

On discrete time ergodic filters with wrong initial data, 2

M. L. Kleptsyna* & A. Yu. Veretennikov†

December 3, 2007

Abstract

For a class of non-uniformly ergodic Markov chains satisfying exponential or polynomial beta-mixing, under observations subject to an IID noise, it is shown that wrong initial data is forgotten in the mean total variation topology, with a certain exponential or polynomial rate. It is allowed that the density of the noise in the signal may vanish.

1 Introduction

We consider a discrete time filter for a Hidden Markov chain (X_n) with values in the Euclidean space R^d , with conditionally Markov observations (Y_n) from R^ℓ , satisfying the system

$$X_{n+1} = X_n + b(X_n) + \sigma(X_n)\xi_{n+1}, \quad (n \geq 0), \quad (1)$$

$$Y_n = h(X_n) + V_n \quad (n \geq 1), \quad (2)$$

where (ξ_n, V_n) is a sequence of IID random vectors of dimension $d + \ell$ with densities $q_\xi(x)q_V(y)$, $b(\cdot)$ is a d -dimensional vector-function, $\sigma(\cdot)$ a $d \times d$ matrix-function, $h(\cdot)$ an ℓ -dimensional vector-function. We assume that the exact initial distribution of X_0 denoted by μ_0 , is known with some error. The main problem addressed in this paper is whether or not this error is forgotten by the optimal filtering algorithm in the long run.

*Université du Maine, Le Mans, France, e-mail: Marina.Kleptsyna at univ.lemans.fr

†University of Leeds, UK, and Institute of Information Transmission Problems, Moscow, Russia, e-mail: A.Veretennikov at leeds.ac.uk

For the review of recent and earlier achievements see [7], [8]. The most important papers for the authors were [1], [2], and [10]. A more complete, – although inevitably not full, – list of relevant references could be found, e.g., in [7]. An important tool in the proof is another distance between equivalent measures, *Birkhoff’s metric*, see below. The latter is known as a useful tool in the theory of positive operators (see, e.g., [9]), and has proved to be very helpful in [7] and in several earlier papers on the subject; its use in this part of filtering theory has been proposed in [1]. In the present paper, we relax considerably the “local mixing condition” of the kernels in compare to [7] and [10], by providing, in particular, an easy way to include into considerations distributions with possibly vanishing density of the noise in the signal. For the purpose of establishing *non-conditional* mixing rate, instead of the frequently used mixing condition

$$\epsilon\lambda(\cdot) \leq Q(x, \cdot) \leq \epsilon^{-1}\lambda(\cdot), \quad (3)$$

which is often not optimal, and in some important cases such as non-degenerate diffusion with a non-smooth diffusion coefficient just fails, one can assume the following integral form of the Doeblin–Doob type condition, due to Dobrushin (cf. [3]),

$$\inf_{\tilde{x}, x} \int \left(\frac{Q(\tilde{x}, dx')}{Q(x, dx')} \wedge 1 \right) Q(x, dx') := \nu > 0,$$

or its localized form, which would be appropriate to call *local uniform Doeblin–Doob–Dobrushin condition*,

$$\inf_{\tilde{x}, x \in B} \int_{B'} \left(\frac{Q(\tilde{x}, dx')}{Q(x, dx')} \wedge 1 \right) Q(x, dx') =: \nu(B, B') > 0, \quad (4)$$

for appropriate sets B and B' . If the latter integral and infimum in (4) are both taken over the whole space, the value $1 - \nu$ is called Dobrushin’s coefficient of ergodicity [3]. The latter condition with arbitrary B has been used earlier in the second author’s papers on mixing since early nineties. This condition is considerably weaker than (3), and at the same time it provides better constants in the bounds of convergence and mixing; this will be shown in one of the sections below. The condition (3) is frequently used for establishing mixing or convergence rate to equilibrium distribution, see, e.g., [4]. It does provide uniform bounds for such convergence to equilibrium (and mixing) on the class of processes (see again [4]). However, the use of (4) instead gives better constants in the bounds, and wider class of processes is covered. It is worth mentioning that the most general Doeblin – Doob condition does provide an exponential convergence rate, however, it does *not* provide any *uniform* convergence rate on any class of processes under this sort of condition. At least, the technique in [4] in the general case which is based on Lebesgue differentiation cannot give

any uniform convergence. Hence, the condition (4) may be considered as a generalization of (3) in all meanings. Concerning non-conditional mixing see [5], [6], [11] – [13].

There is an *open problem*: are there weaker conditions that still allow to control the rate of convergence and mixing for some appropriate classes of processes? Perhaps, some further localization of the condition (4) could be useful.

For the *conditional* setting of this paper we will use the following version of the mixing assumption intermediate between (3) and (4): for every $R > 0$ large enough, there exists $L_R > 0$ such that

$$p_R := \inf_{x \in B_R} Q(x, A_R) > 1/2,$$

where $A_R \subset B_R$ satisfies

$$0 < a_R^- \leq q(x, x') \leq a_R^+ < \infty, \quad \forall x \in B_R, x' \in A_R,$$

see below the assumption (A3). Using this condition, we establish results on forgetting property of the filter under less restrictive assumptions in compare to [8]. Notice that a more general condition (4) fails to work when we are to verify contraction in the Birkhoff metric, see below.

Further generalization of recurrence conditions in terms of Lyapunov functions is possible, but the point of view of the authors is that our type of conditions allows an easier verification and, hence, is more appropriate. The meaning of the condition (5) is that for every *couple* (x, \tilde{x}) from some appropriate ball, their transition measures should have components that are comparable in the sense of equivalence with bounded derivatives. There is no need to require such a comparability for the whole family of transition measures with respect to some reference measure, but only couples of those measures are relevant.

