integer and $\ell \geq 1$. Then,*

$$\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \leq 2 \prod_{k=0}^{\lfloor n/m \rfloor} \left\{ 1 - \varepsilon_{k,m|n}(\mathsf{X} \times \mathsf{X}) \right\} .$$

Example 2 (Uniformly ergodic kernel). When X is a strong small set then one may set $m = 1$ and, using Proposition 5, the coupling constant $\varepsilon_{k,1|n}(\mathsf{X} \times \mathsf{X})$ of the set $\mathsf{X} \times \mathsf{X}$ is lower bounded by $\sigma_-(\mathsf{X})/\sigma_+(\mathsf{X})$, where the constants $\sigma_-(\mathsf{X})$ and $\sigma_+(\mathsf{X})$ are defined in (15). In such a case, Theorem 6 shows that $\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \leq \{1 - \sigma_-(\mathsf{X})/\sigma_+(\mathsf{X})\}^n$.

Example 3 (Bounded observation noise). Assume that a Markov chain $\{X_k\}_{k\geq 0}$ in $\mathsf{X} = \mathbb{R}^d$ is observed in a bounded noise. The case of bounded error is of course particular, because the observations of the Y's allow to locate the corresponding X's within a set. More precisely, we assume that $\{X_k\}_{k\geq 0}$ is a Markov chain with transition kernel Q having density q with respect to the Lebesgue measure and $Y_k = b(X_k) + V_k$ where,

- $\{V_k\}$ is an i.i.d., independent of $\{X_k\}$, with density p_V. In addition, $p_V(|x|) = 0$ for $|x| \geq M$.
- the transition density $(x, x') \mapsto q(x, x')$ is strictly positive and continuous.
- The level sets of b, $\{x \in \mathsf{X} : |b(x)| \leq K\}$ are compact.

This case has already been considered by [3], using projective Hilbert metrics techniques. We will compute an explicit lower bound for the coupling constant $\varepsilon_{k,2|n}(\mathsf{X} \times \mathsf{X})$, and will then prove, under mild additional assumptions on the distribution of the Y's that the chain forgets its initial conditions geometrically fast. For $y \in \mathsf{Y}$, denote $C(y) \stackrel{\mathrm{def}}{=} \{x \in \mathsf{X}, |b(x)| \leq |y| + M\}$. Note that, for any $x \in \mathsf{X}$ and $A \in \mathscr{B}(\mathsf{X})$,

$$\mathrm{F}_{k|n}\mathrm{F}_{k+1|n}(x, A) = \frac{\iint q(x, x_{k+1}) g_{k+1}(x_{k+1}) q(x_{k+1}, x_{k+2}) g_{k+2}(x_{k+2}) \mathbb{1}_A(x_{k+2}) \beta_{k+2|n}(x_{k+2}) \mathrm{d}x_{k+1} \mathrm{d}x_{k+2}}{\iint q(x, x_{k+1}) g_{k+1}(x_{k+1}) q(x_{k+1}, x_{k+2}) g_{k+2}(x_{k+2}) \beta_{k+2|n}(x_{k+2}) \mathrm{d}x_{k+1} \mathrm{d}x_{k+2}} ,$$

where $g_{k+1}(x)$ is a shorthand notation for $g(x, Y_{k+1})$. Since q is continuous and positive, for any compact sets C and C', $\inf_{C\times C'} q(x, x') > 0$ and $\sup_{C\times C'} q(x, x') < \infty$. On the other hand, because the observation noise is bounded, $g(x, y) = g(x, y)\mathbb{1}_{C(y)}(x)$. Therefore,

$$\mathrm{F}_{k|n}\mathrm{F}_{k+1|n}(x, A) \geq \rho(Y_{k+1}, Y_{k+2}) \nu_{k|n}(A) ,$$

where

$$\rho(y, y') = \frac{\inf_{C(y)\times C(y')} q(x, x')}{\sup_{C(y)\times C(y')} q(x, x')} ,$$

and

$$\nu_{k|n}(A) \stackrel{\mathrm{def}}{=} \frac{\int g_{k+2}(x_{k+2}) \mathbb{1}_A(x_{k+2}) \beta_{k+2|n}(x_{k+2}) \nu(\mathrm{d}x_{k+2})}{\int g_{k+2}(x_{k+2}) \beta_{k+2|n}(x_{k+2}) \nu(\mathrm{d}x_{k+2})} .$$

This shows that the coupling constant of $\mathsf{X} \times \mathsf{X}$ is lower bounded by $\rho(Y_k, Y_{k+1})$. By applying Theorem 6, we obtain that

$$\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \leq 2 \prod_{k=0}^{\lfloor n/2 \rfloor} \{1 - \rho(Y_{2k}, Y_{2k+1})\} .$$

Hence, the posterior chain forgets its initial condition provided that

$$\liminf_{n\to\infty} \sum_{k=0}^{\lfloor n/2 \rfloor} \rho(Y_{2k}, Y_{2k+1}) = \infty \;, \quad \mathrm{P}^Y \ a.s. \; .$$

This property holds under many different assumptions on the observations $Y_{0:n}$.

To go beyond these examples, we have to find alternate verifiable conditions upon which we may control the coupling constant of the set $\mathsf{X}\times\mathsf{X}$. An interesting way of achieving this goal is to identify a uniformly accessible strong small set.

Definition 7 (Uniform accessibility). *Let j, ℓ, n be integers satisfying $\ell \geq 1$ and $j \in \{0,\dots,\lfloor n/\ell \rfloor\}$. A set C is uniformly accessible for the product of forward smoothing kernels $\mathrm{F}_{j|n} \dots \mathrm{F}_{j+\ell-1|n}$ if*

$$\inf_{x\in\mathsf{X}} \mathrm{F}_{j|n} \dots \mathrm{F}_{j+\ell-1|n}(x, C) > 0 \; . \tag{21}$$

Proposition 8. *Let k, ℓ, n be integers satisfying $\ell \geq 1$ and $k \in \{0,\dots,\lfloor n/\ell \rfloor - 1\}$. Assume that there exists a set C which is uniformly accessible for the forward smoothing kernels $\mathrm{F}_{k,\ell|n}$ and which is strongly small set for Q. Then, the coupling constant of $\mathsf{X}\times\mathsf{X}$ is lower bounded by*

$$\varepsilon_{k,\ell+1|n}(\mathsf{X}\times\mathsf{X}) \geq \frac{\sigma_-(C)}{\sigma_+(C)} \inf_{x\in\mathsf{X}} \mathrm{F}_{k(\ell+1)|n} \dots \mathrm{F}_{k(\ell+1)+\ell-1|n}(x, C) \; . \tag{22}$$

