Math. Control Signals Systems (2000) 13: 63–93
© 2000 Springer-Verlag London Limited

Mathematics of Control, Signals, and Systems

Exponential Forgetting and Geometric Ergodicity in Hidden Markov Models*

François Le Gland† and Laurent Mevel†

Abstract. We consider a hidden Markov model with multidimensional observations, and with misspecification, i.e., the *assumed* coefficients (transition probability matrix and observation conditional densities) are possibly different from the *true* coefficients. Under mild assumptions on the coefficients of both the true and the assumed models, we prove that: (i) the prediction filter, and its gradient with respect to some parameter in the model, forget almost surely their initial condition exponentially fast, and (ii) the extended Markov chain, whose components are the unobserved Markov chain, the observation sequence, the prediction filter, and its gradient, is geometrically ergodic and has a unique invariant probability distribution.

Key words. HMM, Misspecified model, Prediction filter, Exponential forgetting, Geometric ergodicity, Product of random matrices.

1. Introduction

This paper is concerned with large time asymptotic properties in hidden Markov models (HMMs), i.e., in partially observed stochastic systems of the following fairly general form.

Let $\{X_n, n \geq 0\}$ and $\{Y_n, n \geq 0\}$ be two random sequences defined on the probability space $(\Omega, \mathscr{F}, \mathbf{P}_\bullet)$, with values in the finite set $S = \{1, \ldots, N\}$ and in $\mathbb{R}^d$, respectively. It is assumed that:

- The unobserved state sequence $\{X_n, n \geq 0\}$ is a time-homogeneous Markov chain with transition probability matrix $Q_\bullet = (q_\bullet^{i,j})$, i.e., for any integer $n \geq 0$, and for any $i, j \in S$,

$$\mathbf{P}_\bullet[X_{n+1} = j \mid X_n = i] = q_\bullet^{i,j},$$

* Date received: May 14, 1998. Date revised: November 18, 1999. This work was partially supported by the Commission of the European Communities, under the SCIENCE project System Identification, Project Number SC1*-CT92-0779, and under the HCM project Statistical Inference for Stochastic Processes, Project Number CHRX-CT92-0078, and by the Army Research Office, under Grant DAAH04-95-1-0164.
† IRISA/INRIA, Campus de Beaulieu, 35042 Rennes Cédex, France. {legland,lmevel}@irisa.fr.

and initial probability distribution $p_\bullet = (p_\bullet^i)$, i.e., for any $i \in S$,

$$\mathbf{P}_\bullet[X_0 = i] = p_\bullet^i.$$

- The observations $\{Y_n, n \geq 0\}$ are mutually independent given the sequence of states of the Markov chain, i.e., for any integer $n \geq 0$, and for any $i_0, \ldots, i_n \in S$,

$$\mathbf{P}_\bullet[Y_n \in dy_n, \ldots, Y_0 \in dy_0 \mid X_n = i_n, \ldots, X_0 = i_0] = \prod_{k=0}^{n} \mathbf{P}_\bullet[Y_k \in dy_k \mid X_k = i_k].$$

For any integer $n \geq 0$, and for any $i \in S$, the conditional probability distribution of the observation Y_n given that $(X_n = i)$ is absolutely continuous with respect to a nonnegative and σ-finite measure λ on $\mathbb{R}^d$, i.e.,

$$\mathbf{P}_\bullet[Y_n \in dy \mid X_n = i] = b_\bullet^i(y)\lambda(dy),$$

with a positive density. For any $y \in \mathbb{R}^d$, let

$$b_\bullet(y) = (b_\bullet^i(y)) \quad \text{and} \quad B_\bullet(y) = \mathrm{diag}(b_\bullet^i(y)).$$

Example 1.1 (Conditionally Gaussian Observations). Assume that the observations are of the form

$$Y_n = h_\bullet(X_n) + V_n,$$

for any integer $n \geq 0$, where $\{V_n, n \geq 0\}$ is a Gaussian white noise sequence independent of $\{X_n, n \geq 0\}$, with identity covariance matrix. The mapping $h_\bullet$ from S to $\mathbb{R}^d$ is equivalently defined as $h_\bullet = (h_\bullet^i)$ where $h_\bullet^i \in \mathbb{R}^d$ for any $i \in S$. In this case, λ is the Lebesgue measure on $\mathbb{R}^d$, the mutual independence condition is satisfied, and

$$b_\bullet^i(y) = (2\pi)^{-d/2} \exp\{-\tfrac{1}{2}|y - h_\bullet^i|^2\},$$

for any $i \in S$, and any $y \in \mathbb{R}^d$. Here and throughout the paper the notation $|\cdot|$ denotes the Euclidean norm.

Throughout the paper we make the following assumption:

Assumption A. The stochastic matrix $Q_\bullet = (q_\bullet^{i,j})$ is primitive.

Remark 1.2. Under Assumption A, there exist constants $0 < \rho_\bullet < 1$ and $K_\bullet \geq 1$ such that, for any integer $n \geq 1$,

$$\tfrac{1}{2} \max_{i,i' \in S} \sum_{j \in S} |q_{\bullet,n}^{i,j} - q_{\bullet,n}^{i',j}| \leq K_\bullet \rho_\bullet^n, \tag{1}$$

and the Markov chain $\{X_n, n \geq 0\}$ is geometrically ergodic, with a unique invariant probability distribution $\mu_\bullet = (\mu_\bullet^i)$ on S. Here and throughout the paper $(q_{\bullet,n}^{i,j})$ denote the entries of the stochastic matrix $Q_\bullet^n$, i.e., for any $i, j \in S$,

$$\mathbf{P}_\bullet[X_n = j \mid X_0 = i] = q_{\bullet,n}^{i,j}.$$

For any integer $n \geq 1$, let $p_n^\bullet = (p_n^i)$ denote the *prediction filter*, i.e., the conditional probability distribution under $\mathbf{P}_\bullet$ of the state X_n given observations $(Y_0, \ldots, Y_{n-1})$: for any $i \in S$,

$$p_n^i = \mathbf{P}_\bullet[X_n = i \mid Y_0, \ldots, Y_{n-1}].$$

The random sequence $\{p_n^\bullet, n \geq 0\}$ takes values in the set $\mathscr{P} = \mathscr{P}(S)$ of probability distributions over the finite set S, and satisfies the forward Baum equation

$$p_{n+1}^\bullet = \frac{Q_\bullet^* B_\bullet(Y_n) p_n^\bullet}{b_\bullet^*(Y_n) p_n^\bullet}, \tag{2}$$

for any integer $n \geq 0$, with initial condition $p_0^\bullet = p_\bullet$. Here and throughout the paper the notation * denotes the transpose of a matrix or a vector.

In practice, the transition probability matrix $Q_\bullet$ and the initial probability distribution $p_\bullet$ of the unobserved Markov chain $\{X_n, n \geq 0\}$, and the vector $b_\bullet$ of conditional densities of the observation sequence $\{Y_n, n \geq 0\}$ are possibly unknown. For this reason, we consider instead of (2) the more general equation

$$p_{n+1} = \frac{Q^* B(Y_n) p_n}{b^*(Y_n) p_n} = f[Y_n, p_n], \tag{3}$$

for any integer $n \geq 0$, with initial condition $p_0 = p$, where $Q = (q^{i,j})$ is an $N \times N$ stochastic matrix, $p = (p^i)$ is a probability vector on S, and $b = (b^i)$ is a vector of positive densities on $\mathbb{R}^d$. To make explicit the dependency with respect to the initial condition and the observations, we introduce the notation

$$p_{n+1} = f[Y_n, \ldots, Y_m, p_m],$$

for any integers n, m such that $n \geq m$.

Notice that these misspecification issues are of a different nature. We expect that a wrong initial condition for the prediction filter is rapidly forgotten, so that we could use any initial condition with practically the same effect. On the other hand, we expect that two different transition probability matrices and two different vectors of observation conditional densities will produce two significantly different observation sequences, so that we could estimate the unknown transition probability matrix and the unknown vector of observation conditional densities accurately, by accumulating observations. Indeed, it can be shown that the log-likelihood function for the estimation of the unknown transition probability matrix and the unknown vector of observation conditional densities, can be expressed as an additive functional of the extended Markov chain $\{Z_n = (X_n, Y_n, p_n), n \geq 0\}$. Consequently, an explicit expression can be obtained for the corresponding Kullback–Leibler information provided some ergodicity property holds. This and other statistical applications mentioned below will be found in a forthcoming work on identification of HMMs, and have been announced in [LM2].

Here is a short overview of the results obtained in the first part of this paper, which have been announced in [LM3]:

Exponential forgetting. In Theorem 2.1 we obtain an explicit bound for the Lipschitz constant of the solution map associated with (3). A similar result has been obtained in Proposition 2.1 of [LM1] for the time-dependent case. This nonlogarithmic and nonasymptotic bound goes to zero at exponential rate as time goes to infinity, and as a consequence in Theorem 2.2 we obtain an upper bound for the $\mathbf{P}_\bullet$-a.s. exponential rate of forgetting of the initial condition for (3). A statistical application is to prove that the prediction filter, hence the log-likelihood function, is Lipschitz continuous with respect to some parameter in the model, uniformly in time.

Geometric ergodicity. In Theorem 3.5 we prove the geometric ergodicity of the extended Markov chain $\{Z_n = (X_n, Y_n, p_n), n \geq 0\}$. We prove also the existence of a unique invariant probability distribution, and the existence of a solution to the associated Poisson equation. From this point, the law of large numbers and the central limit theorem can be proved for the extended Markov chain $\{Z_n = (X_n, Y_n, p_n), n \geq 0\}$. A statistical application is to obtain an explicit expression for the $\mathbf{P}_\bullet$-a.s. limit of the log-likelihood function (suitably normalized), i.e., for the Kullback–Leibler information.

Notice that

$$f[y_n, \ldots, y_m, p] = \frac{Q^* B(y_n) \cdots Q^* B(y_m) p}{e^*[Q^* B(y_n) \cdots Q^* B(y_m) p]} = \frac{M_{n,m} p}{e^* M_{n,m} p} = M_{n,m} \cdot p,$$

where $\cdot$ denotes the projective product, and where $e = (1, \ldots, 1)^*$ denotes the N-dimensional vector with all entries equal to 1, provided we define

$$M_{n,m} = Q^* B(y_n) \cdots Q^* B(y_m),$$

and our results will be based on auxiliary estimates for products of column-allowable nonnegative matrices, which improve earlier estimates obtained by Furstenberg and Kesten [FK], Kaijser [K1], and Arapostathis and Marcus [AM]. Our results are stated and proved in the companion paper [LM4].

The large time asymptotic properties of products of random matrices with Markovian dependence is thoroughly studied by Guivarc'h [G] and Bougerol [B1], [B2], see also additional references therein. In these works, the existence and properties of the Lyapunov spectrum are proved under the assumption that all the matrices considered are invertible, so that products of random matrices can be seen as random walks on a group, the group of invertible matrices.

In [H] the invertibility assumption is replaced by a positivity assumption, which is much more natural in the context of HMMs. Indeed, for an HMM without misspecification, with finite state space, positive transition probability matrix, and conditionally Gaussian observations, the exact exponential rate of forgetting is expressed in Corollary 2.1 of [AZ] as the difference between the two top Lyapunov exponents of (2). However, no explicit expression is available in general for these Lyapunov exponents, and estimates for the exponential rate are given in Theorems 1.3 and 1.4 of [AZ] when the signal-to-noise ratio is large or small, respectively.