The paper is arranged as follows: the section 2 contains the assumptions and the main result; the section 3 is devoted to the proof of the main result. The paper may be considered as a continuation of [7] which explains the title; however, it may be read independently. In some cases we used reduced form of the calculus where it is similar to that in [7].

2 Assumptions, main result, auxiliaries

2.1 Assumptions

(A1) We assume that

$$0 < \inf_x \inf_{|\lambda|=1} \lambda^* \sigma \sigma^*(x) \lambda \leq \sup_x \sup_{|\lambda|=1} \lambda^* \sigma \sigma^*(x) \lambda < \infty,$$

where $\lambda \in R^d$, the function b is locally bounded, and there exist $p \in \{0, 1\}$, $M > 0$ and $r \in (0, +\infty]$ such that

$$\left(\frac{|b(x)|}{|x|} - 1\right) |x|^{1+p} \leq -r, \quad |x| \geq M; \quad (5)$$

if $p = 1$ then we understand this as a limit with $r = +\infty$, that is,

$$\limsup_{|x| \rightarrow \infty} \left(\frac{|b(x)|}{|x|} - 1\right) |x|^2 = -\infty. \quad (6)$$

(A2) The noise (ξ_n) is a sequence of i.i.d. random vectors with

$$E\xi_k = 0;$$

and such that in the case $p = 0$,

$$E \exp(c|\xi_1|) < \infty,$$

while in the case $p = 1$, for every $m > 0$,

$$E|\xi_1|^m < \infty.$$

The noise (V_n) is an IID centered sequence; the density q_V is assumed to be positive everywhere.

The function h is locally bounded.

(A3) In terms of the *transition density* of the signal process,

$$q(x, x') dx' := \frac{1}{\sqrt{\det(\sigma^* \sigma)(x)}} q_\xi(\sigma(x)^{-1}(x' - x - g(x))) dx',$$

we assume for every R large enough there exists a Borel set $A_R \subset B_R$ such that

$$0 < a_R^- := \inf_{|x| \leq R} \inf_{x' \in A_R} q(x, x') \leq \sup_{|x| \leq R} \sup_{x' \in A_R} q(x, x') =: a_R^+ < \infty, \quad (7)$$

$$\& p_R = \inf_{|x| \leq R} P_x(X_1 \in A_R) > 1/2. \quad (8)$$

In particular, the latter condition does not include any density q_ξ with a compact support. However, it does include, e.g., many continuous densities with unbounded support, and many more, because continuity here may be relaxed. For example, for finite state Markov chain it is sufficient to have one column with entries greater than $1/2$ in the transition matrix. At the same time, it is worth mentioning that a density with a *compact* support may satisfy the condition (A3) with *some* positive R ; in this case, *some* polynomial rate of convergence may be achieved.

(A4) The measure μ_0 is absolute continuous with respect to ν_0 , and, moreover,

$$\left\| \frac{d\mu_0}{d\nu_0} \right\|_{L^\infty(\nu_0)} < \infty. \quad (9)$$

Moreover, both initial measures μ_0 and ν_0 possess some exponential moment, that is, there exists $c > 0$ such that

$$\int e^{c|x|} (\mu_0(dx) + \nu_0(dx)) < \infty. \quad (10)$$

For $p = 1$ the latter condition may be relaxed.

2.2 Setting and Main Results

Using the Bayes formula, the exact a posteriori filtering conditional measure can be represented as a probability measure for any Y , via the following non-linear operator \bar{S}_n^{Y, μ_0} , applied to the initial measure μ_0 ,

$$\begin{aligned} P_{\mu_0}(X_n \in dx_n \mid Y_1, \dots, Y_n) &= \int \prod_{i=1}^n Q(x_{i-1}, dx_i) c_i^{\mu_0} \Psi(x_i, Y_i) \mu_0(dx_0) \\ &= d_n^{\mu_0} \int \prod_{i=1}^n Q(x_{i-1}, dx_i) \Psi(x_i, Y_i) \mu_0(dx_0) =: \mu_0 \bar{S}_n^{Y, \mu_0}(dx_n). \end{aligned} \quad (11)$$

Here $\Psi(x_i, y_i)$ is a conditional density of Y_i at y_i , given $X_i = x_i$, and $Q(x, dx')$ is a transition kernel for the Markov chain X_n , $n \geq 0$. The random normalization constant $d_i^{\mu_0}$ is defined as follows,

$$d_i^{\mu_0} = \left(E_{\mu_0} \left(\prod_{j=1}^i \Psi(X_j, y_j) \right) \Big|_{y_1=Y_1, \dots, y_i=Y_i} \right)^{-1},$$

and, correspondingly,

$$c_i^{\mu_0} = \frac{d_i^{\mu_0}}{d_{i-1}^{\mu_0}} = \frac{E_{\mu_0} \left(\prod_{j=1}^{i-1} \Psi(X_j, y_j) \right) \Big|_{y_1=Y_1, \dots, y_{i-1}=Y_{i-1}}}{E_{\mu_0} \left(\prod_{j=1}^i \Psi(X_j, y_j) \right) \Big|_{y_1=Y_1, \dots, y_i=Y_i}}.$$

Now, the “wrong initialization” problem can be formulated more precisely as follows. One does not know the measure μ_0 exactly, but only some its approximation ν_0 . Hence, one plugs in the observed values Y 's and this new measure

ν_0 into the formula (11). The problem is whether in the long run the difference between the conditional measures provided by the algorithms with the exact and wrong initial data converges to zero in some suitable topology. However, even before we pose this question about convergence, we shall decide whether this operation of using ν_0 instead of μ_0 is well-defined. In which case it is well-defined and in which it is not? The answer is that it is not well-defined if and only if our actually observed vector \tilde{Y}_n is impossible under ν_0 for some n , or, equivalently, if the vector-value (X_0, \dots, X_n) starting from the distribution ν_0 is impossible under the observed \tilde{Y}_n for some n . Since clearly any value of (X_0, \dots, X_n) with $X_0 \in \text{supp}(\mu_0)$ is possible, we have a sufficient condition for our operation to be well-defined, $\text{supp}(\nu_0) \subset \text{supp}(\mu_0)$, or, equivalently,

$$\nu_0 \ll \mu_0. \tag{12}$$

This condition is sufficient whatever all other distributions are. Notice that in many papers on the subject this is, indeed, assumed. However, it is not necessary if we impose some other additional requirements, e.g., if the density of V_1 is positive everywhere, which we have assumed in (A2).