The proof is given in Section 6. Using this Proposition with Theorem 6 provides a mean to derive non-trivial rate of convergence, as illustrated in Example 4. The idea amounts to find conditions upon which a set is uniformly accessible. In the discussion below, it is assumed that the kernel Q has a density with respect to a σ-finite measure μ on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$, *i.e.,* for all $x \in \mathsf{X}$, $Q(x,\cdot)$ is absolutely continuous with respect to μ. For any set $A \in \mathscr{B}(\mathsf{X})$, define the function $\alpha : \mathsf{Y}^\ell \to [0,1]$

$$\alpha(y_{1:\ell}; A) \stackrel{\text{def}}{=} \inf_{x_0, x_{\ell+1}\in\mathsf{X}\times\mathsf{X}} \frac{W[y_{1:\ell}](x_0, x_{\ell+1}; A)}{W[y_{1:\ell}](x_0, x_{\ell+1}; \mathsf{X})} = \left\{1 + \tilde{\alpha}(y_{1:\ell}; A)\right\}^{-1} \;, \tag{23}$$

where we have set

$$W[y_{1:\ell}](x_0, x_{\ell+1}; A) \stackrel{\text{def}}{=} \int \cdots \int q(x_\ell, x_{\ell+1}) \mathbb{1}_A(x_\ell) \prod_{i=1}^{\ell} q(x_{i-1}, x_i) g(x_i, y_i) \mu(\mathrm{d}x_i) \;, \tag{24}$$

and

$$\tilde{\alpha}(y_{1:\ell}; A) \stackrel{\text{def}}{=} \sup_{x_0, x_{\ell+1}\in\mathsf{X}\times\mathsf{X}} \frac{W[y_{1:\ell}](x_0, x_{\ell+1}; A^c)}{W[y_{1:\ell}](x_0, x_{\ell+1}; A)} \; . \tag{25}$$

Of course, the situations of interest are when $\alpha(y_{1:\ell}; A)$ is positive or, equivalently, $\tilde{\alpha}(y_{1:\ell}; A) < \infty$. In such case, we may prove the following uniform accessibility condition:

Proposition 9. *For any integer n and any $j \in \{0,\dots,n-\ell\}$,*

$$\inf_{x\in\mathsf{X}} \mathrm{F}_{j|n} \cdots \mathrm{F}_{j+\ell-1|n}(x, C) \geq \alpha(Y_{j+1:j+\ell}; C) \; . \tag{26}$$

The proof is given in Section 6.

Example 4 (Functional autoregressive in noise). It is also of interest to consider cases where both the X's and the Y's are unbounded. We consider a non-linear non-Gaussian state space model (borrowed from [13, Example 5.8]). We assume that $X_0 \sim \xi$ and for $k \geq 1$,

$$X_k = a(X_{k-1}) + U_k \,,$$
$$Y_k = b(X_k) + V_k \,,$$

where $\{U_k\}$ and $\{V_k\}$ are two independent sequences of random variables, with probability densities $\bar{p}_U$ and $\bar{p}_V$ with respect to the Lebesgue measure on $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$. We denote by $|x|$ the norm of the vector x. In addition, we assume that

- For any $x \in \mathsf{X} = \mathbb{R}^d$, $\bar{p}_U(x) = p_U(|x|)$ where p_U is a bounded, bounded away from zero on $[0, M]$, is non increasing on $[M, \infty[$, and for some positive constant γ, and all $\alpha \geq 0$ and $\beta \geq 0$,
$$\frac{p_U(\alpha+\beta)}{p_U(\alpha)p_U(\beta)} \geq \gamma > 0 \,. \tag{27}$$
,
- the function a is Lipshitz, *i.e.,* there exists a positive constant a_+ such that $|a(x) - a(x')| \leq a_+|x - x'|$, for any $x, x' \in \mathsf{X}$,
- the function b is one-to-one differentiable and its Jacobian is bounded and bounded away from zero.
- For any $y \in \mathsf{Y} = \mathbb{R}^d$, $\bar{p}_V(y) = p_V(|y|)$ where p_V is a bounded positive lower semi-continuous function, p_V is non increasing on $[M, \infty[$, and satisfies
$$\Upsilon \stackrel{\text{def}}{=} \int_0^\infty [p_U(x)]^{-1} p_V(b_- x)[p_U(a_+ x)]^{-1} \mathrm{d}x < \infty \,, \tag{28}$$
where b_- is the lower bound for the Jacobian of the function b.

The condition on the state noise $\{U_k\}$ is satisfied by Pareto-type, exponential and logistic densities but obviously not by Gaussian density, because the tails are in such case too light.

The fact that the tails of the state noise U are heavier than the tails of the observation noise V (see (28)) plays a key role in the derivations that follow. In Section 5 we consider a case where this restriction is not needed (e.g., normal).

The following technical lemma (whose proof is postponed to section 7), shows that any set with finite diameter is a strong small set.

Lemma 10. *Assume that* $\mathrm{diam}(C) < \infty$. *Then, for all* $x_0 \in C$ *and* $x_1 \in \mathsf{X}$,

$$\varepsilon(C) h_C(x_1) \leq q(x_0, x_1) \leq \varepsilon^{-1}(C) h_C(x_1) \,, \tag{29}$$

with

$$\varepsilon(C) \stackrel{\text{def}}{=} \gamma p_U(\mathrm{diam}(C)) \wedge \inf_{u \leq \mathrm{diam}(C)+M} p_U(u) \wedge \left(\sup_{u \leq \mathrm{diam}(C)+M} p_U(u) \right)^{-1} \,, \tag{30}$$

$$h_C(x_1) \stackrel{\text{def}}{=} \mathbb{1}(d(x_1, a(C)) \leq M) + \mathbb{1}(d(x_1, a(C)) > M) p_U(|x_1 - a(z_0)|) \,, \tag{31}$$

where γ is defined in (27) *and z_0 is an arbitrary element of C. In addition, for all $x_0 \in \mathsf{X}$ and $x_1 \in C$,*

$$\nu(C)k_C(x_0) \le q(x_0, x_1) , \tag{32}$$

with

$$\nu(C) \stackrel{\text{def}}{=} \inf_{|u|\le \text{diam}(C)+M} p_U , \tag{33}$$

$$k_C(x_0) \stackrel{\text{def}}{=} \mathbb{1}(d(a(x_0), C) < M) + \mathbb{1}(d(a(x_0), C) \ge M) p_U(|z_1 - a(x_0)|) , \tag{34}$$

where z_1 is an arbitrary point in C.

By Lemma 10, the denominator of (25) is lower bounded by

$$W[y](x_0, x_2; C) \ge \varepsilon(C)\, \nu(C) k_C(x_0) h_C(x_2) \int_C g(x_1, y) \mathrm{d}x_1 , \tag{35}$$

where we have set $z_0 = b^{-1}(y)$ in the definition (31) of h_C and $z_1 = b^{-1}(y)$ in the definition (34) of k_C. Therefore, we may bound $\tilde{\alpha}(y_1, C)$, defined in (25), by

$$\tilde{\alpha}(y_1, C) \le \left(\varepsilon(C)\, \nu(C) \int_C g(x_1, y_1) \mathrm{d}x_1 \right)^{-1} I(y_1, C) \tag{36}$$

$$I(y_1, C) \stackrel{\text{def}}{=} \sup_{x_0, x_2 \in \mathsf{X}} \left([k_C(x_0)]^{-1} [h_C(x_2)]^{-1} W[y_1](x_0, x_2; C^c) \right) . \tag{37}$$

In the sequel, we set $C = C_K(y) \stackrel{\text{def}}{=} \{x, |x - b^{-1}(y)| \le K\}$, where K is a constant which will be chosen later. Since, by construction, the diameter of the set $C_K(y)$ is $2K$ uniformly with respect to y, the constants $\varepsilon(C_K(y))$ (defined in (30)) and $\nu(C_K(y))$ (defined in (33)) are functions of K only and are therefore uniformly bounded from below with respect to y. The following Lemma shows that, for K large enough, $\int_{C_K(y)} g(x_1, y) \mathrm{d}x_1$ is uniformly bounded from below:

Lemma 11.

$$\lim_{K\to\infty} \inf_{\mathsf{Y}} \int_{C_K(y)} g(x, y) \mathrm{d}x > 0 .$$

The proof is postponed to Section 7. The following Lemma shows that K may be chosen large enough so that $I(y, C_K(y))$ is uniformly bounded.