It must be pointed out that our model is very general, in the sense that we consider a misspecified HMM, with primitive transition probability matrices (for both the true and the assumed models) and arbitrary observation conditional densities. Another important issue is that we not only obtain explicit bounds for the exponential rate of forgetting, but we also obtain nonlogarithmic and nonasymptotic bounds. Our proof of the geometric ergodicity of the extended Markov chain is based on these explicit bounds.

Concerning the existence of an invariant probability distribution, most results found in the literature and discussed below are formulated for the filter, i.e., for the probability distribution of the state X_n given the observations $(Y_0, \dots, Y_n)$, rather than for the prediction filter introduced above, but this does not make any significant difference. Our motivation for considering the prediction filter comes from the same statistical applications as in [AM].

For an HMM without misspecification, the optimal, i.e., correctly initialized, filter is a Markov chain under the probability measure $\mathbf{P}_\bullet$, and the existence of an invariant probability distribution is proved by Kunita [K2] in the case of a compact state space, and generalized by Stettner [S1]. Another early result can be found in [K1] in the special case of a finite state space and *noise-free* observations, under the additional assumption of *subrectangularity*. Under the probability measure $\mathbf{P}_\bullet$, the wrongly initialized filter is not a Markov chain, but the pair (state, filter) is a Markov chain, and the existence of an invariant probability distribution for the pair is proved by Stettner [S2], in the special case of a finite state space S, primitive transition probability matrix, and one-dimensional conditionally Gaussian observations, under the additional assumption that the mapping $h_\bullet = (h_\bullet^i)$ from S to $\mathbb{R}$ is *injective*.

Similarly for a misspecified HMM, the filter is not a Markov chain under the probability measure $\mathbf{P}_\bullet$, but the pair (state, wrong filter) is a Markov chain, and the existence of an invariant probability distribution for the pair is proved by Di Masi and Stettner [DS] in the special case of a finite state space S, positive transition probability matrices (for both the true and the assumed models), and d-dimensional conditionally Gaussian observations, under the assumption that the mapping $h_\bullet = (h_\bullet^i)$ from S to $\mathbb{R}^d$ is *injective* with $d = |S|$.

Closest to our results are those obtained by Arapostathis and Marcus [AM], where the existence of an invariant probability distribution for the triple (state, observation, wrong prediction filter) is proved in the special case of a finite state space, positive transition probability matrices (for both the true and the assumed model), and *binary* observations. Our methods of proof are also very close.

In the second part of this paper, with a view toward identification of HMMs we consider, as in [AM], the following more general problem. Assume that the coefficients of (3), i.e., the transition probability matrix Q and the vector b of observation conditional densities, depend on some one-dimensional parameter. Differentiating (3) with respect to this parameter yields

$$\partial p_{n+1} = Q^* \left[I - \frac{B(Y_n) p_n e^*}{b^*(Y_n) p_n} \right] \frac{B(Y_n)\ \partial p_n}{b^*(Y_n) p_n} + \partial f[Y_n, p_n],$$

where

$$\partial f[y,p] = \frac{\partial Q^* B(y)p}{b^*(y)p} + Q^* \left[I - \frac{B(y)pe^*}{b^*(y)p} \right] \frac{\partial B(y)p}{b^*(y)p},$$

for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$. More generally, we consider the following linear equation:

$$w_{n+1} = Q^* \left[I - \frac{B(Y_n)p_n e^*}{b^*(Y_n)p_n} \right] \frac{B(Y_n)w_n}{b^*(Y_n)p_n} + u[Y_n, p_n] = F[Y_n, p_n, w_n], \tag{4}$$

where

$$Q^* \left[I - \frac{B(y)pe^*}{b^*(y)p} \right] \frac{B(y)}{b^*(y)p}$$

is the Jacobian matrix at the point $p \in \mathscr{P}$ of the mapping

$$p \mapsto \frac{Q^* B(y)p}{b^*(y)p} = f[y,p],$$

and where

$$e^* u[y,p] = 0,$$

for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$. The random sequence $\{w_n, n \geq 0\}$ takes values in the linear space

$$\Sigma = \{w \in \mathbb{R}^N : e^* w = 0\},$$

which is the linear tangent space to $\mathscr{P}$. To make explicit the dependency with respect to the initial condition and the observations, we introduce the notation

$$w_{n+1} = F[Y_n, \ldots, Y_m, p_m, w_m],$$

for any integers n, m such that $n \geq m$.

Here is a short overview of the results obtained in the second part of this paper:

Exponential forgetting. In Theorem 4.5 we obtain an explicit bound for the Lipschitz constant of the solution map associated with (4). A similar result has been obtained in Proposition 4.1 of [LM1] for the time-dependent case. This nonlogarithmic and nonasymptotic bound goes to zero at exponential rate as time goes to infinity, and as a consequence in Theorem 4.6 we obtain an upper bound for the $\mathbf{P}_\bullet$-a.s. exponential rate of forgetting of the initial condition for (4), under some integrability assumption on the vectors $b_\bullet$ and b of observation conditional densities, and on the function u. A statistical application is to prove that the gradient of the prediction filter, hence the score function, is Lipschitz continuous with respect to some parameter in the model, uniformly in time.

Geometric ergodicity. In Theorem 5.4 we prove the geometric ergodicity of the extended Markov chain $\{Z'_n = (X_n, Y_n, p_n, w_n), n \geq 0\}$, under some integrability assumption on the vectors $b_\bullet$ and b of observation conditional densities, and on the function u. We prove also the existence of a unique invariant probability distribution, and the existence of a solution to the associated Poisson equation. From this point, the law of large numbers and the central limit theorem can be proved for the extended Markov chain $\{Z'_n = (X_n, Y_n, p_n, w_n), n \geq 0\}$. A statistical application is to prove the asymptotic normality of the score function (suitably normalized) and to obtain an explicit expression for the asymptotic covariance matrix, i.e., for the Fisher information matrix.

2. Exponential Forgetting for the Prediction Filter

Let $\|\cdot\|$ denote the L_1-norm, i.e., for any $u = (u^i)$ in $\mathbb{R}^N$,

$$\|u\| = \sum_{i \in S} |u^i|,$$

and for any $N \times N$ matrix $M = (M^{i,j})$,

$$\|M\| = \sup_{0 \neq u \in \mathbb{R}^N} \frac{\|Mu\|}{\|u\|} = \max_{j \in S} \sum_{i \in S} |M^{i,j}|.$$

By definition, the $N \times N$ stochastic matrix $Q = (q^{i,j})$ is primitive if there exists an integer r such that the matrix Q^r is positive, and the smallest such integer is called the index of primitivity of Q. Using the notation $\min^+$ to denote the minimum over positive elements, we define

$$\varepsilon = \min_{i,j \in S}{}^{+} q^{i,j} > 0 \quad \text{and} \quad \delta(y) = \frac{\max_{i \in S} b^i(y)}{\min_{i \in S} b^i(y)} < \infty,$$

for any $y \in \mathbb{R}^d$.

Theorem 2.1. *If the stochastic matrix Q is primitive, with index of primitivity r, then for any $p, p' \in \mathscr{P}$, any integers n, m such that $n \geq m + r - 1$, and any sequence $y_m, \ldots, y_n \in \mathbb{R}^d$,*

$$\begin{aligned} &\|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \\ &\quad \leq \varepsilon^{-r}\delta(y_m) \cdots \delta(y_{m+r-1}) \\ &\qquad \times \prod_{\kappa=1}^{[n,m]} (1 - \varepsilon^r [\delta(y_{m+(\kappa-1)r+1}) \cdots \delta(y_{m+\kappa r-1})]^{-1}) \|p - p'\| \end{aligned}$$

and

$$\begin{aligned} &\|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \\ &\qquad \leq 2 \prod_{\kappa=1}^{[n,m]} (1 - \varepsilon^r [\delta(y_{m+(\kappa-1)r+1}) \cdots \delta(y_{m+\kappa r-1})]^{-1}), \end{aligned}$$

where $[n, m] = \lfloor (n - m + 1)/r \rfloor$ is the maximum number of disjoint blocks of length r in the set $\{m, \ldots, n\}$.

This is an immediate application of Theorem 3.5 in the companion paper [LM4].

For any $p, p' \in \mathscr{P}$, and any infinite sequence $\{y_n, n \geq m\}$ such that

$$\limsup_{\ell \to \infty} \frac{1}{\ell} \sum_{\kappa=1}^{\ell} \log(1 - \varepsilon^r [\delta(y_{m+(\kappa-1)r+1}) \cdots \delta(y_{m+\kappa r-1})]^{-1}) < 0,$$

the difference $\|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\|$ goes to zero at exponential rate

$$\begin{aligned}\limsup_{n \to \infty} & \frac{1}{n} \log \|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \\ & \leq \limsup_{n \to \infty} \frac{1}{n} \sum_{\kappa=1}^{[n,m]} \log(1 - \varepsilon^r [\delta(y_{m+(\kappa-1)r+1}) \cdots \delta(y_{m+\kappa r-1})]^{-1}) \\ & = \frac{1}{r} \limsup_{\ell \to \infty} \frac{1}{\ell} \sum_{\kappa=1}^{\ell} \log(1 - \varepsilon^r [\delta(y_{m+(\kappa-1)r+1}) \cdots \delta(y_{m+\kappa r-1})]^{-1}) < 0.\end{aligned}$$

To obtain an estimate of the almost sure exponential rate of forgetting, we define

$$\Delta_{-1} = \min_{i \in S} \int_{\mathbb{R}^d} \delta^{-1}(y) b_\bullet^i(y) \lambda(dy).$$

Notice that $0 < \Delta_{-1} \leq 1$, hence for any sequence $i_1, \ldots, i_r \in S$,

$$\begin{aligned}0 & \leq \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} (1 - \varepsilon^r [\delta(y_2) \cdots \delta(y_r)]^{-1}) b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_r}(y_r) \lambda(dy_1) \cdots \lambda(dy_r) \\ & = 1 - \varepsilon^r \prod_{k=2}^{r} \int_{\mathbb{R}^d} \delta^{-1}(y) b_\bullet^{i_k}(y) \lambda(dy) \leq 1 - \varepsilon^r \Delta_{-1}^{r-1} = 1 - R < 1, \end{aligned} \tag{5}$$

with $R = \varepsilon^r \Delta_{-1}^{r-1} > 0$. Notice that when the matrix Q is positive, i.e., when $r = 1$, then $R = \varepsilon$.

Theorem 2.2. *If the stochastic matrix Q is primitive, with index of primitivity r, and if Assumption A holds, then for any $p, p' \in \mathscr{P}$, and any integer m,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \|f[Y_n, \ldots, Y_m, p] - f[Y_n, \ldots, Y_m, p']\| \leq \frac{1}{r} \log(1 - R), \qquad \mathbf{P}_\bullet\text{-}a.s.,$$

where $R = \varepsilon^r \Delta_{-1}^{r-1}$.