Another issue is that while using the Birkhoff metric and induction we will need *equivalent* measures with bounded derivatives, see (20) below. So it looks as if (12) should have been assumed, at least. However, recall that induction can be started not necessarily from zero. On the other hand, after the first application of the “mixing inequality” (21) we will get comparable measures, that is, equivalent measures with bounded derivatives, as required. Thus, we can start our induction (20) from $n = 1$. This is why the condition (12), which seems so natural and nearly indispensable, in fact, may be relaxed. Nevertheless, in some further studies it will be very desirable.

Now we shall explain how one can interpret this setting in a probabilistic way, using again some Markov dynamics and conditioning. In fact, for the initial distribution ν_0 , we have another sequences of measures and observations,

$$d_n^{\nu_0} \int \prod_{i=1}^n Q(x_{i-1}, dx_i) \Psi(x_i, \tilde{Y}_i) \nu_0(dx_0) = \nu_0 \tilde{S}_n^{\tilde{Y}, \nu_0}(dx_n) \equiv \nu_n(dx_n).$$

This can be, indeed, regarded as another conditional expectation, for the same Markov process starting from another initial distribution ν_0 , given some new observations $(\tilde{Y}_1, \dots, \tilde{Y}_n)$. Without losing a generality, we can and will assume that this pair, (\tilde{X}, \tilde{Y}) , is defined on some *independent probability space*; we will not change our notation for the probability measure, nor for expectation, though, both now apply to the process $(X, Y, \tilde{X}, \tilde{Y})$. However, due to the setting, only original observations Y are available, so that we are obliged to identify $(\tilde{Y}_1, \dots, \tilde{Y}_n)$ with (Y_1, \dots, Y_n) , that is, we *keep the original observations* that

have risen from the original initial data μ_0 , as though they were initialized by its substitution ν_0 . The result, $\nu_0 \bar{S}_n^{Y, \nu_0}$, is still some conditional probability, namely, the conditional distribution of ν_n given $(\tilde{Y}_1, \dots, \tilde{Y}_n)$, after the values $(\tilde{Y}_1, \dots, \tilde{Y}_n)$ have been replaced by (Y_1, \dots, Y_n) . This operation is well defined almost surely with respect to the measure P_{μ_0} , due to our assumptions on the density q_V .

The *main question* here is about a discrepancy of the filter with a wrong measure ν_0 instead of μ_0 and the exact one, or, in other words, about the difference of the two measures,

$$(\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0})(dx_n),$$

whether it is reasonably small for large values of n . We will be interested in the distance in the mean total variation norm with respect to the original initial measure μ_0 .

Theorem 1 *1. Under the assumptions (A1) – (A3) above, the following bounds hold true:*

$$E_{\mu_0} \|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq \begin{cases} C_m n^{-m}, & p = 1, \quad \forall m > 0, \\ C \exp(-cn), & p = 0. \end{cases} \quad (13)$$

2. In addition, the following pathwise inequalities hold true:

(i) *If*

$$E_{\mu_0} \|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq C n^{-m},$$

then, for every $m' < m$, not necessarily integer, there exists a (random) n_0 such that

$$\|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq n^{-m'+1}, \quad n \geq n_0.$$

(ii) *If*

$$E_{\mu_0} \|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq C \exp(-cn),$$

then for any $c' < c$, there exists a (random) n_0 such that

$$\|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq C \exp(-c'n), \quad n \geq n_0.$$

Remark 1 *Notice that one can state uniform bounds for appropriate classes of problems in the Theorem 1.*

Remark 2 *If in (A4) we assume additionally*

$$0 < C_1^{-1} := \operatorname{ess\,inf} \frac{d\mu_0}{d\nu_0} \leq \operatorname{ess\,sup} \frac{d\mu_0}{d\nu_0} \leq C_1 < \infty, \quad (14)$$

then it follows from the same calculus in the next section that

$$E_{\mu_0} \|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \leq \begin{cases} C_m n^{-m} \rho(\mu_0, \nu_0), & p = 1, \\ C \exp(-cn) \rho(\mu_0, \nu_0), & p = 0. \end{cases} \quad \forall m > 0, \quad (15)$$

where ρ is the Birkhoff metric, see (18) below.