Lemma 12.

$$\limsup_{K\to\infty} \sup_{y\in\mathsf{Y}} I(y, C_K(y)) < \infty . \tag{38}$$

The proof is postponed to Section 7. Combining the previous results, $\tilde{\alpha}(y_1, C_K(y_1))$ is uniformly bounded in y_1 for large enough K, and therefore $\alpha(y_1, C_K(y_1))$ is uniformly bounded away from zero. Using Proposition 8 with $C = C_K(y)$ shows that the coupling constant of $\mathsf{X} \times \mathsf{X}$ is bounded away from zero uniformly in y. Hence, Proposition 6 shows that there exists $\varepsilon > 0$, such that for any probability measures ξ and ξ',

$$\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\text{TV}} \le 2(1-\varepsilon)^{\lfloor n/2 \rfloor} .$$

5 Pairwise drift conditions

5.1 The pair-wise drift condition

In the situations where coupling over the whole state-space leads to trivial result, one may still use the coupling argument, but this time over smaller sets. In such cases, however, we need a device to control the return time of the joint chain to the set where the two chains are allowed to couple. In this section we obtain results that are general enough to include the autoregression model with Gaussian innovations and Gaussian measurement error. Drift conditions are used to obtain bounds on the coupling time. Consider the following drift condition.

Definition 13 (Pair-wise drift conditions toward a set)**.** *Let n be an integer and $k \in \{0, \ldots, n-1\}$ and let $\bar{C}_{k|n}$ be a set valued function $\bar{C}_{k|n} : \mathsf{Y}^{n+1} \to \mathscr{B}(\mathsf{X}) \times \mathscr{B}(\mathsf{X})$. We say that the forward smoothing kernel $\mathrm{F}_{k|n}$ satisfies the* pair-wise drift condition *toward the set $\bar{C}_{k|n}$ if there exist functions $V_{k|n} : \mathsf{X} \times \mathsf{X} \times \mathsf{Y}^{n+1} \to \mathbb{R}$, $V_{k|n} \geq 1$, functions $\lambda_{k|n} : \mathsf{Y}^{n+1} \to [0,1)$, $\rho_{k|n} : \mathsf{Y}^{n+1} \to \mathbb{R}^+$ such that, for any sequence $y_{0:n} \in \mathsf{Y}^n$,*

$$\bar{\mathrm{R}}_{k|n} V_{k+1|n}(x,x') \leq \rho_{k|n} \qquad (x,x') \in \bar{C}_{k|n} \tag{39}$$

$$\bar{\mathrm{F}}_{k|n} V_{k+1|n}(x,x') \leq \lambda_{k|n} V_{k|n}(x,x') \qquad (x,x') \notin \bar{C}_{k|n} \,. \tag{40}$$

where $\bar{\mathrm{R}}_{k|n}$ is defined in (16) *and $\bar{\mathrm{F}}_{k|n}$ is defined in* (17).

We set $\varepsilon_{k|n} = \varepsilon_{k|n}(\bar{C}_{k|n})$, the coupling constant of the set $\bar{C}_{k|n}$, and we denote

$$B_{k|n} \stackrel{\text{def}}{=} 1 \vee \rho_{k|n}(1-\varepsilon_{k|n})\lambda_{k|n} \,. \tag{41}$$

For any vector $\{a_{i,n}\}_{1 \leq i \leq n}$, denotes by $[\downarrow a]_{(i,n)}$ the i-th largest order statistics, *i.e.*, $[\downarrow a]_{(1,n)} \geq [\downarrow a]_{(2,n)} \geq \cdots \geq [\downarrow a]_{(n,n)}$ and $[\uparrow a]_{(i,n)}$ the i-th smallest order statistics, *i.e.*, $[\uparrow a]_{(1,n)} \leq [\uparrow a]_{(2,n)} \leq \cdots \leq [\uparrow a]_{(n,n)}$.

Theorem 14. *Let n be an integer. Assume that for each $k \in \{0, \ldots, n-1\}$, there exists a set-valued function $\bar{C}_{k|n} : \mathsf{Y}^{n+1} \to \mathscr{B}(\mathsf{X}) \otimes \mathscr{B}(\mathsf{X})$ such that the forward smoothing kernel $\mathrm{F}_{k|n}$ satisfies the pairwise drift condition toward the set $\bar{C}_{k|n}$. Then, for any probability ξ, ξ' on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$,*

$$\left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \leq \min_{1 \leq m \leq n} A_{m,n} \tag{42}$$

where

$$A_{m,n} \stackrel{\text{def}}{=} \prod_{i=1}^{m} (1 - [\uparrow \varepsilon]_{(i|n)}) + \prod_{i=0}^{n} \lambda_{i|n} \prod_{i=0}^{m} [\downarrow B]_{(i|n)} \xi \otimes \xi'(V_0) \tag{43}$$

The proof is in section 6.

Corollary 15. *If there exists a sequence $\{m(n)\}$ of integers satisfying, $m(n) \leq n$ for any integer n, $\lim_{n\to\infty} m(n) = \infty$, and, P^Y-a.s.*

$$\limsup \left(\sum_{i=0}^{m(n)} \log(1 - [\uparrow \varepsilon]_{(i|n)}) + \sum_{i=0}^{n} \log \lambda_{i|n} + \sum_{i=0}^{m(n)} \log[\downarrow B_{(i,n)}] \right) = -\infty \,,$$

then

$$\limsup_n \left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \xrightarrow{\text{a.s.}} 0 \,, \quad \mathrm{P}^Y - a.s. \,.$$

Corollary 16. *If there exists a sequence $\{m(n)\}$ of integers such that $m(n) \leq n$ for any integer n, $\liminf m(n)/n = \alpha > 0$ and P^Y-a.s.*

$$\limsup \left(\frac{1}{n} \sum_{i=0}^{m(n)} \log(1 - [\uparrow \varepsilon]_{(i|n)}) + \frac{1}{n} \sum_{i=1}^{n} \log \lambda_{i|n} + \frac{1}{n} \sum_{i=1}^{n-m(n)} \log[\downarrow B_{(i|n)}] \right) \leq -\lambda \,,$$

then there exists $\nu \in (0,1)$ such that

$$\nu^{-n} \left\| \phi_{\xi,n} - \phi_{\xi',n} \right\|_{\mathrm{TV}} \xrightarrow{\text{a.s.}} 0 \,, \quad \mathrm{P}^Y - a.s. \;.$$