Proof. If Assumption A holds, then the Markov chain $\{(X_n, Y_n), n \geq 0\}$ is geometrically ergodic under the true probability measure $\mathbf{P}_\bullet$, with a unique invariant probability distribution $\nu_\bullet = (\nu_\bullet^i)$ on $S \times \mathbb{R}^d$, and for any $i \in S$,

$$\nu_\bullet^i(dy) = \mu_\bullet^i b_\bullet^i(y) \lambda(dy).$$

Therefore, $\mathbf{P}_\bullet$-a.s,

$$\begin{aligned}
&\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{\kappa=1}^{\ell}\log(1-\varepsilon^r[\delta(Y_{m+(\kappa-1)r+1})\cdots\delta(Y_{m+\kappa r-1})]^{-1})\\
&\qquad=\sum_{i_1,\dots,i_r\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}\log(1-\varepsilon^r[\delta(y_2)\cdots\delta(y_r)]^{-1})\\
&\qquad\quad\times\mu_\bullet^{i_1}q_\bullet^{i_1,i_2}\cdots q_\bullet^{i_{r-1},i_r}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_r}(y_r)\lambda(dy_1)\cdots\lambda(dy_r)\\
&\qquad\le\log(1-R)<0,
\end{aligned}$$

using the Jensen inequality, and estimate (5) above. ■

3. Geometric Ergodicity of the Markov Chain $\{X_n, Y_n, p_n\}$

Under the probability measure $\mathbf{P}_\bullet$ corresponding to the *true* transition probability matrix $Q_\bullet$ and the *true* vector $b_\bullet$ of observation conditional densities, the extended Markov chain $\{Z_n=(X_n,Y_n,p_n), n\ge 0\}$ with values in $E=S\times\mathbb{R}^d\times\mathscr{P}$, has the following transition probability matrix/kernel:

$$\begin{aligned}
\Pi^{i,j}(y,p,dy',dp')&=\mathbf{P}_\bullet[X_{n+1}=j,\ Y_{n+1}\in dy',p_{n+1}\in dp'\mid X_n=i,\ Y_n=y,p_n=p]\\
&=q_\bullet^{i,j}b_\bullet^j(y')\lambda(dy')\delta_{f[y,\,p]}(dp').
\end{aligned}$$

For any real-valued function g defined on $E=S\times\mathbb{R}^d\times\mathscr{P}$, which is equivalently defined as a collection $g=(g^i)$ of real-valued functions defined on $\mathbb{R}^d\times\mathscr{P}$, we have

$$\begin{aligned}
(\Pi g)^i(y,p)&=\mathbf{E}_\bullet[g(X_{n+1},Y_{n+1},p_{n+1})\mid X_n=i,Y_n=y,p_n=p]\\
&=\sum_{j\in S}\mathbf{E}_\bullet[g^j(Y_{n+1},p_{n+1})\mathbf{1}_{[X_{n+1}\,=j]}\mid X_n=i,\ Y_n=y,p_n=p]\\
&=\sum_{j\in S}\int_{\mathbb{R}^d}g^j(y',f[y,p])q_\bullet^{i,j}b_\bullet^j(y')\lambda(dy')\\
&=\sum_{j\in S}\int_{\mathbb{R}^d\times\mathscr{P}}g^j(y',p')\Pi^{i,j}(y,p,dy',dp'),
\end{aligned}$$

for any $i\in S$, any $y\in\mathbb{R}^d$, and any $p\in\mathscr{P}$. More generally, for any integer $n\ge 2$ we have

$$\begin{aligned}
(\Pi^n g)^i(y,p)&=\mathbf{E}_\bullet[g(X_n,Y_n,p_n)\mid X_0=i,Y_0=y,p_0=p]\\
&=\sum_{i_n\in S}\mathbf{E}_\bullet[g^{i_n}(Y_n,p_n)\mathbf{1}_{[X_n\,=\,i_n]}\mid X_0=i,\ Y_0=y,p_0=p]\\
&=\sum_{i_1,\dots,i_n\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}g^{i_n}(y_n,f[y_{n-1},\dots,y_1,y,p])\\
&\quad\times q_\bullet^{i,i_1}\cdots q_\bullet^{i_{n-1},i_n}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n),
\end{aligned}$$

for any $i\in S$, any $y\in\mathbb{R}^d$, and any $p\in\mathscr{P}$.

Recall that the set of bounded and Lipschitz continuous functions $g = (g^i)$ defined on $E = S \times \mathbb{R}^d \times \mathscr{P}$, is a Banach space for the norm $\| \cdot \|_{\mathrm{BL}}$ defined by

$$\|g\|_{\mathrm{BL}} = \max_{i \in S} \sup_{(y,p) \in \mathbb{R}^d \times \mathscr{P}} |g^i(y,p)| + \max_{i \in S} \sup_{(y,p) \neq (y',p') \in \mathbb{R}^d \times \mathscr{P}} \frac{|g^i(y,p) - g^i(y',p')|}{|y - y'| + \|p - p'\|},$$

and the set $\mathscr{P}(E)$ of probability distributions on E is a Banach space for the dual norm $\| \cdot \|_{\mathrm{BL}}^*$ defined by

$$\|\mu - \mu'\|_{\mathrm{BL}}^* = \sup_{\|g\|_{\mathrm{BL}} = 1} |\langle \mu, g \rangle - \langle \mu', g \rangle|,$$

which defines the same topology as the topology associated with the weak convergence of probability distributions, see [D1] and [D2]. This framework is somehow too strong for our purpose, and we consider instead the following set of test functions.

Definition 3.1. Let L denote the set of functions $g = (g^i)$ defined on $E = S \times \mathbb{R}^d \times \mathscr{P}$, such that for any $i \in S$ and any $y \in \mathbb{R}^d$ the partial mapping $p \mapsto g^i(y,p)$ is Lipschitz continuous, hence bounded since $\mathscr{P}$ is compact, i.e., there exist constants $\mathrm{K}(g^i, y)$ and $\mathrm{Lip}(g^i, y)$ such that

$$\mathrm{K}(g^i, y) = \sup_{p \in \mathscr{P}} |g^i(y,p)| < \infty,$$

$$\mathrm{Lip}(g^i, y) = \sup_{p \neq p' \in \mathscr{P}} \frac{|g^i(y,p) - g^i(y,p')|}{\|p - p'\|} < \infty,$$

and such that

$$\mathrm{Lip}(g) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{Lip}(g^i, y) b_\bullet^i(y) \lambda(dy) < \infty,$$

$$\mathrm{K}(g) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{K}(g^i, y) b_\bullet^i(y) \lambda(dy) < \infty.$$

Remark 3.2. The set L of test functions is sufficiently large to contain some functions related to the identification of HMMs, see Example 3.3 below, and to contain also any bounded and Lipschitz continuous function g defined on E, since

$$\mathrm{Lip}(g) + \mathrm{K}(g) \leq \|g\|_{\mathrm{BL}}.$$

Example 3.3 (Log-Likelihood Function). If Δ is finite, see Definition 4.1 below, and if

$$\max_{i \in S} \int_{\mathbb{R}^d} \left[\max_{j \in S} |\log b^j(y)| \right] b_\bullet^i(y) \lambda(dy) < \infty,$$

then the function g defined by

$$g(y,p) = \log[b^*(y)p],$$

for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$, i.e., constant over S, belongs to the set L. Indeed, for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$,

$$\min_{j \in S} b^j(y) \le b^*(y)p \le \max_{j \in S} b^j(y),$$

hence

$$\begin{aligned}|\log[b^*(y)p]| &= \log^+[b^*(y)p] + \log^-[b^*(y)p] \\ &\le \log^+\left[\max_{j \in S} b^j(y)\right] + \log^-\left[\min_{j \in S} b^j(y)\right] \\ &\le \max_{j \in S} \log^+[b^j(y)] + \max_{j \in S} \log^-[b^j(y)] \le 2 \max_{j \in S} |\log b^j(y)|,\end{aligned}$$

whereas using the refined estimate of Lemma A.3 in the companion paper [LM4] yields, for any $y \in \mathbb{R}^d$ and any $p, p' \in \mathscr{P}$,

$$\begin{aligned}\log[b^*(y)p] - \log[b^*(y)p'] &= \log\left[1 + \frac{b^*(y)(p-p')}{b^*(y)p'}\right] \\ &\le \frac{b^*(y)(p-p')}{b^*(y)p'} \le \tfrac{1}{2}[\delta(y) - 1]\|p - p'\|,\end{aligned}$$

hence

$$|\log[b^*(y)p] - \log[b^*(y)p']| \le \tfrac{1}{2}[\delta(y) - 1]\,\|p - p'\|.$$

Let $\rho_{\max} = \max(\rho_*, \rho_\bullet)$, where $\rho_* = (1-R)^{1/r}$, and where the constants $\rho_\bullet$ and $R = \varepsilon^r \Delta_{-1}^{r-1}$ are defined in Remark 1.2 and in Theorem 2.2, respectively.

Remark 3.4. For any integers n, m such that $n \ge m + r - 1$,

$$r[n,m] = r\left\lfloor \frac{n-m+1}{r} \right\rfloor > n - m + 1 - r,$$

hence

$$(1-R)^{[n,m]} \le \rho_*^{n-m+1-r}. \tag{6}$$

It follows from the estimate

$$\rho_* = (1-R)^{1/r} \le 1 - \frac{R}{r},$$

that

$$\frac{1}{1-\rho_{\max}} = \max\left(\frac{1}{1-\rho_*}, \frac{1}{1-\rho_\bullet}\right) \le \max\left(\frac{r\varepsilon^{-r}}{\Delta_{-1}^{r-1}}, \frac{1}{1-\rho_\bullet}\right) = \frac{r\varepsilon^{-r}}{\Delta_{-1}^{r-1}},$$

for $\varepsilon > 0$ small enough. Finally, another useful estimate is

$$\rho_{\max}^{-r} \le \rho_*^{-r} = \frac{1}{1-R}. \tag{7}$$

Theorem 3.5. *If the stochastic matrix Q is primitive, with index of primitivity r, and if Assumption A holds, then there exists a constant $C > 0$ such that, for any*

$z, z' \in E$ and for any function g in L,

$$|\Pi^n g(z) - \Pi^n g(z')| \leq C[\mathrm{Lip}(g) + \mathrm{K}(g)] n \rho_{\max}^{n-1}, \tag{8}$$

where the product $C(1-R)$ depends only on $K_\bullet$.

The following corollary holds, which is similar to Proposition 2 in Chapter 2, Part II, of [BMP].

Corollary 3.6. *With the assumptions of Theorem* 3.5:

(i) *For any $z \in E$, and for any function g in L, there exists a constant $\lambda(g)$ independent of z, such that*

$$|\Pi^n g(z) - \lambda(g)| \leq C[\mathrm{Lip}(g) + \mathrm{K}(g)] \frac{n \rho_{\max}^{n-1}}{(1-\rho_{\max})^2}, \tag{9}$$

and there exists a (not necessarily unique) solution $V = (V^i)$ defined on E of the Poisson equation

$$[I - \Pi] V(z) = g(z) - \lambda(g).$$

(ii) *The Markov chain $\{Z_n = (X_n, Y_n, p_n), n \geq 0\}$ has, under the true probability measure $\mathbf{P}_\bullet$, a unique invariant probability distribution $\mu = (\mu^i)$ on E, and the constant $\lambda(g)$ is defined as*

$$\lambda(g) = \int_E g(z) \mu(dz).$$

Proof. For any function g in L, and any $z \in E$,

$$\begin{aligned} \Pi^{n+1} g(z) - \Pi^n g(z) &= \int_E \Pi^n g(z') \Pi(z, dz') - \Pi^n g(z) \\ &= \int_E [\Pi^n g(z') - \Pi^n g(z)] \Pi(z, dz'), \end{aligned}$$

and it follows from estimate (8) that

$$|\Pi^{n+1} g(z) - \Pi^n g(z)| \leq C[\mathrm{Lip}(g) + \mathrm{K}(g)] n \rho_{\max}^{n-1}. \tag{10}$$

Therefore $\{\Pi^n g(z), n \geq 0\}$ is a Cauchy sequence, hence converges to a limit $\lambda(g)$ which is independent of $z \in E$ by estimate (8). The other points of (i) can be checked along the same lines as in the proof of Proposition 2 in Chapter 2, Part II, of [BMP].