3 The proof of Theorem 1

1. First of all, let us introduce some indicators. Note that in this proof, x stands for the whole sequence (x_0, x_1, \dots, x_n) , and likewise for \tilde{x} , and the same for the random sequences X , \tilde{X} , and Y . As suggested above in the setting, we consider independent couples (X, Y) and (\tilde{X}, \tilde{Y}) , with initial distributions of the first components, $\mathcal{L}(X_0) = \mu_0$ and $\mathcal{L}(\tilde{X}_0) = \nu_0$. For every $i \geq 0$, let $M_i := \max(|X_i|, |\tilde{X}_i|)$. For fixed R and n , we denote by δ , $\delta = (\delta_0, \delta_1, \dots, \delta_n)$ a (non-random) vector of dimension $n+1$ with coordinates 1 or 0 at every place, and by δ' , $\delta' = (\delta'_1, \delta'_2, \dots, \delta'_n)$ any other non-random vector of dimension n with coordinates 1 or 0 such that all zeros in δ remain zeros in δ' . More than that, δ'_i may be equal to 1 only if $\delta_i = 1$, that is,

$$(\delta'_i = 1) \implies (\delta_i = 1).$$

Further, consider the following indicators, with a convention $0^0 = 1$,

$$\begin{aligned} 1_{\delta, \delta'}(X, \tilde{X}) &:= \prod_{i=0}^n (\mathbf{1}(M_i \leq R))^{\delta_i} (1 - \mathbf{1}(M_i \leq R))^{1-\delta_i} \\ &\times \prod_{i=1}^n \left(\mathbf{1}((X_i, \tilde{X}_i) \in A_R \times A_R) \right)^{\delta'_i} \left(1 - \mathbf{1}((X_i, \tilde{X}_i) \in A_R \times A_R) \right)^{1-\delta'_i} \\ &= 1_{\delta}(X, \tilde{X}) \chi_{\delta'}(X, \tilde{X}) \end{aligned}$$

In fact, the part $\prod_{i=0}^n (\mathbf{1}(M_i \leq R))^{\delta_i} (1 - \mathbf{1}(M_i \leq R))^{1-\delta_i}$ in the latter formula is superficial, however, there is no harm to leave it.

Remind that this indicator function depends on R and n as parameters, which are dropped from the notation. In some cases it will be useful to present the latter indicator as

$$1_{\delta}(X, \tilde{X}) = \prod_{i=0}^n 1_{\delta_i}(M_i), \quad \chi_{\delta'}(X, \tilde{X}) = \prod_{i=1}^{n-1} \chi_{\delta'_i}(X_i, \tilde{X}_i),$$

where for any $M > 0$,

$$\mathbf{1}_{\delta_i}(M) = (\mathbf{1}(\delta_i = 1)\mathbf{1}(M \leq R) + \mathbf{1}(\delta_i = 0)\mathbf{1}(M > R)),$$

and

$$\chi_{\delta'_i}(x) = (\mathbf{1}(\delta'_i = 1)\mathbf{1}(x \in A_R) + \mathbf{1}(\delta'_i = 0)\mathbf{1}(x \notin A_R)).$$

Let us emphasize a very important property of these indicators:

$$1 = \sum_{\delta, \delta' \in \Delta} \mathbf{1}_{\delta, \delta'}(x, \tilde{x}), \quad \forall x, \tilde{x} \in R^{2(n+1)},$$

where Δ the set of all possible values of the vector δ, δ' .

For every δ, δ' let us define

$$J(\delta, \delta') := \{i : 1 \leq i \leq n, \delta_{i-1} = 1, \delta'_i = 1\}, \quad \#1(\delta, \delta') := \sum_{j=1}^n \mathbf{1}(j \in J(\delta, \delta')), \quad (16)$$

$$J(\delta) := \{i : 0 \leq i \leq n, \delta_i = 1\}, \quad \#1(\delta) := \sum_{j=1}^n \mathbf{1}(j \in J(\delta)). \quad (17)$$

Now let us define new operators on the spaces of normalized and non-normalized measures on $R^{2d} = R^d \times R^d$, or, rather, on the space of pairs of measures, each on R^d , as follows, (we use a double integral notation just to emphasize that we integrate with respect to the variables x and \tilde{x})

$$\begin{aligned} & (\mu, \nu) \bar{S}_n^{Y; \mu_0, \nu_0}(A \times B) = \int \int \mathbf{1}(x_n \in A, \tilde{x}_n \in B) \\ & \times \left(\prod_{i=1}^n c_i^{\mu_0} c_i^{\nu_0} \Psi(x_i, Y_i) \Psi(\tilde{x}_i, Y_i) Q(x_{i-1}, dx_i) Q(\tilde{x}_{i-1}, d\tilde{x}_i) \right) \mu(dx_0) \nu(d\tilde{x}_0), \end{aligned}$$

and

$$\begin{aligned} & (\mu, \nu) \bar{S}_n^{Y; R; \delta, \delta'; \mu_0, \nu_0}(A \times B) \\ & = \int \int \mathbf{1}(x_n \in A, \tilde{x}_n \in B) \mathbf{1}_{\delta, \delta'}(x, \tilde{x}) \\ & \times \left(\prod_{i=1}^n c_i^{\mu_0} c_i^{\nu_0} \Psi(x_i, Y_i) \Psi(\tilde{x}_i, Y_i) Q(x_{i-1}, dx_i) Q(\tilde{x}_{i-1}, d\tilde{x}_i) \right) \mu(dx_0) \nu(d\tilde{x}_0), \end{aligned}$$

and

$$\begin{aligned}
& (\mu, \nu) S_n^{Y;R;\delta,\delta'}(A \times B) \\
&= \int \int 1(x_n \in A, \tilde{x}_n \in B) 1_{\delta,\delta'}(x, \tilde{x}) \\
&\times \left(\prod_{i=1}^n \Psi(x_i, Y_i) \Psi(\tilde{x}_i, Y_i) Q(x_{i-1}, dx_i) Q(\tilde{x}_{i-1}, d\tilde{x}_i) \right) \mu(dx_0) \nu(d\tilde{x}_0).
\end{aligned}$$