5.2 Examples

5.2.1 Gaussian autoregression

Let

$$X_i = \alpha X_{i-1} + \sigma U_i$$
$$Y_i = X_i + \tau V_i$$

where $|\alpha| < 1$ and $\{U_i\}_{i \geq 0}$ and $\{V_i\}$ are i.i.d. standard Gaussian and are independent from X_0. Let n be an integer and $k \in \{0, \ldots, n-1\}$. The backward functions are given by

$$\beta_{k|n}(x) \propto \exp\left(-(\alpha x - m_{k|n})^2/(2\rho^2_{k|n})\right) \,, \tag{44}$$

where $m_{k|n}$ and $\rho_{k|n}$ can be computed for $k = \{0, \ldots, n-2\}$ using the following backward recursions (see (6))

$$m_{k|n} = \frac{\rho^2_{k+1|n} Y_{k+1} + \alpha \tau^2 m_{k+1|n}}{\rho^2_{k+1|n} + \alpha^2 \tau^2} \,, \qquad \rho^2_{k|n} = \frac{(\tau^2 + \sigma^2)\rho^2_{k+1|n} + \alpha^2 \sigma^2 \tau^2}{\rho^2_{k+1|n} + \alpha^2 \tau^2} \,, \tag{45}$$

initialized with $m_{n-1|n} = Y_n$ and $\rho_{n-1|n} = \sigma^2 + \tau^2$. The forward smoothing kernel $\mathrm{F}_{i|n}(x, \cdot)$ has a density with respect to to the Lebesgue measure given by $\phi(\cdot; \mu_{i|n}(x), \gamma^2_{i|n})$, where $\phi(z; \mu, \sigma^2)$ is the density of a Gaussian random variable with mean μ and variance σ^2 and

$$\mu_{i|n}(x) = \frac{\tau^2 \rho^2_{i+1|n} \alpha x + \sigma^2 \rho^2_{i+1|n} Y_{i+1} + \sigma^2 \alpha \tau m_{i+1|n}}{(\sigma^2 + \tau^2)\rho^2_{i+1|n} + \tau^2 \alpha^2 \sigma^2} \,,$$

$$\gamma^2_{i|n} = \frac{\sigma^2 \tau^2 \rho^2_{i+1|n}}{(\tau^2 + \sigma^2)\rho^2_{i+1|n} + \alpha^2 \tau^2 \sigma^2} \,.$$

From (45), it follows that for any $i \in \{0, \ldots, n-1\}$, $\sigma^2 \leq \rho^2_{i|n} \leq \sigma^2 + \tau^2$. This implies that, for any $(x, x') \in \mathsf{X} \times \mathsf{X}$, and any $i \in \{0, \ldots, n-1\}$, the function $\mu_{i|n}$ is Lipshitz and with Lipshitz constant which is uniformly bounded by some $\beta < |\alpha|$,

$$|\mu_{i|n}(x) - \mu_{i|n}(x')| \leq \beta |x - x'| \,, \quad \beta \stackrel{\text{def}}{=} |\alpha| \frac{\tau^2(\sigma^2 + \tau^2)}{(\sigma^2 + \tau^2)^2 + \tau^2 \alpha^2 \sigma^2} \,, \tag{46}$$

and that the variance is uniformly bounded

$$\gamma_-^2 \stackrel{\text{def}}{=} \frac{\sigma^2\tau^2}{(1+\alpha^2)\tau^2+\sigma^2} \le \gamma_{i|n}^2 \le \gamma_+^2 \stackrel{\text{def}}{=} \frac{\sigma^2\tau^2(\sigma^2+\tau^2)}{(\tau^2+\sigma^2)^2+\alpha^2\tau^2\sigma^2} . \tag{47}$$

For $0 < c < \infty$, consider the set

$$C \stackrel{\text{def}}{=} \{(x,x') \in \mathsf{X}\times\mathsf{X} : \; |x-x'| \le c\} , \tag{48}$$

For any $i \in \{0,\dots,n-1\}$, the coupling constant of C is uniformly lower-bounded,

$$1 - \frac{1}{2}\left\|\mathrm{F}_{i|n}(x,\cdot) - \mathrm{F}_{i|n}(x',\cdot)\right\|_{\mathrm{TV}} = 1 - \mathrm{erf}\left(\gamma_{i|n}^{-1}|\mu_{i|n}(x) - \mu_{i|n}(x')|\right) \le 1 - \mathrm{erf}(\gamma_-^{-1}\beta c) ,$$

where erf is the error function. For c large enough, the drift condition is satisfied with $V(x,x') = 1 + (x-x')^2$:

$$\bar{\mathrm{F}}_{i|n} V(x,x') = 1 + \left\{\mu_{i|n}(x) - \mu_{i|n}(x')\right\}^2 + 2\gamma_{i|n}^2 \le 1 + \beta^2|x-x'|^2 + \gamma_+^2 .$$

The condition (39) is satisfied with

$$\rho_{i|n} \le \rho \stackrel{\text{def}}{=} \frac{\left(1+\beta^2c^2+\gamma_+^2\right)}{\mathrm{erf}(\gamma_-^{-1}\beta c)} , \tag{49}$$

where c is defined in (48). The condition (40) is satisfied with $\lambda_{i|n} = \tilde{\beta}^2$ for any $\tilde{\beta}$ and c satisfying $\beta < \tilde{\beta} < 1$ and $c^2 > (1-\tilde{\beta}^2+\gamma_+^2)/(\tilde{\beta}^2-\beta^2)$. All these bounds are uniform with respect to n, $i \in \{0,\dots,n-1\}$ and the observations $y_{0:n}$. Therefore, for any $m \in \{0,\dots,n\}$,

$$\left\|\phi_{\xi,n} - \phi_{\xi',n}\right\|_{\mathrm{TV}} \le \min_{1\le m\le n}\left\{(1-\varepsilon)^m + B^m\tilde{\beta}^{2n}\left(1 + 2\int \xi(dx)x^2 + 2\int \xi'(dx)x^2\right)\right\}$$

with $\varepsilon = \mathrm{erf}(\gamma_-^{-1}\beta c)$, $B = 1 \vee \rho(1-\varepsilon)\tilde{\beta}^2$ where ρ is defined in (49). Taking $m = [\delta n]$ for some $\delta > 0$ such that $B^\delta\tilde{\beta}^2 < 1$, this upper bound may be shown to go to zero exponentially fast and uniformly with respect to the observations $y_{0:n}$.

5.2.2 State space models with strongly unimodal distributions

The Gaussian example can be generalized to the more general case where the distribution of the state noise and the measurement noise are strongly unimodal. Recall that a density is strongly unimodal if the log of its density is concave.

Note that if f and g are two strongly unimodal density, then the density $h = fg/\int fg$ is also strongly unimodal, with mode that lies between the two modes; its second-order derivative of $\log h$ is smaller that the sum of the second-order derivative of $\log f$ and $\log g$. Let the state noise density be denoted by $p_U(\cdot) = \mathrm{e}^{\varphi(\cdot)}$ and that of the measurements' errors be $p_V(\cdot) = \mathrm{e}^{\psi(\cdot)}$. Define by the recursion operating on the decreasing indices

$$\bar{\beta}_{i|n}(x) = p_V(y_i - x)\int q(x, x_{i+1})\bar{\beta}_{i+1|n}(x_{i+1})\mathrm{d}x_{i+1} , \tag{50}$$

with the initial condition $\bar{\beta}_{n|n}(x) = p_V(y_n - x)$. These functions are the conditional distribution of the observations $Y_{i:n}$ given $X_i = x$. They are related to the backward function through the relation $\bar{\beta}_{i|n}(x) \stackrel{\text{def}}{=} \beta_{i|n}(x) p_V(y_i - x)$. We denote $\psi_{i|n}(x) \stackrel{\text{def}}{=} \log \bar{\beta}_{i|n}(x)$. Now,

$$\psi_{i|n}(x) = \psi(Y_i - x) + \log \int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\mathrm{d}z \ .$$

Under the stated assumptions, the forward smoothing kernel $\mathrm{F}_{i|n}$ has a density with respect to the Lebesgue measure which is given by

$$\mathrm{f}_{i|n}(x_i, x_{i+1}) = p_U(x_{i+1} - \alpha x_i)\bar{\beta}_{i+1|n}(x_{i+1}) \Big/ \int p_U(z - \alpha x_i)\bar{\beta}_{i+1|n}(z)\mathrm{d}z \ . \tag{51}$$