To show the existence and uniqueness of the invariant probability distribution, let μ_n denote the probability distribution of Z_n on E, with initial probability distribution μ_0. For any function g in L,

$$\langle \mu_{n+1}, g \rangle - \langle \mu_n, g \rangle = \int_E [\Pi^{n+1} g(z) - \Pi^n g(z)] \mu_0(dz),$$

hence it follows from estimate (10) and Remark 3.2 that

$$\|\mu_{n+1} - \mu_n\|^*_{\mathrm{BL}} = \sup_{\|g\|_{\mathrm{BL}}=1} |\langle \mu_{n+1}, g\rangle - \langle \mu_n, g\rangle| \le C n \rho_{\max}^{n-1}.$$

Therefore $\{\mu_n, n \ge 0\}$ is a Cauchy sequence in the Banach space $(\mathscr{P}(E), \|\cdot\|^*_{\mathrm{BL}})$, hence converges weakly to a probability distribution μ on E, which is invariant with respect to $\{\Pi^n, n \ge 0\}$. To see that the limit does not depend on the initial distribution μ_0, let μ'_n denote the probability distribution of Z_n on E, with initial distribution μ'_0. For any function g in L,

$$\begin{aligned}\langle \mu_n, g\rangle - \langle \mu'_n, g\rangle &= \int_E \Pi^n g(z)[\mu_0(dz) - \mu'_0(dz)] \\ &= \int_E [\Pi^n g(z) - \lambda(g)][\mu_0(dz) - \mu'_0(dz)],\end{aligned}$$

hence it follows from estimate (9) and Remark 3.2 that

$$\|\mu_n - \mu'_n\|^*_{\mathrm{BL}} = \sup_{\|g\|_{\mathrm{BL}}=1} |\langle \mu_n, g\rangle - \langle \mu'_n, g\rangle| \le C\|\mu_0 - \mu'_0\|_{\mathrm{TV}} \frac{n\rho_{\max}^{n-1}}{(1-\rho_{\max})^2}.$$

This finishes the proof of (ii). ∎

The proof of Theorem 3.5 is given in Appendix A, and is based on the next result.

Proposition 3.7. *If the stochastic matrix Q is primitive, with index of primitivity r, then for any $p, p' \in \mathscr{P}$, for any integers n, m such that $n \ge m + r - 1$, and for any function $g = (g^i)$ in L,*

$$\begin{aligned}\max_{i_m,\ldots,i_{n+1}\in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} &|g^{i_{n+1}}(y_{n+1}, f[y_n,\ldots,y_m,p]) - g^{i_{n+1}}(y_{n+1}, f[y_n,\ldots,y_m,p'])| \\ &\times b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_{n+1}}(y_{n+1})\lambda(dy_m)\cdots\lambda(dy_{n+1}) \\ &\le 2\,\mathrm{Lip}(g)\rho_*^{n-m+1-r},\end{aligned}$$

where $\rho_* = (1-R)^{1/r}$.

Proof. For any sequence $i_m, \ldots, i_{n+1} \in S$,

$$\begin{aligned}\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} &|g^{i_{n+1}}(y_{n+1}, f[y_n,\ldots,y_m,p]) - g^{i_{n+1}}(y_{n+1}, f[y_n,\ldots,y_m,p'])| \\ &\times b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_{n+1}}(y_{n+1})\lambda(dy_m)\cdots\lambda(dy_{n+1}) \\ \le &\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \mathrm{Lip}(g^{i_{n+1}}, y_{n+1})\|f[y_n,\ldots,y_m,p] - f[y_n,\ldots,y_m,p']\| \\ &\times b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_{n+1}}(y_{n+1})\lambda(dy_m)\cdots\lambda(dy_{n+1}) \\ \le &\,\mathrm{Lip}(g) \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|f[y_n,\ldots,y_m,p] - f[y_n,\ldots,y_m,p']\| \\ &\times b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_m)\cdots\lambda(dy_n).\end{aligned}$$

Recall from Theorem 2.1 that

$$\|f[y_n,\dots,y_m,p]-f[y_n,\dots,y_m,p']\| \\ \leq 2\prod_{\kappa=1}^{[n,m]}(1-\varepsilon^r[\delta(y_{m+(\kappa-1)r+1})\cdots\delta(y_{m+\kappa r-1})]^{-1}),$$

hence integration is straightforward, and the upper bound (6) yields

$$\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}\|f[y_n,\dots,y_m,p]-f[y_n,\dots,y_m,p']\| \\ \times b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_m)\cdots\lambda(dy_n) \\ \leq 2(1-R)^{[n,m]}\leq 2\rho_*^{n-m+1-r},$$

for any sequence $i_m,\dots,i_n\in S$. ■

4. Exponential Forgetting for the Linear Tangent Prediction Filter

Recall that $\{p_n, n\geq 0\}$ and $\{w_n, n\geq 0\}$ satisfy (3) and (4) respectively, i.e.,

$$p_{n+1}=\frac{Q^*B(Y_n)p_n}{b^*(Y_n)p_n}=f[Y_n,p_n]=f[Y_n,\dots,Y_m,p_m]$$

and

$$w_{n+1}=Q^*\left[I-\frac{B(Y_n)p_ne^*}{b^*(Y_n)p_n}\right]\frac{B(Y_n)w_n}{b^*(Y_n)p_n}+u[Y_n,p_n] \\ =F[Y_n,p_n,w_n]=F[Y_n,\dots,Y_m,p_m,w_m],$$

for any integers n,m such that $n\geq m$.

In addition to the Assumption A on the *true* transition probability matrix $Q_\bullet$, we shall need the following *integrability* assumption on the vectors $b_\bullet$ and b of observation conditional densities:

$$\Delta=\max_{i\in S}\int_{\mathbb{R}^d}\delta(y)b_\bullet^i(y)\lambda(dy)<\infty.$$

More generally, we introduce the following definition.

Definition 4.1. For any $s\geq 0$, let

$$\Delta_s=\max_{i\in S}\int_{\mathbb{R}^d}\delta^s(y)b_\bullet^i(y)\lambda(dy).$$

Notice that $\Delta_1=\Delta$, and that the mapping $s\mapsto\Delta_s$ is nondecreasing on $[0,\infty)$, since $\delta(y)\geq 1$ for any $y\in\mathbb{R}^d$.

Example 4.2. If the observation conditional densities are Gaussian for both the true and the assumed models, i.e., in particular

$$b^i(y)=(2\pi)^{-d/2}\exp\{-\tfrac{1}{2}|y-h^i|^2\},$$

for any $i \in S$ and any $y \in \mathbb{R}^d$, then Δ_s is finite for any $s \geq 0$. Indeed, for any $i, j \in S$ and any $y \in \mathbb{R}^d$,

$$\frac{b^i(y)}{b^j(y)} = \exp\{-\tfrac{1}{2}|y - h^i|^2 + \tfrac{1}{2}|y - h^j|^2\} = \exp\{y^*(h^i - h^j) - \tfrac{1}{2}(h^i + h^j)^*(h^i - h^j)\},$$

hence

$$\delta(y) \leq \exp\left\{\max_{i,j \in S} |h^i - h^j| \left[|y| + \max_{i \in S} |h^i|\right]\right\}.$$

Assumption B. For any $y \in \mathbb{R}^d$, the partial mapping $p \mapsto u[y, p]$ with values in Σ is Lipschitz continuous, hence bounded since $\mathscr{P}$ is compact, i.e., there exist constants $\mathrm{K}(u, y)$ and $\mathrm{Lip}(u, y)$ such that

$$\mathrm{K}(u, y) = \sup_{p \in \mathscr{P}} \|u[y, p]\| < \infty,$$

$$\mathrm{Lip}(u, y) = \sup_{p \neq p' \in \mathscr{P}} \frac{\|u[y, p] - u[y, p']\|}{\|p - p'\|} < \infty.$$

Definition 4.3. Under Assumption B, let

$$\mathrm{Lip}(u) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{Lip}(u, y) b^i_\bullet(y) \lambda(dy),$$

$$\mathrm{K}(u) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{K}(u, y) b^i_\bullet(y) \lambda(dy).$$

Example 4.4. The function u defined by

$$u[y, p] = \frac{\partial Q^* B(y) p}{b^*(y) p},$$

for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$, satisfies Assumption B, and if Δ is finite, then $\mathrm{Lip}(u)$ and $\mathrm{K}(u)$ are finite. Indeed, for any $y \in \mathbb{R}^d$ and any $p \in \mathscr{P}$,

$$\left\| \frac{\partial Q^* B(y) p}{b^*(y) p} \right\| \leq \|\partial Q^*\|,$$

whereas it follows from the Lipschitz estimate of Lemma A.1 in the companion paper [LM4] that for any $y \in \mathbb{R}^d$ and any $p, p' \in \mathscr{P}$,

$$\left\| \frac{\partial Q^* B(y) p}{b^*(y) p} - \frac{\partial Q^* B(y) p'}{b^*(y) p'} \right\| \leq \delta(y) \|\partial Q^*\| \, \|p - p'\|.$$

Theorem 4.5. *If the stochastic matrix Q is primitive, with index of primitivity r, and if Assumption B holds, then for any $p, p' \in \mathscr{P}$, any $w, w' \in \Sigma$, any integers n, m*

such that $n \geq m + r - 1$, *and any sequence* $y_m, \ldots, y_n \in \mathbb{R}^d$,

$$\begin{aligned}
&\|F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w']\| \\
&\qquad \leq \tfrac{1}{2}([\varepsilon^{-r}\delta(y_m)\cdots\delta(y_{m+r-1})]^2 + 1) \\
&\qquad\quad \times \prod_{\kappa=1}^{[n,m]} (1 - \varepsilon^r[\delta(y_{m+(\kappa-1)r+1})\cdots\delta(y_{m+\kappa r-1})]^{-1}) \\
&\qquad\quad \times [\|w - w'\| + \|p - p'\|(\|w\| + \|w'\|)] + \varepsilon^{-r}\delta(y_m)\cdots\delta(y_{m+r-1}) \\
&\qquad\quad \times \Bigg[\mathrm{Lip}(u, y_n) + \sum_{l=m}^{n-1} \varepsilon^{-r}\delta(y_{l+1})\cdots\delta(y_{\min(l+r,n)})\,\mathrm{Lip}(u, y_l) \\
&\qquad\qquad + \tfrac{1}{2}\sum_{l=m}^{n-1}([\varepsilon^{-r}\delta(y_{l+1})\cdots\delta(y_{\min(l+r,n)})]^2 - 1)\,\mathrm{K}(u, y_l)\Bigg] \\
&\qquad\quad \times (1 - \varepsilon^r)^{-1} \prod_{\kappa=1}^{[n,m]} (1 - \varepsilon^r[\delta(y_{m+(\kappa-1)r+1})\cdots\delta(y_{m+\kappa r-1})]^{-1})\|p - p'\|.
\end{aligned}$$

This is an immediate application of Theorem 5.7 in the companion paper [LM4].