The last linear non-normalized operator $S_n^{Y;R;\delta,\delta'}$ can be equivalently presented as

$$(\mu, \nu) S_n^{Y;R;\delta,\delta'}(A \times B) = (\mu, \nu) \prod_{i=0}^{n-1} S_{i:i+1}^{Y;R;\delta,\delta'}(A \times B),$$

with

$$\begin{aligned}
& (\mu_i, \nu_i) S_{i:i+1}^{Y;R;\delta,\delta'}(A \times B) \\
&= \int \int 1(x_{i+1} \in A, \tilde{x}_{i+1} \in B) 1_{\delta}(x_i, \tilde{x}_i) \chi_{\delta'}(x_{i+1}, \tilde{x}_{i+1}) \\
&\times \Psi(x_{i+1}, Y_{i+1}) \Psi(\tilde{x}_{i+1}, Y_{i+1}) Q(x_i, dx_{i+1}) Q(\tilde{x}_i, d\tilde{x}_{i+1}) \mu_i(dx_i) \nu_i(d\tilde{x}_i).
\end{aligned}$$

Next, for every δ, δ' , let

$$e_n^{Y;\delta,\delta';\mu_0,\nu_0} := (\mu_0, \nu_0) \bar{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(R^{2d}) \equiv E_{\mu_0,\nu_0}(1_{\delta,\delta'}(Z) \mid Y, \tilde{Y}) \Big|_{\tilde{Y}=Y},$$

where $Z = (X, \tilde{X})$. Due to the assumption on the density q_V , these random variables are well-defined. Notice that the symmetry in the definition of \bar{S} implies an identity crucial for the following calculus,

$$e_n^{Y;\delta,\delta';\mu_0,\nu_0} = e_n^{Y;\delta,\delta';\nu_0,\mu_0}.$$

Next, denote

$$(\mu, \nu) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(A \times B) := (e_n^{Y;\delta,\delta';\mu_0,\nu_0})^{-1} (\mu, \nu) \bar{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(A \times B).$$

The sense of the last notation is that the result of this action is a *normalized* measure restricted to the event $1_{\delta,\delta'}(X, \tilde{X}) = 1$.

Next important step is due to the fact that the distance in total variation for the measures in R^d can be estimated from above via the correspondingly duplicated measures, and the latter can be split into different terms as follows,

$$\begin{aligned}
& \|\mu_0 \bar{S}_n^{Y;\mu_0,\nu_0} - \nu_0 \bar{S}_n^{Y;\nu_0,\mu_0}\|_{TV} \leq \|(\mu_0, \nu_0) \bar{S}_n^{Y;\mu_0,\nu_0} - (\nu_0, \mu_0) \bar{S}_n^{Y;\nu_0,\mu_0}\|_{TV} \\
& \leq \sum_{\delta, \delta' \in \Delta} \|(\mu_0, \nu_0) \bar{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0} - (\nu_0, \mu_0) \bar{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}\|_{TV} \\
& = 2 \sum_{\delta, \delta' \in \Delta} e_n^{Y;\delta;\mu_0,\nu_0} \sup_D ((\mu_0, \nu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(D) - (\nu_0, \mu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(D)),
\end{aligned}$$

where D runs all Borel sets $\mathcal{B}(R^{2d})$ (see [7] for the details). We will use the Birkhoff metric for positive measures, see [9], and also [1], [10] (where it is called Hilbert metric; one more synonym is the projective metric),

$$\rho(\mu, \nu) = \begin{cases} \ln \frac{(\inf s : \mu \leq s\nu)}{(\sup t : \mu \geq t\nu)}, & \text{if finite,} \\ +\infty, & \text{otherwise.} \end{cases} \quad (18)$$

Another equivalent definition reads,

$$\rho(\mu, \nu) = \begin{cases} \ln \sup(d\mu/d\nu) + \ln \sup(d\nu/d\mu), & \text{if finite,} \\ +\infty, & \text{otherwise.} \end{cases}$$

Due to the inequality for the total variation norm and the Birkhoff metric (see [1] and [10]), and since both measures below, – that is, $(\mu_0, \nu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}$ and $(\nu_0, \mu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}$, – are normalized, we have,

$$\begin{aligned}
& 2 \sup_D ((\mu_0, \nu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(D) - (\nu_0, \mu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}(D)) \\
& \leq \rho((\mu_0, \nu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}, (\nu_0, \mu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}). \quad (19)
\end{aligned}$$

We claim that there exists $\pi_R < 1$ such that if $k = \#1(\delta, \delta') > 1$, then

$$\begin{aligned}
& \rho((\mu_0, \nu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}, (\nu_0, \mu_0) \hat{S}_n^{Y;R;\delta,\delta';\mu_0,\nu_0}) \\
& \equiv \rho((\mu_0, \nu_0) S_n^{Y;R;\delta,\delta'}, (\nu_0, \mu_0) S_n^{Y;R;\delta,\delta'}) \leq C \pi_R^{k-1}, \quad (20)
\end{aligned}$$

$k = \#1(\delta, \delta')$ is defined in (16).

This follows by induction from the following two inequalities, see, e.g., [10]; we use here short notations $(\mu_i, \nu_i) = (\mu_0 \nu_0) S_i^{Y;R;\delta,\delta'}$.

(1°) For every i ,

$$\rho\left((\mu_i, \nu_i)S_{i:i+1}^{Y;R;\delta,\delta'}, (\nu_i, \mu_i)S_{i:i+1}^{Y;R;\delta,\delta'}\right) \leq \rho((\mu_i, \nu_i), (\nu_i, \mu_i)).$$

(2°) There exists $\pi_R < 1$ such that if $i + 1 \in J(\delta, \delta')$,

$$\rho\left((\mu_i, \nu_i)S_{i:i+1}^{Y;R;\delta,\delta'}, (\nu_i, \mu_i)S_{i:i+1}^{Y;R;\delta,\delta'}\right) \leq \pi_R \rho((\mu_i, \nu_i), (\nu_i, \mu_i)).$$