Denote by $\widetilde{\mathrm{Cov}}_{i|n,x}$ the covariance function with respect to the forward smoothing kernel density. We recall that for any probability distribution μ on $(\mathsf{X}, \mathscr{B}(\mathsf{X}))$ and any two increasing measurable functions f and g which are square integrable with respect to μ, the covariance of f and g with respect to μ, is non-negative. Hence,

$$\begin{aligned}
&\psi''_{i|n}(x) \\
&\quad = \psi''(Y_i - x) + \alpha^2 \frac{\int p''_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz}{\int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz} - \alpha^2 \Big(\frac{\int p'_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz}{\int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz} \Big)^2 \\
&\quad = \psi''(Y_i - x) - \alpha^2 \frac{\int p'_U(z - \alpha x)\bar{\beta}'_{i+1|n}(z)\,dz}{\int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz} \\
&\qquad\qquad + \alpha^2 \Big(\frac{\int p'_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz}{\int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz} \Big) \Big(\frac{\int p_U(z - \alpha x)\bar{\beta}'_{i+1|n}(z)\,dz}{\int p_U(z - \alpha x)\bar{\beta}_{i+1|n}(z)\,dz} \Big) \\
&\quad = \psi''(Y_i - x) - \alpha^2 \widetilde{\mathrm{Cov}}_{i|n,x}\left(\varphi'(\cdot - \alpha x), \psi'_{i+1|n}(\cdot)\right) \\
&\quad \leq \psi''(Y_i - x),
\end{aligned} \tag{52}$$

where we used a direct differentiation, integration by parts, and the fact that both ϕ' and $\psi'_{i+1|n}$ are monotone non-increasing functions (the last statement follows by applying (52) inductively from n backward).

We conclude that $\psi_{i|n}$ is strongly unimodal with curvature at least as that of the original likelihood function. Hence the curvature of the logarithm of the forward smoothing density is smaller than the sum of the curvature of the state and of the measurement noise,

$$\left[\log \mathrm{f}_{i|n}(x_i, x_{i+1})\right]'' \leq \varphi''(x_{i+1} - \alpha x_i) + \psi''(Y_{i+1} - x_{i+1}) \leq -c \ , \tag{53}$$

where

$$c = -\max_{x_{i+1}} \varphi''(x_{i+1}) - \max_{x_{i+1}} \psi''(x_{i+1}) \ . \tag{54}$$

Lemma 17 shows that the variance of X_{i+1} given X_i and $Y_{i+1:n}$ is uniformly bounded

$$v_{i|n}(x) \stackrel{\text{def}}{=} \int \left(x_{i+1} - \int x_{i+1}\mathrm{f}_{i|n}(x, x_{i+1})\mathrm{d}x_{i+1} \right)^2 \mathrm{f}_{i|n}(x, x_{i+1})\mathrm{d}x_{i+1} \leq c^{-1} \ .$$

where c is defined in (54). Now let

$$e_{i|n}(x) \stackrel{\text{def}}{=} \int x_{i+1} \mathrm{f}_{i|n}(x, x_{i+1}) \mathrm{d}x_{i+1} \; .$$

Similarly as above

$$\frac{\mathrm{d}e_{i|n}}{\mathrm{d}x}(x) = -\alpha \, \widetilde{\mathrm{Cov}}_{i|n,x} \left(Z, \varphi'(Z - \alpha x) \right) \; .$$

Note that $x_{i+1} \mapsto e_{i|n}(x) - x_{i+1}$, $x_{i+1} \mapsto \varphi'(x_{i+1} - \alpha x)$, and $x_{i+1} \mapsto \psi'_{i+1|n}(x_{i+1})$ are monotone non-increasing and therefore their correlation is positive with respect to any probability measure. Hence

$$\begin{aligned}
&\left| \frac{\mathrm{d}e_{i|n}}{\mathrm{d}x}(x) \right| \\
&= |\alpha| \frac{\int \left(e_{i|n}(x) - x_{i+1} \right) \varphi'(x_{i+1} - \alpha x) \mathrm{e}^{\varphi(x_{i+1} - \alpha x) + \psi_{i+1|n}(x_{i+1})} \mathrm{d}x_{i+1}}{\int \mathrm{e}^{\varphi(x_{i+1} - \alpha x) + \psi_{i+1|n}(x_{i+1})} \mathrm{d}x_{i+1}} \\
&\leq |\alpha| \frac{\int \left(e_{i|n}(x) - x_{i+1} \right) \left(\varphi'(x_{i+1} - \alpha x) + \psi'_{i+1|n}(x_{i+1}) \right) \mathrm{e}^{\phi(x_{i+1} - \alpha x) + \psi_{i+1|n}(x_{i+1})} \mathrm{d}x_{i+1}}{\int \mathrm{e}^{\varphi(x_{i+1} - \alpha x) + \psi_{i+1|n}(x_{i+1})} \mathrm{d}x_{i+1}} \\
&= |\alpha| \; .
\end{aligned}$$

by integration by parts. Put as before $V(x, x') = 1 + (x - x')^2$. It follows from the discussion above that

$$\bar{\mathrm{F}}_{i|n} V(x, x') = 1 + (e_{i|n}(x) - e_{i|n}(x'))^2 + v_{i|n}(x) + v_{i|n}(x') \; ,$$

where $v_{i|n}(x)$ and $v_{i|n}(x')$ are uniformly bounded with respect to x and x' and $|e_{i|n}(x) - e_{i|n}(x')| \leq \alpha |x - x'|$. The rest of the argument is like that for the normal-normal case.

We conclude the argument by stating and proving a lemma which was used above.

Lemma 17. *Suppose that Z is a random variable with probability density function f satisfying $\sup_x (\partial^2 / \partial x^2) \log f \leq -c$. Then, Z is square integrable and* $\mathrm{Var}(Z) \leq c^{-1}$.

Proof. Suppose, w.l.o.g., that the maximum of f is at 0. Under the stated assumption, there exist constants $a \geq 0$ and b such that $f(x) \leq a \mathrm{e}^{-c(x-b)^2}$. This implies that Z is square integrable. Denote $z \mapsto \zeta(z) = \log f(z) + cz^2/2$ which by assumption is a concave function. Let m be the mean of Z.

$$\begin{aligned}
\mathrm{E}[(Z - m)^2] = \int (z - m) z \mathrm{e}^{\xi(z) - cz^2/2} \mathrm{d}z = \\
c^{-1} \int (z - m) \left(cz - \xi'(z) \right) \mathrm{e}^{\xi(z) - cz^2/2} \mathrm{d}z + c^{-1} \int (z - m) \xi'(z) \mathrm{e}^{\xi(z) - cz^2/2} \mathrm{d}z.
\end{aligned}$$

By construction, $z \mapsto \xi'(z)$ is a non-increasing function. Since the inequality $\mathrm{Cov}(\varphi(Z), \psi(Z)) \geq 0$ holds for any two non-decreasing function φ and ψ which have finite second moment, the second term in the RHS of the previous equation is negative. Since $\left(cz - \xi'(z) \right) \mathrm{e}^{\xi(z) - cz^2/2} = -f'(z)$, the proof follows by integration by parts:

$$\mathrm{Var}(Z) \leq -c^{-1} \int (z - m) f'(z) \mathrm{d}z = c^{-1} \int f(z) \mathrm{d}z = c^{-1} \; .$$