For any $p, p' \in \mathscr{P}$, any $w, w' \in \Sigma$, and any infinite sequence $\{y_n, n \geq m\}$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{l=m}^{n} \delta(y_{l+1})\cdots\delta(y_{l+r})\,\mathrm{Lip}(u, y_l) < \infty,$$

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{l=m}^{n} ([\varepsilon^{-r}\delta(y_{l+1})\cdots\delta(y_{l+r})]^2 - 1)\,\mathrm{K}(u, y_l) < \infty,$$

and

$$\limsup_{\ell\to\infty} \frac{1}{\ell} \sum_{\kappa=1}^{\ell} \log(1 - \varepsilon^r[\delta(y_{m+(\kappa-1)r+1})\cdots\delta(y_{m+\kappa r-1})]^{-1}) < 0,$$

the difference $\|F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w']\|$ goes to zero at exponential rate

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log\|F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w']\| \\
&\qquad \leq \frac{1}{r} \limsup_{\ell\to\infty} \frac{1}{\ell} \sum_{\kappa=1}^{\ell} \log(1 - \varepsilon^r[\delta(y_{m+(\kappa-1)r+1})\cdots\delta(y_{m+\kappa r-1})]^{-1}) < 0.
\end{aligned}$$

Theorem 4.6. *If the stochastic matrix* Q *is primitive, with index of primitivity* r, *if Assumptions A and B hold, and if* Δ_2, $\mathrm{Lip}(u)$, *and* $\mathrm{K}(u)$ *are finite, then for any*

$p, p' \in \mathscr{P}$, any $w, w' \in \Sigma$, and any integer m,

$$\limsup_{n\to\infty} \frac{1}{n} \log\|F[Y_n, \ldots, Y_m, p, w] - F[Y_n, \ldots, Y_m, p', w']\| \\ \leq \frac{1}{r} \log(1 - R), \qquad \mathbf{P}_\bullet\text{-}a.s.,$$

where $R = \varepsilon^r \Delta_{-1}^{r-1}$.

Proof. It has already been proved in Theorem 2.2 that $\mathbf{P}_\bullet$-a.s.

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{\kappa=1}^{\ell} \log(1 - \varepsilon^r [\delta(Y_{m+(\kappa-1)r+1}) \cdots \delta(Y_{m+\kappa r-1})]^{-1}) \leq \log(1 - R).$$

Following the same lines as in the proof of Theorem 2.2, if Δ_2, $\mathrm{Lip}(u)$, and $\mathrm{K}(u)$ are finite, then $\mathbf{P}_\bullet$-a.s.

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \sum_{l=m}^{n} \delta(Y_{l+1}) \cdots \delta(Y_{l+r}) \,\mathrm{Lip}(u, Y_l) \\ &\quad = \sum_{i_0,\ldots,i_r\in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \delta(y_1) \cdots \delta(y_r) \,\mathrm{Lip}(u, y_0) \\ &\qquad \times \mu_\bullet^{i_0} q_\bullet^{i_0,i_1} \cdots q_\bullet^{i_{r-1},i_r} b_\bullet^{i_0}(y_0) \cdots b_\bullet^{i_r}(y_r) \lambda(dy_0) \cdots \lambda(dy_r) \\ &\quad \leq \Delta^r \,\mathrm{Lip}(u) < \infty, \end{aligned}$$

and similarly

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \sum_{l=m}^{n} ([\varepsilon^{-r} \delta(Y_{l+1}) \cdots \delta(Y_{l+r})]^2 - 1) \,\mathrm{K}(u, Y_l) \\ &\quad = \sum_{i_0,\ldots,i_r\in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} ([\varepsilon^{-r} \delta(y_1) \cdots \delta(y_r)]^2 - 1) \,\mathrm{K}(u, y_0) \\ &\qquad \times \mu_\bullet^{i_0} q_\bullet^{i_0,i_1} \cdots q_\bullet^{i_{r-1},i_r} b_\bullet^{i_0}(y_0) \cdots b_\bullet^{i_r}(y_r) \lambda(dy_0) \cdots \lambda(dy_r) \\ &\quad \leq (\varepsilon^{-2r} \Delta_2^r - 1) \,\mathrm{K}(u) < \infty. \end{aligned}$$

■

Remark 4.7. This bound for the exponential rate of forgetting is exactly the same as the bound obtained in Theorem 2.2.

5. Geometric Ergodicity of the Markov Chain $\{X_n, Y_n, p_n, w_n\}$

Under the probability measure $\mathbf{P}_\bullet$ corresponding to the *true* transition probability matrix $Q_\bullet$ and the *true* vector $b_\bullet$ of observation conditional densities, the extended Markov chain $\{Z_n' = (X_n, Y_n, p_n, w_n), n \geq 0\}$ with values in $E' = S \times \mathbb{R}^d \times$

$\mathscr{P} \times \Sigma$, has the following transition probability matrix/kernel:

$$\Pi^{i,j}(y, p, w, dy', dp', dw') = \mathbf{P}_\bullet[X_{n+1} = j,\, Y_{n+1} \in dy', p_{n+1} \in dp', w_{n+1} \in dw' \mid X_n = i,\, Y_n = y, p_n = p, w_n = w] = q_\bullet^{i,j} b_\bullet^j(y') \lambda(dy') \delta_{f[y,\,p]}(dp') \delta_{F[y,\,p,\,w]}(dw').$$

For any real-valued function g defined on $E' = S \times \mathbb{R}^d \times \mathscr{P} \times \Sigma$, which is equivalently defined as a collection $g = (g^i)$ of real-valued functions defined on $\mathbb{R}^d \times \mathscr{P} \times \Sigma$, we have

$$\begin{aligned}(\Pi g)^i(y, p, w) &= \mathbf{E}_\bullet[g(X_{n+1}, Y_{n+1}, p_{n+1}, w_{n+1}) \mid X_n = i,\, Y_n = y, p_n = p, w_n = w] \\ &= \sum_{j \in S} \mathbf{E}_\bullet[g^j(Y_{n+1}, p_{n+1}, w_{n+1}) \mathbf{1}_{[X_{n+1} = j]} \mid X_n = i,\, Y_n = y, p_n = p, w_n = w] \\ &= \sum_{j \in S} \int_{\mathbb{R}^d} g^j(y', f[y, p], F[y, p, w]) q_\bullet^{i,j} b_\bullet^j(y') \lambda(dy') \\ &= \sum_{j \in S} \int_{\mathbb{R}^d \times \mathscr{P} \times \Sigma} g^j(y', p', w') \Pi^{i,j}(y, p, w, dy', dp', dw'),\end{aligned}$$

for any $i \in S$, any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$. More generally, for any integer $n \geq 2$ we have

$$\begin{aligned}&(\Pi^n g)^i(y, p, w) \\ &\quad = \mathbf{E}_\bullet[g(X_n, Y_n, p_n, w_n) \mid X_0 = i,\, Y_0 = y, p_0 = p, w_0 = w] \\ &\quad = \sum_{i_n \in S} \mathbf{E}_\bullet[g^{i_n}(Y_n, p_n, w_n) \mathbf{1}_{[X_n = i_n]} \mid X_0 = i,\, Y_0 = y, p_0 = p, w_0 = w] \\ &\quad = \sum_{i_1, \ldots, i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g^{i_n}(y_n, f[y_{n-1}, \ldots, y_1, y, p], F[y_{n-1}, \ldots, y_1, y, p, w]) \\ &\qquad \times q_\bullet^{i, i_1} \cdots q_\bullet^{i_{n-1}, i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_1) \cdots \lambda(dy_n),\end{aligned}$$

for any $i \in S$, any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$.

For any $p \in \mathscr{P}$, any $w \in \Sigma$, and any $y \in \mathbb{R}^d$, let

$$f \otimes F[y, p, w] = (f[y, p], F[y, p, w]).$$

More generally, for any $p \in \mathscr{P}$, any $w \in \Sigma$, and any sequence $y_m, \ldots, y_n \in \mathbb{R}^d$, let

$$f \otimes F[y_n, \ldots, y_m, p, w] = (f[y_n, \ldots, y_m, p], F[y_n, \ldots, y_m, p, w]).$$

With this definition, we can write the transition probability matrix/kernel equivalently as

$$\Pi^{i,j}(y, p, w, dy', dp', dw') = q_\bullet^{i,j} b_\bullet^j(y') \lambda(dy') \delta_{f \otimes F[y,\,p,\,w]}(dp', dw'),$$

and

$$(\Pi^n g)^i(y, p, w) = \sum_{i_1, \dots, i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g^{i_n}(y_n, f \otimes F[y_{n-1}, \dots, y_1, y, p, w])$$
$$\times q_\bullet^{i, i_1} \cdots q_\bullet^{i_{n-1}, i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_1) \cdots \lambda(dy_n),$$

for any $i \in S$, any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$.

By analogy with Definition 3.1, we introduce the following set of test functions.

Definition 5.1. Let L' denote the set of functions $g = (g^i)$ defined on $E' = S \times \mathbb{R}^d \times \mathscr{P} \times \Sigma$, such that for any $i \in S$ and any $y \in \mathbb{R}^d$ the partial mapping $(p, w) \mapsto g^i(y, p, w)$ is locally Lipschitz continuous, with linear growth, i.e., there exist constants $\mathrm{K}(g^i, y)$ and $\mathrm{Lip}(g^i, y)$ such that

$$\mathrm{K}(g^i, y) = \sup_{(p, w) \in \mathscr{P} \times \Sigma} \frac{|g^i(y, p, w)|}{1 + \|w\|} < \infty,$$

$$\mathrm{Lip}(g^i, y) = \sup_{(p, w) \neq (p', w') \in \mathscr{P} \times \Sigma} \frac{|g^i(y, p, w) - g^i(y, p', w')|}{\|w - w'\| + \|p - p'\|(1 + \|w\| + \|w'\|)} < \infty,$$

and such that

$$\mathrm{Lip}(g) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{Lip}(g^i, y) b_\bullet^i(y) \lambda(dy) < \infty,$$

$$\mathrm{K}(g) = \max_{i \in S} \int_{\mathbb{R}^d} \mathrm{K}(g^i, y) b_\bullet^i(y) \lambda(dy) < \infty.$$

Remark 5.2. The set L' of test functions is sufficiently large to contain some functions related to identification of HMMs, see Example 5.3 below, and to contain also any bounded and Lipschitz continuous function g defined on E', with

$$\mathrm{Lip}(g) + \mathrm{K}(g) \le \|g\|_{\mathrm{BL}}.$$

Example 5.3 (Score Function). If Δ_2 is finite, then the function g defined by

$$g(y, p, w) = \frac{b^*(y) w}{b^*(y) p},$$

for any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$, i.e., constant over S, belongs to the set L'. Indeed, using the refined estimate of Lemma A.3 in the companion paper [LM4] yields for any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$,

$$\left| \frac{b^*(y) w}{b^*(y) p} \right| \le \tfrac{1}{2}[\delta(y) - 1] \|w\|,$$

whereas for any $y \in \mathbb{R}^d$, any $p, p' \in \mathscr{P}$, and any $w, w' \in \Sigma$,

$$\frac{b^*(y) w}{b^*(y) p} - \frac{b^*(y) w'}{b^*(y) p'} = \frac{b^*(y)(w - w')}{b^*(y) p} - \frac{b^*(y)(p - p')}{b^*(y) p} \frac{b^*(y) w'}{b^*(y) p'},$$

hence using the refined estimate of Lemma A.3 in the companion paper [LM4] yields

$$\left| \frac{b^*(y)w}{b^*(y)p} - \frac{b^*(y)w'}{b^*(y)p'} \right| \leq \tfrac{1}{2}[\delta(y) - 1]\|w - w'\| + \tfrac{1}{4}[\delta(y) - 1]^2 \|p - p'\| \|w'\|$$
$$\leq \tfrac{1}{4}[\delta^2(y) - 1][\|w - w'\| + \|p - p'\|(\|w\| + \|w'\|)],$$

since $\max(\frac{1}{2}[\delta - 1], \frac{1}{4}[\delta - 1]^2) \leq \frac{1}{4}[\delta^2 - 1]$ for any $\delta \geq 1$.