The latter follows from the Proposition 3.9 from [10], with the contraction constant, $\pi_R \leq (1 - \tilde{C}_R^{-2})/(1 + \tilde{C}_R^{-2})$, due to the “mixing condition”

$$\tilde{C}_R =: \sup_{D_R} \frac{Q_{i:i+1}(x_0, \tilde{x}_0, dx', d\tilde{x}')}{Q_{i:i+1}(v_0, \tilde{v}_0, dx', d\tilde{x}')} \equiv \sup_{D_R} \frac{Q(x_0, \tilde{x}_0, dx', d\tilde{x}')}{Q(v_0, \tilde{v}_0, dx', d\tilde{x}')} < \infty, \quad (21)$$

with

$$D_R := \{(x_0, \tilde{x}_0, v_0, \tilde{v}_0, x', \tilde{x}') : |x_0|, |\tilde{x}_0|, |v_0|, |\tilde{v}_0| \leq R, x', \tilde{x}' \in A_R \times A_R\}.$$

Then, the meaning of the inequality (2°) is that the replacement of non-random kernels Q by random ones $Q\Psi$ does not change the supremum of the derivative of one measure with respect to another.

For the completeness, the proof of the inequality (1°) can be found in [7].

The induction base $k = 1$ (not $k = 0$) in (20) is valid due to the fact that after the first pair of ones, the measures become comparable, by virtue of (21); and the induction step follows from (2°) directly.

Now we can estimate as follows :

$$\begin{aligned} & E_{\mu_0, \nu_0} \|\mu_0 \bar{S}_n^{Y, \mu_0} - \nu_0 \bar{S}_n^{Y, \nu_0}\|_{TV} \\ & \leq \sum_{\delta, \delta' \in \Delta, \#1(\delta, \delta') \geq 1} C_R \pi_R^{\#1(\delta, \delta') - 1} E_{\mu_0, \nu_0} e_n^{Y; \delta, \delta'; \mu_0, \nu_0} + 2 \sum_{\delta, \delta' \in \Delta; \#1(\delta, \delta') = 0} E e_n^{Y; \delta, \delta'; \mu_0, \nu_0} \\ & \leq \sum_{\delta, \delta' \in \Delta} C_R (\pi_R^{\#1(\delta, \delta') - 1} \wedge 1) E_{\mu_0, \nu_0} e_n^{Y; \delta, \delta'; \mu_0, \nu_0}. \end{aligned} \quad (22)$$

2. Let us split the sum $\sum_{\delta, \delta' \in \Delta}$ into three parts: (1) with $\#1(\delta, \delta') \geq \epsilon' n$; (2) with $\#1(\delta, \delta') < \epsilon' n$ & $\#1(\delta) \leq \epsilon n$; and (3) with $\#1(\delta, \delta') < \epsilon' n$ & $\#1(\delta) > \epsilon n$,

where $\epsilon' > 0$ and another small $\epsilon > 0$ such that $\epsilon' \ll \epsilon$ are to be chosen. Correspondingly, we will estimate the three sums,

$$S^1 = \sum_{\delta: \#1(\delta, \delta') \geq \epsilon' n}, \quad S^2 = \sum_{\delta: \#1(\delta) \leq \epsilon n}, \quad \text{and} \quad S^3 = \sum_{\delta: \#1(\delta) > \epsilon n, \#1(\delta, \delta') < \epsilon' n}.$$

For the first one we have contraction in the Birkhoff metric; the second sum is small whatever $\epsilon < 1$ for R large enough, due to recurrence; the third one turns out to be exponentially small if ϵ' is small enough, due to the assumption (A3). Notice that the third sum is a new term in compare to [7]; the first two are similar to the terms in [7], however, due to a new object δ' we have to repeat partially the calculus.

3. We have, for any $0 < \epsilon' < 1$,

$$\begin{aligned} S^1 &= \sum_{\delta, \delta': \#1(\delta, \delta') \geq \epsilon' n} (\pi_R^{\#1(\delta, \delta')-1} \wedge 1) E_{\mu_0} e_n^{Y; \delta; \delta' \mu_0, \nu_0} \\ &= \sum_{\delta: \#1(\delta, \delta') \geq \epsilon' n} (\pi_R^{\#1(\delta, \delta')-1} \wedge 1) E_{\mu_0} E_{\mu_0, \nu_0} (1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y}) \Big|_{\tilde{Y}=Y} \\ &\leq \pi_R^{\epsilon' n - 1} \sum_{\delta, \delta': \#1(\delta, \delta') \geq \epsilon' n} E_{\mu_0} E_{\mu_0, \nu_0} (1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y}) \Big|_{\tilde{Y}=Y} \\ &= \pi_R^{\epsilon' n} E_{\mu_0} \sum_{\delta, \delta': \#1(\delta, \delta') \geq \epsilon' n} P_{\mu_0, \nu_0} (1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y}) \Big|_{\tilde{Y}=Y} \\ &= \pi_R^{\epsilon' n} E_{\mu_0} P_{\mu_0, \nu_0} \left(\bigcup_{\delta': \#1(\delta, \delta') \geq \epsilon' n} 1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \leq \pi_R^{\epsilon' n}. \end{aligned} \quad (23)$$