□

6 Proofs

Proof of Proposition 5. The proof is similar to the one done in [11]. For $x \in C$, the condition, (15) implies that

$$\sigma_-(C) \le \frac{\mathrm{d}Q(x,\cdot)}{\mathrm{d}\nu_C}(dx') \le \sigma_+(C)\,.$$

Plugging the lower and upper bounds in the numerator and the denominator of (8) yields,

$$\mathrm{F}_{k|n}(x_k,A) \ge \frac{\sigma_-(C)}{\sigma_+(C)} \frac{\int_A \beta_{k+1|n}(x_{k+1})\nu_C(\mathrm{d}x_{k+1})}{\int_\mathsf{X} \beta_{k+1|n}(x_{k+1})\nu_C(\mathrm{d}x_{k+1})}$$

The result is established with

$$\nu_{k|n}(A) \stackrel{\text{def}}{=} \frac{\int_A \beta_{k+1|n}(x_{k+1})\mu(\mathrm{d}x_{k+1})}{\int_\mathsf{X} \beta_{k+1|n}(x_{k+1})\mu(\mathrm{d}x_{k+1})}\,.$$

□

Proof of Proposition 8. For any $x \in \mathsf{X}$, by the Chapman-Kolmogorov equations,

$$\mathrm{F}_{k,l+1|n}(x,A) = \int \mathrm{F}_{k(\ell+1)|n} \dots \mathrm{F}_{k(\ell+1)+\ell-1|n}(x,dx')\mathrm{F}_{k(\ell+1)+\ell|n}(x',A)$$

$$\ge \inf_{x\in\mathsf{X}} \mathrm{F}_{k(\ell+1)|n} \dots \mathrm{F}_{k(\ell+1)+\ell-1|n}(x,C) \inf_{x'\in C} \mathrm{F}_{k(\ell+1)+\ell|n}(x',A)\,.$$

The proof follows from Proposition 5, showing that the infimum in the last term is greater or equal that $\sigma_-(C)/\sigma_+(C)\nu_C(A)$. □

Proof of Proposition 9. For simplicity, the dependence of $W(;)$ in the observations is implicit. Also, we set $g_i(x) = g(x, y_i)$. For any $x_i \in \mathsf{X}$,

$$\begin{aligned}
&\mathrm{P}\left(X_{i+\ell} \in C \,\middle|\, X_i = x_i, Y_{1:n}\right) \\
&= \frac{\int \cdots \int W(x_i, x_{i+\ell+1}; C) g_{i+\ell+1}(x_{i+\ell+1})\beta_{i+\ell+1|n}(x_{i+\ell+1})\mu(\mathrm{d}x_{i+\ell+1})}{\int \cdots \int W(x_i, x_{i+\ell+1}; \mathsf{X}) g_{i+\ell+1}(x_{i+\ell+1})\beta_{i+\ell+1|n}(x_{i+\ell+1})\mu(\mathrm{d}x_{i+\ell+1})}\,, \\
&= \frac{\int \cdots \int \frac{W(x_i,x_{i+\ell+1};C)}{W(x_i,x_{i+\ell+1};\mathsf{X})} W(x_i, x_{i+\ell+1}; \mathsf{X}) g_{i+\ell+1}(x_{i+\ell+1})\beta_{i+\ell+1|n}(x_{i+\ell+1})\mu(\mathrm{d}x_{i+\ell+1})}{\int \cdots \int W(x_i, x_{i+\ell+1}; \mathsf{X}) g_{i+\ell+1}(x_{i+\ell+1})\beta_{i+\ell+1|n}(x_{i+\ell+1})\mu(\mathrm{d}x_{i+\ell+1})}\,,
\end{aligned}$$

where W is defined in (24). The proof is concluded by noting that, under the stated assumptions,

$$\inf_{(x_i,x_{i+\ell+1})\in\mathsf{X}\times\mathsf{X}} \frac{W(x_i, x_{i+\ell+1}; C)}{W(x_i, x_{i+\ell+1}; \mathsf{X})} \ge \alpha(Y_{i+1:i+\ell}; C)\,,$$

□

Proof of Theorem 14. For notational simplicity, we drop the dependence in the sample size n. Denote $N_n \stackrel{\text{def}}{=} \sum_{j=0}^n \mathbb{1}_{\bar{C}_j}(X_j, X'_j)$ and $\varepsilon_i \stackrel{\text{def}}{=} \varepsilon(\bar{C}_i)$. For any $m \in \{1, \dots, n+1\}$, we have:

$$\mathrm{P}^Y_{\xi,\xi',0}(T \ge n) \le \mathrm{P}^Y_{\xi,\xi',0}\left(T \ge n, N_{n-1} \ge m\right) + \mathrm{P}^Y_{\xi,\xi',0}\left(T \ge n, N_{n-1} < m\right) . \tag{55}$$

The first term on the RHS of the previous equation is the probability that we fail to couple the chains after at least m independent trial. It is bounded by

$$\mathrm{P}^Y_{\xi,\xi',0}\left(T \ge n, N_{n-1} \ge m\right) \le \prod_{i=0}^{m-1}\left(1 - [\uparrow \varepsilon]_{(i)}\right) . \tag{56}$$

where $[\uparrow \varepsilon]_{(i)}$ are the smallest-order statistics of $(\varepsilon_0, \dots, \varepsilon_{n-1})$. We consider now the second term in the RHS of (55). Set $B_j \stackrel{\text{def}}{=} 1 \vee \rho_j(1-\varepsilon_j)\lambda_j^{-1}$. On the event $\{N_{n-1} \le m-1\}$,

$$\prod_{j=0}^{n-1} B_j^{\mathbb{1}_{\bar{C}_j}(X_j, X'_j)} \le \prod_{j=0}^{m-2} [\downarrow B]_{(j)} ,$$

where $[\downarrow B]_{(j)}$ is the j-th largest order statistics of $B_1, \dots, B_n$. Hence,

$$\mathbb{1}\{N_{n-1} \le m-1\} \le \left(\prod_{j=0}^{n-1} B_j^{\mathbb{1}_{\bar{C}_j}(X_j, X'_j)}\right)^{-1} \prod_{j=0}^{m-2} [\downarrow B]_{(j)} ,$$

which implies that:

$$\mathrm{P}^Y_{\xi,\xi',0}\left(T \ge n, N_{n-1} < m\right) \le \prod_{j=0}^{n-1} \lambda_j \prod_{j=0}^{m-2} [\downarrow B]_{(j)} \, \mathrm{E}^Y_{\xi\otimes\xi'\otimes\delta_0}[M_n] \tag{57}$$

where, $M_0 = V_0(X_0, X'_0)\mathbb{1}\{d_0 = 0\}$ and for $k \in \{1, \dots, n\}$:

$$M_k \stackrel{\text{def}}{=} \left(\prod_{j=0}^{k-1} \lambda_j\right)^{-1} \prod_{j=0}^{k-1} B_j^{-\mathbb{1}_{\bar{C}_j}(X_j, X'_j)} V_k(X_k, X'_k)\mathbb{1}\{d_k = 0\} . \tag{58}$$

Since, by construction,

$$\mathrm{E}_{\xi,\xi',0}\left[V_{k+1}(X_{k+1}, X'_{k+1})\mathbb{1}\{d_{k+1} = 0\} \,\middle|\, \mathscr{F}_k\right]$$
$$(1-\varepsilon_k)\bar{R}_k V_k(X_k, X'_k)\mathbb{1}_{\bar{C}_k^c}(X_k, X'_k) + \lambda_k V_k(X_k, X'_k)\mathbb{1}_{\bar{C}_k}(X_k, X'_k) ,$$

it is easily shown that $(M_k, k \ge 0)$ is a $(\mathscr{F}, \mathrm{P}^Y_{\xi,\xi',0})$-supermartingale w.r.t. where $\mathscr{F} \stackrel{\text{def}}{=} (\mathscr{F}_k)_{1 \le k \le n}$ with for $k \ge 0$, $\mathscr{F}_k \stackrel{\text{def}}{=} \sigma\left[(X_j, X'_j, d_j), 0 \le j \le k\right]$. Therefore,

$$\mathrm{E}^Y_{\xi,\xi',0}(M_n) \le \mathrm{E}^Y_{\xi,\xi',0}(M_0) = \xi \otimes \xi'(V_0) .$$

This establishes (42) and concludes the proof.