Theorem 5.4. *If the stochastic matrix Q is primitive, with index of primitivity r, if Assumptions A and B hold, and if Δ_2, $\mathrm{Lip}(u)$, and $\mathrm{K}(u)$ are finite, then there exists a constant $C > 0$ such that, for any $z, z' \in E'$ and for any function g in L',*

$$|\Pi^n g(z) - \Pi^n g(z')| \leq C\varepsilon^{-3r}\Delta_2^{3r/2}[\mathrm{Lip}(g) + \mathrm{K}(g)]$$
$$\times [1 + \|F[y, p, w]\| + \|F[y', p', w']\|] n(n-1)\rho_{\max}^{n-2},$$

where the product $C(1 - R)^3$ depends only on $K_\bullet$, $\mathrm{Lip}(u)$, and $\mathrm{K}(u)$.

Notice that

$$\sum_{j \in S} \int_{\mathbb{R}^d} \|F[y', p', w']\| \Pi^{i,j}(y, p, w, dy', dp', dw') \leq \Delta \|F[y, p, w]\| + \mathrm{K}(u),$$

for any $i \in S$, any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$, hence the following corollary holds, which is similar to Proposition 2 in Chapter 2, Part II, of [BMP] and whose proof follows the same lines as the proof of Corollary 3.6.

Corollary 5.5. *With the assumptions of Theorem* 5.4:

(i) *For any $z \in E'$, and for any function g in L', there exists a constant $\lambda(g)$ independent of z, such that*

$$|\Pi^n g(z) - \lambda(g)| \leq C\varepsilon^{-3r}\Delta_2^{3r/2}[\mathrm{Lip}(g) + \mathrm{K}(g)][1 + \|F[y, p, w]\|] \frac{n(n-1)\rho_{\max}^{n-2}}{(1 - \rho_{\max})^3},$$

and there exists a (not necessarily unique) solution $V = (V^i)$ defined on E' of the Poisson equation

$$[I - \Pi]V(z) = g(z) - \lambda(g).$$

(ii) *The Markov chain $\{Z'_n = (X_n, Y_n, p_n, w_n), n \geq 0\}$ has, under the true probability measure $\mathbf{P}_\bullet$, a unique invariant probability distribution $\mu = (\mu^i)$ on E', and the constant $\lambda(g)$ is defined as*

$$\lambda(g) = \int_{E'} g(z)\mu(dz).$$

The proof of Theorem 5.4 is given in Appendix C, and is based on the next two results.

Proposition 5.6. *If the stochastic matrix Q is primitive, with index of primitivity r, if Assumption B holds, and if Δ_2, $\mathrm{Lip}(u)$, and $\mathrm{K}(u)$ are finite, then there exists a constant $C > 0$ such that, for any $p, p' \in \mathscr{P}$, for any $w, w' \in \Sigma$, for any integers n, m such that $n \geq m + 3r - 1$, and for any function $g = (g^i)$ in L',*

$$\max_{i_m, \ldots, i_{n+1} \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p, w]) - g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p', w'])| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \leq C\varepsilon^{-2r} \Delta_2^r \, \mathrm{Lip}(g) [\|w\| + \|w'\| + (n - m + 1)] \rho_*^{n-m-3r},$$

where $\rho_* = (1 - R)^{1/r}$, and where the constant C depends only on $\mathrm{Lip}(u)$ and $\mathrm{K}(u)$.

The proof of Proposition 5.6 is given in Appendix B.

Proposition 5.7. *If the stochastic matrix Q is primitive, with index of primitivity r, if Assumption B holds, and if Δ and $\mathrm{K}(u)$ are finite, then there exists a constant $C > 0$ such that, for any $p \in \mathscr{P}$, for any $w \in \Sigma$, for any integers n, m such that $n \geq m$, and for any function $g = (g^i)$ in L',*

$$\max_{i_m, \ldots, i_{n+1} \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p, w])| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \leq C\varepsilon^{-r} \Delta^r \mathrm{K}(g) [\|w\| + (n - m + 1)],$$

where the constant C depends only on $\mathrm{K}(u)$.

Proof. For any sequence $i_m, \ldots, i_{n+1} \in S$,

$$\begin{aligned} &\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p, w])| \\ &\qquad \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \\ &\leq \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \mathrm{K}(g^{i_{n+1}}, y_{n+1})(1 + \|F[y_n, \ldots, y_m, p, w]\|) \\ &\qquad \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \\ &\leq \mathrm{K}(g) \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} (1 + \|F[y_n, \ldots, y_m, p, w]\|) \\ &\qquad \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n). \end{aligned}$$

It follows immediately from the rough estimate of Theorem 5.5 in the companion paper [LM4] that

$$\|F[y_n, \dots, y_m, p, w]\| \le \varepsilon^{-r}\delta(y_m)\cdots\delta(y_{\min(m+r-1,n)})\|w\| + \mathrm{K}(u, y_n) + \sum_{l=m}^{n-1} \mathrm{K}(u, y_l)\varepsilon^{-r}\delta(y_{l+1})\cdots\delta(y_{\min(l+r,n)}).$$

Integration is straightforward in this case, and yields

$$\begin{aligned}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} &\|F[y_n, \dots, y_m, p, w]\| b_\bullet^{i_m}(y_m)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_m)\cdots\lambda(dy_n) \\ &\le \varepsilon^{-r}\Delta^r\|w\| + \mathrm{K}(u) + \mathrm{K}(u)\varepsilon^{-r}\Delta^r(n-m) \\ &\le \varepsilon^{-r}\Delta^r[\|w\| + \mathrm{K}(u)(n-m+1)], \qquad (11)\end{aligned}$$

for any sequence $i_m, \dots, i_n \in S$. ■

Acknowledgments. The authors gratefully acknowledge Jan van Schuppen for bringing the paper [AM] to their attention, Ofer Zeitouni for providing a preprint of the paper [AZ] prior to publication, and Christiane Cocozza for her very careful reading of the manuscript, and especially of the proof of Theorems 3.5 and 5.4.

Appendix A. Proof of Theorem 3.5

Recall that

$$\begin{aligned}(\Pi^n g)^i(y, p) = \sum_{i_1,\dots,i_n \in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} & g^{i_n}(y_n, f[y_{n-1}, \dots, y_1, y, p]) \\ &\times q_\bullet^{i,i_1}\cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n),\end{aligned}$$

for any $i \in S$, any $y \in \mathbb{R}^d$, and any $p \in \mathscr{P}$. The following decomposition holds:

$$\begin{aligned}(\Pi^n g)^i(y, p) &- (\Pi^n g)^{i'}(y', p') \\ &= (\Pi^n g)^{i'}(y, p) - (\Pi^n g)^{i'}(y', p') + (\Pi^n g)^i(y, p) - (\Pi^n g)^{i'}(y, p),\end{aligned}$$

and we estimate separately the two terms in the right-hand side.

To estimate the first term, we use the exponential forgetting of the prediction filter. Using Proposition 3.7 and estimate (7) yields

$$\begin{aligned}|(\Pi^n g)^{i'}(y, p) &- (\Pi^n g)^{i'}(y', p')| \\ \le \sum_{i_1,\dots,i_n \in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d} & |g^{i_n}(y_n, f[y_{n-1}, \dots, y_1, y, p]) \\ &- g^{i_n}(y_n, f[y_{n-1}, \dots, y_1, y', p'])| \\ \times q_\bullet^{i',i_1}\cdots q_\bullet^{i_{n-1},i_n} & b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n)\end{aligned}$$

$$\leq \sum_{i_1,\ldots,i_n\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}|g^{i_n}(y_n,f[y_{n-1},\ldots,y_1,f[y,p]]) - g^{i_n}(y_n,f[y_{n-1},\ldots,y_1,f[y',p']])|$$
$$\times q_\bullet^{i',i_1}\cdots q_\bullet^{i_{n-1},i_n}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n)$$
$$\leq 2\,\mathrm{Lip}(g)\rho_*^{n-1-r}\leq \frac{2}{1-R}\mathrm{Lip}(g)\rho_{\max}^{n-1}.$$

To estimate the second term, we use the geometric convergence of the *true* transition probabilities of the chain $\{X_n, n\geq 0\}$, which is a consequence of Assumption A, and we use again the exponential forgetting of the prediction filter. Recall that

$$(\Pi^n g)^i(y,p)-(\Pi^n g)^{i'}(y,p)$$
$$=\sum_{i_1,\ldots,i_n\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}g^{i_n}(y_n,f[y_{n-1},\ldots,y_1,p_1])[q_\bullet^{i,i_1}-q_\bullet^{i',i_1}]$$
$$\times q_\bullet^{i_1,i_2}\cdots q_\bullet^{i_{n-1},i_n}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n),$$

where the notation $p_1=f[y,p]$ is used. Notice that, for any integer m such that $m\leq n-1$, and for any sequence $z_1,\ldots,z_m\in\mathbb{R}^d$,

$$g^{i_n}(y_n,f[y_{n-1},\ldots,y_1,p_1])$$
$$=\sum_{k=1}^{m}[g^{i_n}(y_n,f[y_{n-1},\ldots,y_k,z_{k-1},\ldots,z_1,p_1]) - g^{i_n}(y_n,f[y_{n-1},\ldots,y_{k+1},z_k,\ldots,z_1,p_1])]$$
$$+g^{i_n}(y_n,f[y_{n-1},\ldots,y_{m+1},z_m,\ldots,z_1,p_1]).$$

Therefore

$$(\Pi^n g)^i(y,p)-(\Pi^n g)^{i'}(y,p)$$
$$=\sum_{k=1}^{m}\sum_{i_1,\ldots,i_n\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}[g^{i_n}(y_n,f[y_{n-1},\ldots,y_k,z_{k-1},\ldots,z_1,p_1]) - g^{i_n}(y_n,f[y_{n-1},\ldots,y_{k+1},z_k,\ldots,z_1,p_1])]$$
$$\times[q_\bullet^{i,i_1}-q_\bullet^{i',i_1}]q_\bullet^{i_1,i_2}\cdots q_\bullet^{i_{n-1},i_n}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n)$$
$$+\sum_{i_1,\ldots,i_n\in S}\int_{\mathbb{R}^d}\cdots\int_{\mathbb{R}^d}g^{i_n}(y_n,f[y_{n-1},\ldots,y_{m+1},z_m,\ldots,z_1,p_1])$$
$$\times[q_\bullet^{i,i_1}-q_\bullet^{i',i_1}]q_\bullet^{i_1,i_2}\cdots q_\bullet^{i_{n-1},i_n}b_\bullet^{i_1}(y_1)\cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n)$$

$$
\begin{aligned}
&= \sum_{k=1}^{m} \sum_{i_k,\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \\
&\quad \times \int_{\mathbb{R}^d} [g^{i_n}(y_n, f[y_{n-1},\dots,y_{k+1}, f[y_k, z_{k-1},\dots,z_1,p_1]]) \\
&\qquad\qquad - g^{i_n}(y_n, f[y_{n-1},\dots,y_{k+1}, f[z_k, z_{k-1},\dots,z_1,p_1]])] \\
&\quad \times [q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}] q^{i_k,i_{k+1}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} b^{i_k}_{\bullet}(y_k) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_k) \cdots \lambda(dy_n) \\
&\quad + \sum_{i_{m+1},\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g^{i_n}(y_n, f[y_{n-1},\dots,y_{m+1}, z_m,\dots,z_1,p_1]) \\
&\quad \times [q^{i,i_{m+1}}_{\bullet,m+1} - q^{i',i_{m+1}}_{\bullet,m+1}] q^{i_{m+1},i_{m+2}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} \\
&\quad \times b^{i_{m+1}}_{\bullet}(y_{m+1}) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_{m+1}) \cdots \lambda(dy_n).
\end{aligned}
$$