4. To estimate the second term of the sum we apply the hint proposed in [7] which gives the following bound:

$$\begin{aligned} S^2 &= \sum_{\delta, \delta': \#1(\delta, \delta') < \epsilon' n \ \& \ \#1(\delta) < \epsilon n} (\pi_R^{\#1(\delta, \delta')-1} \wedge 1) E_{\mu_0} \left(E_{\mu_0, \nu_0} (1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y}) \Big|_{\tilde{Y}=Y} \right) \\ &\leq \sum_{\delta, \delta': \#1(\delta) < \epsilon n} E_{\mu_0} \left(E_{\mu_0, \nu_0} (1_{\delta, \delta'}(X, \tilde{X}) | Y, \tilde{Y}) \Big|_{\tilde{Y}=Y} \right) \\ &\leq E_{\mu_0} \left(E_{\mu_0, \nu_0} \left(1(\sum_{k=0}^n \mathbf{1}(M_k \leq R) < \epsilon n) | Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \right) \\ &\leq E_{\mu_0} \left(E_{\mu_0, \nu_0} \left(1(\sum_{k=0}^n \mathbf{1}(|X_k| \leq R) < \frac{1+\epsilon}{2} n) | Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \right) \\ &+ E_{\mu_0} \left(E_{\mu_0, \nu_0} \left(1(\sum_{k=0}^n \mathbf{1}(|\tilde{X}_k| \leq R) < \frac{1+\epsilon}{2} n) | Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \right) \end{aligned}$$

$$\begin{aligned}
&\leq E_{\mu_0} \left(\mathbf{1}(\sum_{k=0}^n \mathbf{1}(|X_k| \leq R) < \frac{1+\epsilon}{2}n) \right) \\
&+ CE_{\nu_0} \left(\mathbf{1}(\sum_{k=0}^n \mathbf{1}(|\tilde{X}_k| \leq R) < \frac{1+\epsilon}{2}n) \right). \tag{24}
\end{aligned}$$

Similarly to the calculus in [7], due to the bounds from [11] and [12], for every $0 < \epsilon < 1$, the latter expectation admits an appropriate bound, exponential or polynomial, depending on the value p , if R is chosen large enough, namely,

$$E_{\mu_0} \left(E_{\mu_0, \nu_0} \left(\sum_{\delta: \#1(\delta) < \epsilon n} \mathbf{1}_{\delta}(X, \tilde{X}) \mid Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \right) \leq \begin{cases} C_m n^{-m}, & p = 1, \\ C \exp(-cn), & p = 0, \end{cases}$$

for all $m > 0$ in the case $p = 1$.

5. The last step is to estimate the third term in the sum. This estimation will be based on the following inequalities:

$$\begin{aligned}
&\sum_{\delta, \delta': \#1(\delta, \delta') < \epsilon' n \ \& \ \#1(\delta) > \epsilon n} E_{\mu_0} \left(E_{\mu_0, \nu_0} \left(\mathbf{1}_{\delta, \delta'}(X, \tilde{X}) \mid Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \right) \\
&= E_{\mu_0} \left(E_{\mu_0, \nu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(M_{i-1} \leq R) > \epsilon n, \right. \right. \\
&\quad \left. \left. \sum_{i=1}^n \mathbf{1}(M_{i-1} \leq R \ \& \ (X_i, \tilde{X}_i) \in A_R \times A_R) < \epsilon' n \right) \mid Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \\
&\leq E_{\mu_0} \left(E_{\mu_0, \nu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R) > \epsilon n, \right. \right. \\
&\quad \left. \left. \sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R \ \& \ X_i \in A_R) < \frac{1+\epsilon'}{2}n \right) \mid Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \\
&\quad + \left(E_{\mu_0, \nu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|\tilde{X}_{i-1}| \leq R) > \epsilon n, \right. \right. \\
&\quad \left. \left. \sum_{i=1}^n \mathbf{1}(|\tilde{X}_{i-1}| \leq R \ \& \ \tilde{X}_i \in A_R) < \frac{1+\epsilon'}{2}n \right) \mid Y, \tilde{Y} \right) \Big|_{\tilde{Y}=Y} \\
&\leq E_{\mu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R) > \epsilon n, \sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R \ \& \ X_i \in A_R) < \frac{1+\epsilon'}{2}n \right) \\
&+ CE_{\nu_0} \left(\mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|\tilde{X}_{i-1}| \leq R) > \epsilon n, \sum_{i=1}^n \mathbf{1}(|\tilde{X}_{i-1}| \leq R \ \& \ \tilde{X}_i \in A_R) < \frac{1+\epsilon'}{2}n \right) \right) \\
&\leq E_{\mu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \geq \epsilon'' n \right) \\
&+ CE_{\nu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|\tilde{X}_{i-1}| \leq R, \tilde{X}_i \notin A_R) \geq \epsilon'' n \right),
\end{aligned}$$

where $\epsilon'' = \epsilon - \frac{1+\epsilon'}{2}$. Two last terms can be estimated using the same calculations, so we deal with only the first one. Remind that so far both ϵ and ϵ' are arbitrary values from $(0, 1)$, – with the only restriction $\epsilon' < \epsilon$, – and it is time to choose them now. Clearly, the more ϵ'' , the less the indicator function $\mathbf{1}(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \geq \epsilon''n)$. Hence, we also minimize the expectation of this indicator if we choose ϵ'' as large as possible; and the supremum of possible is $1/2$, although, of course, it is not attained, for $\epsilon < 1$, and $\epsilon' > 0$. In the other words, we should choose ϵ close to 1, and ϵ' close to zero, then ϵ'' will be just slightly less than $1/2$. What is important here is that by this choice we can assure the inequality

$$p_R + \epsilon'' > 1. \quad (25)$$

Now, by exponential Bienaimé–Chebyshev, we estimate, with any $\lambda > 0$,

$$\begin{aligned} & E_{\mu_0} \mathbf{1} \left(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \geq \epsilon''n \right) \\ & \leq \exp(-\lambda\epsilon''n) E_{\mu_0} \exp \left(\lambda \sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \right) \\ & = \exp(-\lambda\epsilon''n) E_{\mu_0} \left(\exp \left(\lambda \sum_{i=1}^{n-1} \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \right) \right. \\ & \quad \left. \times E_{X_{n-1}} \exp(\lambda \mathbf{1}(|X_{n-1}| \leq R, X_n \notin A_R)) \right). \end{aligned}$$