□

7 Proofs of Example 4

To simplify the notations, the dependence of $C(y)$ in K is implicit throughout the section.

Proof of Lemma 10. Consider first the case $d(x_1, a(C)) \geq M$. For any $z_1 \in a(C)$,

$$M \leq |x_1 - a(x_0)| \leq |x_1 - z_1| + |z_1 - a(x_0)| \leq \mathrm{diam}(C) + |x_1 - z_1| \,,$$
$$M \leq |x_1 - z_1| \leq |x_1 - a(x_0)| + |z_1 - a(x_0)| \leq \mathrm{diam}(C) + |x_1 - a(x_0)| \,.$$

Using that p_U is non-increasing for $u \geq M$ and (27), we obtain

$$p_U(|x_1 - a(x_0)|) \geq p_U(\mathrm{diam}(C) + |x_1 - z_1|) \geq \gamma p_U(\mathrm{diam}(C)) p_U(|x_1 - z_1|) \,,$$

and similarly,

$$p_U(|x_1 - z_1|) \geq \gamma p_U(\mathrm{diam}(C)) p_U(|x_1 - a(x_0)|) \,,$$

which establishes that (29) holds when $d(x_0, a(C)) \geq M$.

Consider now the case $d(x_1, a(C)) \leq M$. Since x_0 belongs to C, then $|x_1 - a(x_0)| \leq M + \mathrm{diam}(C)$, which implies that

$$\inf_{u \leq M + \mathrm{diam}(C)} p_U(u) \leq p_U(|x_1 - a(x_0)|) \leq \sup_{u \leq M + \mathrm{diam}(C)} p_U(u) \,,$$

(29) holds for $d(x_1, a(C)) \leq M$.

Consider now the second assertion. Assume first that x_0 is such that $d(a(x_0), C) \geq M$ and let z_1 be an arbitrary point of C. Then, for any $x_1 \in C$,

$$M \leq |x_1 - a(x_0)| \leq |x_1 - z_1| + |z_1 - a(x_0)| \leq \mathrm{diam}(C) + |z_1 - a(x_0)| \,.$$

Using that p_U is monotone decreasing on $[M, \infty)$ and (27),

$$\begin{aligned} p_U(|x_1 - a(x_0)|) &\geq p_U(\mathrm{diam}(C) + |z_1 - a(x_0)|) \\ &\geq \gamma p_U[\mathrm{diam}(C)] p_U(|z_1 - a(x_0)|) \,. \end{aligned} \tag{59}$$

If $d(a(x_0), C) \leq M$, then for any $x_1 \in C$, $|x_1 - a(x_0)| \leq \mathrm{diam}(C) + M$, so that

$$\inf_{|u| \leq \mathrm{diam}(C) + M} p_U \leq p_U\left(|x_1 - a(x_0)|\right) \,. \tag{60}$$

□

Proof of Lemma 11. Let $b_1^{-1} > 0$ be a lower bound for the Jacobian of b and choose K such that $b_1^{-1} K \geq M$. If $|b^{-1}(y) - x| \geq K$, then,

$$|y - b(x)| = |b(b^{-1}(y)) - b(x)| \geq b_1^{-1} |b^{-1}(y) - x| \geq M \,, \tag{61}$$

and since p_V is non-increasing on the interval $[M, \infty[$, the following inequality holds

$$\begin{aligned} \int_{|x - b^{-1}(y)| \geq K} p_V(|y - b(x)|) \mathrm{d}x &\leq \int_{|x - b^{-1}(y)| \geq K} p_V(b_1^{-1} |b^{-1}(y) - x|) \mathrm{d}x \\ &\leq \int_{|x| > K}^{\infty} p_V(b_1^{-1} |x|) \mathrm{d}x \,. \end{aligned}$$

Since the Jacobian of b is bounded, $\int p_V(|y-b(x)|)\mathrm{d}x$ is bounded away from zero by change of variables. The proof follows.

$\square$

Proof of Lemma 12. We will establish the results by considering independently the following cases:

1. For any y and any (x_0, x_2) such that $d(a(x_0), C(y)) \le M$ and $d(x_2, a[C(y)]) \le M$,

$$I(x_0, x_2; y) \le \left(\sup_{\mathbb{R}} p_U\right)^2 .$$

2. For any y and any (x_0, x_2) such that $d(a(x_0), C(y)) > M$ and $d(x_2, a[C(y)]) \le M$,

$$I(x_0, x_2; y) \le \gamma^{-1} \ (\sup p_U) \int_K^\infty [p_U(x)]^{-1} p_V(b_- x) dx \,.$$

3. For any y and any (x_0, x_2) such that $d(a(x_0), C(y)) \le M$ and $d(x_2, a[C(y)]) > M$

$$I(x_0, x_2; y) \le \gamma^{-1} (\sup p_U) \left\{ b_-^{-1} + \int_K^\infty p_V(b_- x) [p_U(a_+ x)]^{-1} \mathrm{d}x \right\}$$

4. For any y and any (x_0, x_2) such that $d(a(x_0), C(y)) > M$ and $d(x_2, a[C(y)]) > M$,

$$I(x_0, x_2; y) \le \gamma^{-2} \times \int_K^\infty [p_U(x)]^{-1} p_V(b_- x) \left\{ \left(\inf_{u \le M} p_U(u) \right)^{-1} + [p_U(a_+ x)]^{-1} \right\} \mathrm{d}x \,.$$

Proof of Assertion 1. On the set $\{x_0, d(a(x_0), C(y)) \le M\}$, $k_{C(y)}(x_0) \equiv 1$; On the set $\{x_2, d(x_2, a[C(y)]) \le M\}$, $h_{C(y)}(x_2) \equiv 1$. Since p_U is uniformly bounded, the bound follows from Lemma 11 and the choice of K. $\square$

Proof of Assertion 2. On the set $\{x_0, d(a(x_0), C(y)) > M\}$, $k_C(x_0) = p_U(|b^{-1}(y) - a(x_0)|)$; On the set $\{x_2, d(x_2, a[C(y)]) \le M\}$, $h_C(x_2) \equiv 1$. Therefore, for such (x_0, x_2),

$$I(x_0, x_2; y) \le (\sup p_U) \quad p_U^{-1}(|b^{-1}(y) - a(x_0)|) \int_{C^c(y)} p_U(|x_1 - a(x_0)|) p_V(|y_1 - b(x_1)|) \mathrm{d}x_1 \,. \quad (62)$$