This holds for any integer m such that $m \leq n-1$. Taking now $m = n-r-1$, using Proposition 3.7 and estimate (1) yields

$$
\begin{aligned}
&|(\Pi^n g)^i(y,p) - (\Pi^n g)^{i'}(y,p)| \\
&\quad \leq \sum_{k=1}^{m} \sum_{i_k,\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \\
&\qquad \times \int_{\mathbb{R}^d} |g^{i_n}(y_n, f[y_{n-1},\dots,y_{k+1}, f[y_k, z_{k-1},\dots,z_1,p_1]]) \\
&\qquad\qquad - g^{i_n}(y_n, f[y_{n-1},\dots,y_{k+1}, f[z_k, z_{k-1},\dots,z_1,p_1]])| \\
&\qquad \times |q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}| q^{i_k,i_{k+1}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} b^{i_k}_{\bullet}(y_k) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_k) \cdots \lambda(dy_n) \\
&\qquad + \sum_{i_{m+1},\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_n}(y_n, f[y_{n-1},\dots,y_{m+1}, z_m,\dots,z_1,p_1])| \\
&\qquad \times |q^{i,i_{m+1}}_{\bullet,m+1} - q^{i',i_{m+1}}_{\bullet,m+1}| q^{i_{m+1},i_{m+2}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} \\
&\qquad\quad \times b^{i_{m+1}}_{\bullet}(y_{m+1}) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_{m+1}) \cdots \lambda(dy_n) \\
&\quad \leq 2\,\mathrm{Lip}(g) \sum_{k=1}^{m} \rho_*^{n-1-k-r} [2K_\bullet \rho_\bullet^k] + \mathrm{K}(g)[2K_\bullet \rho_\bullet^{m+1}],
\end{aligned}
$$

hence, using $m = n-r-1$ and estimate (7) yields

$$
\begin{aligned}
|(\Pi^n g)^i(y,p) - (\Pi^n g)^{i'}(y,p)| &\leq 4K_\bullet\,\mathrm{Lip}(g)(n-r-1)\rho_{\max}^{n-r-1} + 2K_\bullet\,\mathrm{K}(g)\rho_{\max}^{n-r} \\
&\leq \frac{4K_\bullet}{1-R}[\mathrm{Lip}(g) + \mathrm{K}(g)](n-1)\rho_{\max}^{n-1}. \qquad \blacksquare
\end{aligned}
$$

Appendix B. Proof of Proposition 5.6

For any sequence $i_m, \ldots, i_{n+1} \in S$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} &|g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p, w]) \\
&- g^{i_{n+1}}(y_{n+1}, f \otimes F[y_n, \ldots, y_m, p', w'])| \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \\
\leq & \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \mathrm{Lip}(g^{i_{n+1}}, y_{n+1}) \| F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w'] \| \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \\
&+ \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \mathrm{Lip}(g^{i_{n+1}}, y_{n+1}) \| f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p'] \| \\
&\times (1 + \| F[y_n, \ldots, y_m, p, w] \| + \| F[y_n, \ldots, y_m, p', w'] \|) \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_{n+1}}(y_{n+1}) \lambda(dy_m) \cdots \lambda(dy_{n+1}) \\
\leq & \,\mathrm{Lip}(g) \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \| F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w'] \| \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n) \\
&+ \mathrm{Lip}(g) \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \| f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p'] \| \\
&\times (1 + \| F[y_n, \ldots, y_m, p, w] \| + \| F[y_n, \ldots, y_m, p', w'] \|) \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n).
\end{aligned}$$

To estimate the first term, we use the estimate in Theorem 5.9 in the companion paper [LM4], and we notice that each term in the sum is a product of factors which span disjoint index sets, hence integration is straightforward, and yields for any sequence $i_m, \ldots, i_n \in S$,

$$\begin{aligned}
\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} & \| F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w'] \| \\
&\times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n) \\
\leq & \tfrac{1}{2}(\varepsilon^{-2r} \Delta_2^r + 1)(1 - R)^{[n,m]-1} (\|w\| + \|w'\|) + 2\,\mathrm{Lip}(u)(1 - R)^{[n-1,m]} \\
&+ 2\,\mathrm{Lip}(u) \varepsilon^{-r} \Delta^r \sum_{l=m}^{n-1} (1 - R)^{[n,l+1]+[l-1,m]-1} \\
&+ \mathrm{K}(u)(\varepsilon^{-2r} \Delta_2^r - 1) \sum_{l=m}^{n-1} (1 - R)^{[n,l+1]+[l-1,m]-1}.
\end{aligned}$$

Using the upper bound (6) yields

$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w']\| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n)$$
$$\leq \tfrac{1}{2}(\varepsilon^{-2r}\Delta_2^r + 1)\rho_*^{n-m-2r}(\|w\| + \|w'\|) + 2\,\mathrm{Lip}(u)\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r} + \mathrm{K}(u)(\varepsilon^{-2r}\Delta_2^r - 1)(n-m)\rho_*^{n-m-3r}.$$

To estimate the second term, we use the estimate of Proposition 5.6 in the companion paper [LM4], and we notice that each term in the sum is a product of factors which span disjoint index sets, hence integration is straightforward, and yields for any sequence $i_m, \ldots, i_n \in S$,

$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \, \|F[y_n, \ldots, y_m, p, w]\| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n)$$
$$\leq 2\varepsilon^{-r}\Delta^r(1-R)^{[n,m]-1}\|w\| + 2\,\mathrm{K}(u)(1-R)^{[n-1,m]} + 2\,\mathrm{K}(u)\varepsilon^{-r}\Delta^r \sum_{l=m}^{n-1} (1-R)^{[n,l+1]+[l-1,m]-1}.$$

Using the upper bound (6) yields

$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \, \|F[y_n, \ldots, y_m, p, w]\| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n)$$
$$\leq 2\varepsilon^{-r}\Delta^r\rho_*^{n-m-2r}\|w\| + 2\,\mathrm{K}(u)\,\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r}.$$

Combining the above estimates yields

$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|F[y_n, \ldots, y_m, p, w] - F[y_n, \ldots, y_m, p', w']\| \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n)$$
$$+ \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|f[y_n, \ldots, y_m, p] - f[y_n, \ldots, y_m, p']\| \times (1 + \|F[y_n, \ldots, y_m, p, w]\| + \|F[y_n, \ldots, y_m, p', w']\|) \times b_\bullet^{i_m}(y_m) \cdots b_\bullet^{i_n}(y_n) \lambda(dy_m) \cdots \lambda(dy_n)$$
$$\leq \tfrac{1}{2}(\varepsilon^{-2r}\Delta_2^r + 1)\rho_*^{n-m-2r}(\|w\| + \|w'\|) + 2\,\mathrm{Lip}(u)\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r} + \mathrm{K}(u)(\varepsilon^{-2r}\Delta_2^r - 1)(n-m)\rho_*^{n-m-3r} + 2\rho_*^{n-m+1-r}$$

$$
\begin{aligned}
&+ 2\varepsilon^{-r}\Delta^r \rho_*^{n-m-2r}\|w\| + 2\,\mathrm{K}(u)\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r} \\
&+ 2\varepsilon^{-r}\Delta^r \rho_*^{n-m-2r}\|w'\| + 2\,\mathrm{K}(u)\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r} \\
\le\ &\tfrac{1}{2}(\varepsilon^{-2r}\Delta_2^r + 1 + 4\varepsilon^{-r}\Delta^r)\rho_*^{n-m-2r}(\|w\| + \|w'\|) + 2\rho_*^{n-m+1-r} \\
&+ 2\,\mathrm{Lip}(u)\varepsilon^{-r}\Delta^r(n-m+1)\rho_*^{n-m-3r} \\
&+ \mathrm{K}(u)(\varepsilon^{-2r}\Delta_2^r - 1 + 4\varepsilon^{-r}\Delta^r)(n-m+1)\rho_*^{n-m-3r}. \qquad \blacksquare
\end{aligned}
$$

Appendix C. Proof of Theorem 5.4

The proof follows the same lines as the proof of Theorem 3.5. Recall that

$$
\begin{aligned}
(\Pi^n g)^i(y,p,w) = \sum_{i_1,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} & g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1,y,p,w]) \\
& \times q_\bullet^{i,i_1} \cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n),
\end{aligned}
$$

for any $i \in S$, any $y \in \mathbb{R}^d$, any $p \in \mathscr{P}$, and any $w \in \Sigma$. The following decomposition holds:

$$
\begin{aligned}
&(\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y',p',w') \\
&\quad = (\Pi^n g)^{i'}(y,p,w) - (\Pi^n g)^{i'}(y',p',w') \\
&\qquad + (\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y,p,w),
\end{aligned}
$$

and we estimate separately the two terms in the right-hand side.

To estimate the first term, we use the exponential forgetting of the prediction filter and its gradient. Using Proposition 5.6 yields

$$
\begin{aligned}
&|(\Pi^n g)^{i'}(y,p,w) - (\Pi^n g)^{i'}(y',p',w')| \\
&\quad \le \sum_{i_1,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1,y,p,w]) \\
&\qquad\qquad - g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1,y',p',w'])| \\
&\qquad \times q_\bullet^{i',i_1} \cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n) \\
&\quad \le \sum_{i_1,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} |g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1, f \otimes F[y,p,w]]) \\
&\qquad\qquad - g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1, f \otimes F[y',p',w']])| \\
&\qquad \times q_\bullet^{i',i_1} \cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n) \\
&\quad \le C'\varepsilon^{-2r}\Delta_2^r\,\mathrm{Lip}(g)[(n-1) + \|F[y,p,w]\| + \|F[y',p',w']\|]\rho_*^{n-2-3r},
\end{aligned}
$$

and using estimate (7) yields

$$|(\Pi^n g)^{i'}(y,p,w) - (\Pi^n g)^{i'}(y',p',w')| \leq \frac{C'}{(1-R)^3}\varepsilon^{-2r}\Delta_2^r \operatorname{Lip}(g)[1 + \|F[y,p,w]\| + \|F[y',p',w']\|]n\rho_{\max}^{n-2},$$

where the constant C' depends only on $\operatorname{Lip}(u)$ and $\mathrm{K}(u)$.