Here we have,

$$\begin{aligned} & E_{X_{n-1}} \exp(\lambda \mathbf{1}(|X_{n-1}| \leq R, X_n \notin A_R)) \\ & = \mathbf{1}(X_{n-1} \in B_R) E_{X_{n-1}} \exp(\lambda \mathbf{1}(|X_{n-1}| \leq R, X_n \notin A_R)) \\ & \quad + \mathbf{1}(X_{n-1} \notin B_R) E_{X_{n-1}} \exp(\lambda \mathbf{1}(|X_{n-1}| \leq R, X_n \notin A_R)) \\ & \leq \mathbf{1}(X_{n-1} \in B_R) \sup_{x \in B_R} E_x \exp(\lambda \mathbf{1}(X_1 \notin A_R)) + \mathbf{1}(X_{n-1} \notin B_R) \\ & \quad \leq \sup_{x \in B_R} E_x \exp(\lambda \mathbf{1}(X_1 \notin A_R)) \\ & = \sup_{x \in B_R} (P_x(X_1 \in A_R) + e^\lambda(1 - P_x(X_1 \in A_R))) \end{aligned}$$

$$= (p_R + e^\lambda(1 - p_R)).$$

By induction,

$$\begin{aligned} & \exp(-\lambda\epsilon''n) E_{\mu_0} \exp\left(\lambda \sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R)\right) \\ & \leq \exp(-\lambda\epsilon''n) (p_R + e^\lambda(1 - p_R))^n \equiv (\exp(-\lambda\epsilon'')(p_R + e^\lambda(1 - p_R)))^n. \end{aligned}$$

Now,

$$(\exp(-\lambda\epsilon'')(p_R + e^\lambda(1 - p_R)))|_{\lambda=0} = 1,$$

and, by virtue of (25),

$$(\exp(-\lambda\epsilon'')(p_R + e^\lambda(1 - p_R)))'_\lambda|_{\lambda=0} = -\epsilon'' + 1 - p_R < 0,$$

Therefore, for λ from some right neighbourhood of zero, we obtain

$$q := (\exp(-\lambda\epsilon'')(p_R + e^\lambda(1 - p_R))) < 1.$$

Let us fix any such $\lambda > 0$. Then, with this λ ,

$$E_{\nu_0} \mathbf{1}\left(\sum_{i=1}^n \mathbf{1}(|X_{i-1}| \leq R, X_i \notin A_R) \geq \epsilon''n\right) \leq q^n,$$

which provides a desired exponential bound for S^3 .

6. The non-averaged bounds follow from Chebyshev's inequality and the Borel–Cantelli lemmata. The Theorem 1 is proved.

Remark 3 *Notice that the inequality with $1/2$ never appears in non-conditional mixing, so it looks a bit mysterious in the assumption (A3). Intuitively, it relates to the “uncoupling” of X and \tilde{X} : after that procedure we still wish to see each component frequently enough in the ball B_R along with its next state, leaving alone that this next state actually should be in A_R . All this requires $> 1/2$ in (5). Possibly, this is due to the method, and could be relaxed by some more clever procedure.*

Acknowledgements

The second author thanks the grant RFBR 05-01-00449 for support.

References

- [1] Atar, R., Zeitouni, O. Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.* 33 (1997), 6, 697–725.
- [2] Baxendale, P., Chigansky, P., Liptser, R. Asymptotic stability of the Wonham filter: ergodic and nonergodic signals. *SIAM J. Control Optim.* 43 (2004), 2, 643–669 , DOI. 10.1137/S0363012902416929.
- [3] Dobrushin, R.L. Central limit theorem for nonstationary Markov chains, I, II, *Theory of probability and its applications*, 1 (1956), 66-80 and 330-385.
- [4] Doob, J. L. *Stochastic Processes*, Wiley, NY, 1953.
- [5] Ibragimov, I. A., Linnik, Yu. V. *Independent and stationary sequences of random variables*. Wolters-Noordhoff Publ., Groningen, 1971.
- [6] Gulinsky, O. V., Veretennikov, A. Yu. *Large deviations for discrete-time processes with averaging*. VSP, Utrecht, 1993.
- [7] Kleptsyna, M. L., Veretennikov, A. Yu. On discrete time ergodic filters with wrong initial data. *Probability Theory and Related Fields*, 2007, to appear; available online at DOI: 10.1007/s00440-007-0089-7.
- [8] Kleptsyna, M. L., Veretennikov, A. Yu. On ergodic filters with wrong initial data, *C.R.Acad. Sci. Paris, Ser. I*, 344(2007), 727-731.
- [9] Krasnosel'skii, M. A., Lifshits, E. A., Sobolev, A. V. *Positive linear systems. The method of positive operators*. Heldermann Verlag, Berlin, 1989.
- [10] Le Gland, F., Oudjane, N. Stability and uniform approximation of nonlinear filters using the Hilbert metric and application to particle filters, *The Annals of Applied Probability*, 14, (2004), 1, 144-187.
- [11] Veretennikov, A. Yu. Estimates of the mixing rate for stochastic equations. (Russian) *Teor. Veroyatnost. Primenen.* 32 (1987), no. 2, 299–308; Engl. transl.: *Theory Probab. Appl.* 32 (1987), no. 2, 273–281.
- [12] Veretennikov, A. Yu. On polynomial mixing and the rate of convergence for stochastic differential and difference equations. *Teor. Veroyatn. Primenen.* 44 (1999), 2, 312–327; Engl. transl.: *Theory Probab. Appl.* 44 (2000), 2, 361–374.
- [13] Veretennikov, A. Yu. On approximations of diffusions with equilibrium, Research Report / Helsinki University of Technology. Institute of Mathematics. Report C017, 2004. <http://math.tkk.fi/visitors0405/AVslides.pdf>