We set $\alpha = x_1 - a(x_0)$ and $\beta = b^{-1}(y) - x_1$. Note that $|\alpha + \beta| = |b^{-1}(y) - a(x_0)| \ge d(a(x_0), C(y)) > M$. Since p_U is non-increasing on $[M, \infty[$, $p_U(|\alpha+\beta|) \ge p_U(|\alpha| + |\beta|)$, and the condition (27) shows that $\left(p_U(|\alpha+\beta|)\right)^{-1} p_U(|\alpha|) \le \gamma^{-1} p_U^{-1}(|\beta|)$ which implies

$$p_U^{-1}(|b^{-1}(y) - a(x_0)|) p_U(|x_1 - a(x_0)|) \le \gamma^{-1} p_U^{-1}(|b^{-1}(y) - x_1|) \,. \quad (63)$$

Therefore, plugging (63) into the RHS of (62) yields

$$
\begin{aligned}
I(x_0,x_2;y) &\le \gamma^{-1}\left(\sup p_U\right)\int_{|x_1-b^{-1}(y)|\ge K} p_U^{-1}(|b^{-1}(y)-x_1|)p_V(b_-|b^{-1}(y)-x|)\mathrm{d}x_1\\
&\le \gamma^{-1}\left(\sup p_U\right)\int_K^{\infty} p_U^{-1}(x)p_V(b_-x)\mathrm{d}x\,.
\end{aligned}
$$

□

Proof of Assertion 3. On the set $\{x_0, d(a(x_0),C(y)) \le M\}$, $k_C(x_0) \equiv 1$; on the set $\{x_2, d(x_2, a[C(y)]) > M\}$, $h_C(x_2) = p_U(|x_2 - a[b^{-1}(y)]|) \equiv 1$. Therefore, for such (x_0,x_2);

$$
\begin{aligned}
&I(x_0,x_2;y) \le \left(\sup p_U\right)\\
&\qquad\times p_U^{-1}(|x_2-a[b^{-1}(y)]|)\int_{C^c(y)} p_V(|y-b(x_1)|)p_U(|x_2-a(x_1)|)\mathrm{d}x_1\,. \quad (64)
\end{aligned}
$$

We set $\alpha = x_2 - a(x_1)$, $\beta = a(x_1) - a[b^{-1}(y)]$. Since $|\alpha+\beta| \ge d(x_2,a[C(y)]) > M$, using as above that $\left(p_U(|\alpha+\beta|)\right)^{-1} p_U(|\alpha|) \le \gamma^{-1} p_U^{-1}(|\beta|)$, we show

$$
p_U^{-1}(|x_2-a[b^{-1}(y)]|)p_U(|x_2-a(x_1)|) \le \gamma^{-1}p_U^{-1}(|a(x_1)-a[b^{-1}(y)]|)\,. \quad (65)
$$

Since for any $x, x' \in \mathsf{X}$,

$$
\begin{aligned}
p_U^{-1}(|a(x)-a(x')|) \le &\left(\inf_{u\le M} p_U(u)\right)^{-1}\mathbb{1}\{|a(x)-a(x')|\le M\}\\
&+p_U^{-1}(a_+|x-x'|)\mathbb{1}\{|a(x)-a(x')|>M\}\,, \quad (66)
\end{aligned}
$$

the RHS of (64) is therefore bounded by

$$
\begin{aligned}
I(x_0,x_2;y) \le \gamma^{-1}\left(\sup p_U\right)&\int_{|x_1-b^{-1}(y)|\ge K} p_V(b_-(|x_1-b^{-1}(y)|))\\
&\left\{\left(\inf_{u\le M} p_U(u)\right)^{-1} + p_U^{-1}(a_+|x_1-b^{-1}(y)|)\right\}\mathrm{d}x_1\,.
\end{aligned}
$$

□

Proof of Assertion 4. On the set $\{x_0, d(a(x_0),C(y)) > M\}$, $k_{C(y)}(x_0) = p_U(|b^{-1}(y) - a(x_0)|)$. On the set $\{x_2, d(x_2, a[C(y)]) > M\}$, $k_{C(y)}(x_2) = p_U(|x_2 - a[b^{-1}(y)]|)$. Therefore, for such (x_0,x_2),

$$
\begin{aligned}
&I(x_0,x_2;y) \le p_U^{-1}(|b^{-1}(y)-a(x_0)|)p_U^{-1}(|x_2-a[b^{-1}(y)]|)\\
&\qquad\times\int_{C^c(y)} p_U(|x_1-a(x_0)|)p_V(|y-b(x_1)|)p_U(|x_2-a(x_1)|)\mathrm{d}x_1\,. \quad (67)
\end{aligned}
$$

Using (61), (63), (65), and (66), the RHS of the previous equation is bounded by

$$
I(x_0,x_2;y) \le \gamma^{-2}\int_K^{\infty} p_U^{-1}(|x|)p_V(b_-|x|)\left\{\left(\inf_{u\le M} p_U(u)\right)^{-1} + p_U^{-1}(a_+x)\right\}\mathrm{d}x\,.
$$

The proof follows. □

□

References

[1] Atar, R. and Zeitouni, O. (1997). Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.* **33**, 6, 697–725.

[2] Baum, L. E., Petrie, T. P., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41**, 1, 164–171.

[3] Budhiraja, A. and Ocone, D. (1997). Exponential stability of discrete-time filters for bounded observation noise. *Systems Control Lett. 30*, 185–193.

[4] Budhiraja, A. and Ocone, D. (1999). Exponential stability in discrete-time filtering for non-ergodic signals. *Stochastic Process. Appl.* **82**, 2, 245–257.

[5] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer. http://www.tsi.enst.fr/ cappe/ihmm/.

[6] Chigansky, P. and Lipster, R. (2004). Stability of nonlinear filters in nonmixing case. *Ann. Appl. Probab.* **14**, 4, 2038–2056.

[7] Chigansky, P. and Liptser, R. (2006). On a role of predictor in the filtering stability. *Electron. Comm. Probab. 11*, 129–140 (electronic). MR2240706K

[8] Del Moral, P. (2004). *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer.

[9] Del Moral, P. and Guionnet, A. (1998). Large deviations for interacting particle systems: applications to non-linear filtering. *Stoch. Proc. App. 78*, 69–95.

[10] Del Moral, P., Ledoux, M., and Miclo, L. (2003). On contraction properties of Markov kernels. *Probab. Theory Related Fields* **126**, 3, 395–420.

[11] Douc, R., Moulines, E., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.* **32**, 5, 2254–2304.

[12] Griffeath, D. (1974/75). A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 31*, 95–106. MR0370771

[13] LeGland, F. and Oudjane, N. (2003). A robustification approach to stability and to uniform particle approximation of nonlinear filters: the example of pseudo-mixing signals. *Stochastic Process. Appl.* **106**, 2, 279–316.

[14] Lindvall, T. (1992). *Lectures on the Coupling Method*. Wiley, New-York.

[15] Ocone, D. and Pardoux, E. (1996). Asymptotic stability of the optimal filter with respect to its initial condition. *SIAM J. Control 34*, 226–243.

[16] Oudjane, N. and Rubenthaler, S. (2005). Stability and uniform particle approximation of nonlinear filters in case of non ergodic signals. *Stoch. Anal. Appl.* **23**, 3, 421–448.

[17] Thorisson, H. (2000). *Coupling, Stationarity and Regeneration*. Probability and its Applications. Springer-Verlag, New-York.