To estimate the second term, we use the geometric convergence of the *true* transition probabilities of the chain $\{X_n, n \geq 0\}$, which is a consequence of Assumption A, and we use again the exponential forgetting of the prediction filter and its gradient. Recall that

$$(\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y,p,w) = \sum_{i_1,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_1,p_1,w_1])[q_\bullet^{i,i_1} - q_\bullet^{i',i_1}] \times q_\bullet^{i_1,i_2} \cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_1}(y_1) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_1)\cdots\lambda(dy_n),$$

where the notation $(p_1,w_1) = f \otimes F[y,p,w] = (f[y,p], F[y,p,w])$ is used. As in the proof of Theorem 3.5 above, for any integer m such that $m \leq n-1$, and for any sequence $z_1,\ldots,z_m \in \mathbb{R}^d$,

$$\begin{aligned}(\Pi^n g)^i(y,p,w) &- (\Pi^n g)^{i'}(y,p,w) \\ &= \sum_{k=1}^m \sum_{i_k,\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \\ &\times \int_{\mathbb{R}^d} [g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_{k+1}, f \otimes F[y_k, z_{k-1},\ldots,z_1,p_1,w_1]]) \\ &\qquad - g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_{k+1}, f \otimes F[z_k, z_{k-1},\ldots,z_1,p_1,w_1]])] \\ &\times [q_{\bullet,k}^{i,i_k} - q_{\bullet,k}^{i',i_k}] q_\bullet^{i_k,i_{k+1}} \cdots q_\bullet^{i_{n-1},i_n} b_\bullet^{i_k}(y_k) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_k)\cdots\lambda(dy_n) \\ &+ \sum_{i_{m+1},\ldots,i_n \in S} \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g^{i_n}(y_n, f \otimes F[y_{n-1},\ldots,y_{m+1},z_m,\ldots,z_1,p_1,w_1]) \\ &\times [q_{\bullet,m+1}^{i,i_{m+1}} - q_{\bullet,m+1}^{i',i_{m+1}}] q_\bullet^{i_{m+1},i_{m+2}} \cdots q_\bullet^{i_{n-1},i_n} \\ &\times b_\bullet^{i_{m+1}}(y_{m+1}) \cdots b_\bullet^{i_n}(y_n)\lambda(dy_{m+1})\cdots\lambda(dy_n).\end{aligned}$$

This holds for any integer m such that $m \leq n-1$. Taking now $m = n - 3r + 1$, using Propositions 5.6 and 5.7 and estimate (1) yields

$$
\begin{aligned}
&|(\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y,p,w)| \\
&\quad \le \sum_{k=1}^{m} \sum_{i_k,\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \\
&\qquad \times \int_{\mathbb{R}^d} |g^{i_n}(y_n, f \otimes F[y_{n-1},\dots,y_{k+1}, f \otimes F[y_k, z_{k-1},\dots,z_1,p_1,w_1]]) \\
&\qquad\qquad - g^{i_n}(y_n, f \otimes F[y_{n-1},\dots,y_{k+1}, f \otimes F[z_k, z_{k-1},\dots,z_1,p_1,w_1]])| \\
&\qquad \times |q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}| q^{i_k,i_{k+1}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} b^{i_k}_{\bullet}(y_k) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_k) \cdots \lambda(dy_n) \\
&\qquad + \sum_{i_{m+1},\dots,i_n \in S} \int_{\mathbb{R}^d} \cdots \\
&\qquad \times \int_{\mathbb{R}^d} |g^{i_n}(y_n, f \otimes F[y_{n-1},\dots,y_{m+1}, f \otimes F[z_m,\dots,z_1,p_1,w_1]])| \\
&\qquad \times |q^{i,i_{m+1}}_{\bullet,m+1} - q^{i',i_{m+1}}_{\bullet,m+1}| q^{i_{m+1},i_{m+2}}_{\bullet} \cdots q^{i_{n-1},i_n}_{\bullet} \\
&\qquad \times b^{i_{m+1}}_{\bullet}(y_{m+1}) \cdots b^{i_n}_{\bullet}(y_n) \lambda(dy_{m+1}) \cdots \lambda(dy_n) \\
&\quad \le C' \varepsilon^{-2r} \Delta_2^r \operatorname{Lip}(g) \sum_{k=1}^{m} \sum_{i_k \in S} |q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}| \\
&\qquad \times \Bigg[\int_{\mathbb{R}^d} (\|F[z_k, z_{k-1},\dots,z_1,p_1,w_1]\| + \|F[y_k, z_{k-1},\dots,z_1,p_1,w_1]\|) \\
&\qquad\qquad \times b^{i_k}_{\bullet}(y_k)\lambda(dy_k) + (n-1-k) \Bigg] \rho_*^{n-k-2-3r} \\
&\qquad + C'' \varepsilon^{-r} \Delta^r \operatorname{K}(g) \sum_{i_{m+1} \in S} |q^{i,i_{m+1}}_{\bullet,m+1} - q^{i',i_{m+1}}_{\bullet,m+1}| \\
&\qquad \times [(n-1-m) + \|F[z_m,\dots,z_1,p_1,w_1]\|] \\
&\quad \le C' \varepsilon^{-2r} \Delta_2^r \operatorname{Lip}(g) \sum_{k=1}^{m} \Bigg[\sum_{i_k \in S} |q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}| \\
&\qquad \times \int_{\mathbb{R}^d} (\|F[z_k, z_{k-1},\dots,z_1,p_1,w_1]\| + \|F[y_k, z_{k-1},\dots,z_1,p_1,w_1]\|) \\
&\qquad \times b^{i_k}_{\bullet}(y_k)\lambda(dy_k) + (n-1-k)[2K_\bullet \rho_\bullet^k] \Bigg] \rho_*^{n-k-2-3r} \\
&\qquad + C'' \varepsilon^{-r} \Delta^r \operatorname{K}(g) [2K_\bullet \rho_\bullet^{m+1}] [(n-1-m) + \|F[z_m,\dots,z_1,p_1,w_1]\|],
\end{aligned}
$$

where the constant C' depends only on $\operatorname{Lip}(u)$ and $\operatorname{K}(u)$, and the constant C'' depends only on $\operatorname{K}(u)$. These estimates hold for any sequence $z_1,\dots,z_m \in \mathbb{R}^d$.

Integrating and using estimate (11) yields, for any sequence $j_1, \ldots, j_m \in S$,

$$\sum_{i_k \in S} |q^{i,i_k}_{\bullet,k} - q^{i',i_k}_{\bullet,k}| \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} (\|F[z_k, z_{k-1}, \ldots, z_1, p_1, w_1]\| + \|F[y_k, z_{k-1}, \ldots, z_1, p_1, w_1]\|) \times b^{i_k}_\bullet(y_k)\lambda(dy_k) b^{j_1}_\bullet(z_1) \cdots b^{j_m}_\bullet(z_m)\lambda(dz_1)\cdots\lambda(dz_m) \le 2C''\varepsilon^{-r}\Delta^r[k + \|w_1\|][2K_\bullet \rho^k_\bullet],$$

and similarly

$$\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \|F[z_m, \ldots, z_1, p_1, w_1]\| b^{j_1}_\bullet(z_1) \cdots b^{j_m}_\bullet(z_m)\lambda(dz_m)\cdots\lambda(dz_m) \le C''\varepsilon^{-r}\Delta^r[m + \|w_1\|],$$

where the constant C'' depends only on $\mathrm{K}(u)$. Using these estimates yields

$$\begin{aligned}
&|(\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y,p,w)| \\
&\le 2C'\varepsilon^{-2r}\Delta_2^r K_\bullet \,\mathrm{Lip}(g)\left[\sum_{k=1}^m (n-1-k) + 2C''\varepsilon^{-r}\Delta^r \sum_{k=1}^m [k + \|w_1\|]\right]\rho_{\max}^{n-2-3r} \\
&\quad + 2C''\varepsilon^{-r}\Delta^r K_\bullet \,\mathrm{K}(g)[(n-1-m) + C''\varepsilon^{-r}\Delta^r[m + \|w_1\|]]\rho_{\max}^{m+1} \\
&\le 2C'\varepsilon^{-2r}\Delta_2^r K_\bullet \,\mathrm{Lip}(g)[\tfrac{1}{2}m(2n-m-3) \\
&\quad + 2C''\varepsilon^{-r}\Delta^r[\tfrac{1}{2}(m+1) + \|w_1\|]m]\rho_{\max}^{n-2-3r} \\
&\quad + 2C''\varepsilon^{-r}\Delta^r K_\bullet \,\mathrm{K}(g)[(n-1-m) + C''\varepsilon^{-r}\Delta^r[m + \|w_1\|]]\rho_{\max}^{m+1},
\end{aligned}$$

hence, using $m = n - 3r + 1$ and estimate (7) yields

$$\begin{aligned}
&|(\Pi^n g)^i(y,p,w) - (\Pi^n g)^{i'}(y,p,w)| \\
&\le 2C'\varepsilon^{-2r}\Delta_2^r \frac{K_\bullet}{(1-R)^3}\mathrm{Lip}(g)[\tfrac{1}{2}n(n-1) \\
&\quad + 2C''\varepsilon^{-r}\Delta^r[\tfrac{1}{2}(n-1) + \|w_1\|]n]\rho_{\max}^{n-2} \\
&\quad + 2C''\varepsilon^{-r}\Delta^r \frac{K_\bullet}{(1-R)^3}\mathrm{K}(g)[(3r-2) \\
&\quad + C''\varepsilon^{-r}\Delta^r[(n-3r+1) + \|w_1\|]]\rho_{\max}^{n-1} \\
&\le \frac{C'''K_\bullet}{(1-R)^3}\varepsilon^{-3r}\Delta_2^{3r/2}[\mathrm{Lip}(g) + \mathrm{K}(g)][1 + \|F[y,p,w]\|]n(n-1)\rho_{\max}^{n-2},
\end{aligned}$$

where the constant C''' depends only on $\mathrm{Lip}(u)$ and $\mathrm{K}(u)$. ■

References

[AM] A. Arapostathis and S. I. Marcus. Analysis of an identification algorithm arising in the adaptive estimation of Markov chains. *Mathematics of Control, Signals, and Systems*, 3(1):1–29, 1990.

[AZ] R. Atar and O. Zeitouni. Lyapunov exponents for finite state nonlinear filtering. *SIAM Journal on Control and Optimization*, 35(1):36–55, January 1997.

[BMP] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Volume 22 of Applications of Mathematics. Springer-Verlag, New York, 1990.

[B1] P. Bougerol. Comparaison des exposants de Lyapounov des processus markoviens multiplicatifs. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 24(4):439–489, 1988.

[B2] P. Bougerol. Théorèmes limites pour les systèmes linéaires à coefficients markoviens. *Probability Theory and Related Fields*, 78(2):193–221, 1988.

[DS] G. B. Di Masi and L. Stettner. On adaptive control of a partially observed Markov chain. *Applicationes Mathematicæ*, 22(2):165–180, 1994.

[D1] R. M. Dudley. Convergence of Baire measures. *Studia Mathematica*, XXVII:251–268, 1966.

[D2] R. M. Dudley. Distance of probability measures and random variables. *The Annals of Mathematical Statistics*, 39(5):1563–1572, October 1968.

[FK] H. Furstenberg and H. Kesten. Product of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, June 1960.

[G] Y. Guivarc'h. Exposants caractéristiques des produits de matrices aléatoires en dépendance markovienne. In H. Heyer, editor, *Probability Measures on Groups VII, Oberwolfach* 1983, pages 161–181. Volume 1064 of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1984.

[H] H. Hennion. Limit theorems for products of positive random matrices. *The Annals of Probability*, 25(4):1545–1587, October 1997.

[K1] T. Kaijser. A limit theorem for partially observed Markov chains. *The Annals of Probability*, 3(4):677–696, August 1975.

[K2] H. Kunita. Asymptotic behavior of the nonlinear filtering errors of Markov processes. *Journal of Multivariable Analysis*, 1(4):365–393, 1971.

[LM1] F. Le Gland and L. Mevel. Geometric ergodicity in hidden Markov models. Rapport de Recherche 2991, INRIA, September 1996.
ftp://ftp.inria.fr/INRIA/publication/RR/RR-2991.ps.gz.

[LM2] F. Le Gland and L. Mevel. Asymptotic behaviour of the MLE in hidden Markov models. In *Proceedings of the* 4*th European Control Conference, Bruxelles*, 1997, Paper FRA–F6.

[LM3] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in HMMs. In *Proceedings of the* 36*th IEEE Conference on Decision and Control, San Diego*, 1997, pages 537–542.

[LM4] F. Le Gland and L. Mevel. Basic properties of the projective product, with application to products of column-allowable nonnegative matrices. *Mathematics of Control, Signals, and Systems*, this issue, pp. 41–62.

[S1] Ł. Stettner. On invariant measures of filtering processes. In N. Christopeit, K. Helmes, and M. Kohlmann, editors, *Stochastic Differential Systems, Bad Honnef, 1988*, pages 279–292. Volume 126 of Lecture Notes in Control and Information Sciences. Springer-Verlag, Berlin, 1989.

[S2] Ł. Stettner. Invariant measures for the pair: state, approximate filtering process. *Colloquium Mathematicum*, LXII(2):347–351, 1991.