Asymptotic filtering for finite state Markov chains

Rafail Khasminskii *
Department of Mathematics
Wayne State University
Detroit, MI 48202
U.S.A.

Ofer Zeitouni †
Department of Electrical Engineering
Technion- Israel Institute of Technology
Haifa 32000, Israel

September 6, 1995. Revised March 6, 1996.

Abstract Asymptotic formulae for the optimal filtering error for finite state space Markov chains observed in independent noise are presented. Asymptotically optimal *simple* filters, which do not depend on the transition rates of the chain, are also presented.

AMS SUBJECT CLASSIFICATION: Primary 93E11, Secondary 62M05,62M02.
KEYWORDS: Nonlinear filtering, hypothesis testing, Markov chains.

1 Introduction and statement of results

Let X_n^ϵ (the "state process") denote a discrete time Markov chain, with state space $S = \{1, \ldots, d\}$, initial distribution p_0^ϵ, and transition probability matrix

$$\Pi = \{\pi_{ij}^\epsilon\} = P(X_{n+1}^\epsilon = j \mid X_n^\epsilon = i) = \begin{cases} \epsilon\lambda_{ij} & , \quad i \neq j \\ 1 - \epsilon\lambda_{ii} & , \quad i = j \end{cases} \tag{1}$$

where $\lambda_{ii} = \sum_{j \neq i} \lambda_{ij}$.

We assume throughout that the Markov chain generated by Π is irreducible (and aperiodic) and denote its stationary distribution (which is independent of ϵ!) by $p_s = \{p_s(i)\}_{i=1}^d$. Let $\{y_n^\epsilon\}_{n=1}^\infty$, (the "observation") denote a sequence of random variables, such that, given the sequence $\{X_n^\epsilon\}_{n=1}^\infty$,

*The work of this author was partially supported by ONR Grant N00014-95-1-0793.
†The work of this author was partially supported by the Israel Science Foundation administered by the Israel Academy of Sciences and Humanities, and by the fund for promotion of research at the Technion.

the random variables y_i^ϵ are independent, with

$$P(y_n^\epsilon \in dx \,|\, \{X_i^\epsilon\}_{i=1}^\infty) = P(y_n^\epsilon \in dx \,|\, X_n^\epsilon)$$

and $P(y_n^\epsilon \in dx \,|\, X_n^\epsilon = i) = \mu_i(dx)$.

Let now $\mathcal{F}_{y,n} = \sigma\{y_i^\epsilon,\, 1 \le i \le n\}$. The *filtering problem* consists of finding the best (in an appropriate sense, see below) estimator for X_n^ϵ given the information $\mathcal{F}_{y,n}$. In the setting considered here, it is natural to consider "best" in the sense of minimum probability of error. That is, for any S-valued estimator $\tilde{X}_n$ which is $\mathcal{F}_{y,n}$ measurable let

$$P_e^{\epsilon,n} = E(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n})\,.$$

(Here and throughout, E denotes expectation with respect to the Markov measure generated by (p_0^ϵ, Π) and with respect to the filter $\tilde{X}_n$).

Let $\hat{P}_e^{\epsilon,n}$ denote the infinum of $P_e^{\epsilon,n}$ over the class of all possible $\mathcal{F}_{y,n}$-measurable estimators. In this note, we consider the asymptotics of $\hat{P}_e^{\epsilon,n}$. Throughout, $E_i g = \int g(x)\mu_i(dx)$, and a function $g(\epsilon,n)$ is denoted $o(k(\epsilon))$ if $\lim_{\epsilon\to 0}\lim_{n\to\infty} g(\epsilon,n)/k(\epsilon) = 0$.

To state our results, we make the following assumption.

(A) For any i,j, the measures μ_i, μ_j are mutually absolutely continuous. Furthermore, the relative entropies

$$a_{ij} = \int \mu_i(dx) \log\left(\frac{d\mu_i}{d\mu_j}(x)\right)$$

exist, with $\min\limits_{i\neq j} a_{ij} > 0$. Finally, $E_i \left|\log \frac{d\mu_i}{d\mu_j}\right|^\alpha < \infty$ for some $\alpha > 2$.

Our main result is the

Theorem 1. *Assume (A). Then, for any initial distribution p_0^ϵ,*

$$\lim_{n\to\infty} \hat{P}_e^{\epsilon,n} = \left(\sum_i p_s(i) \sum_{j\neq i} \frac{\lambda_{ij}}{a_{ji}}\right) \epsilon \log\left(\frac{1}{\epsilon}\right)(1+o(1))\,, \quad \epsilon \to 0\,.$$

A similar result holds also for (the simpler) continuous time model. That is, let X_t^ϵ denote a continuous time Markov chain with state space $S = \{1, \ldots, d\}$, initial law p_0^ϵ, and jump rates $\epsilon\lambda_{ij}$, and let $dy_t^\epsilon = h(X_t^\epsilon)dt + d\nu_t$, where ν_t is a Brownian motion independent of the process $\{X_t^\epsilon\}$. Let $\mathcal{F}_{y,t} = \sigma\{y_s^\epsilon, 0 \le s \le t\}$, and define $\hat{P}_e^{\epsilon,t}$ as in the discrete time case. Finally, modify (A) to (A'):

(A') let $h(i) \neq h(j)$ for $i \neq j$, and define

$$a_{ij} = \frac{(h(i) - h(j))^2}{2}\,.$$

Theorem 1'. *Assume (A'). Then, for any initial law p_0^ϵ,*

$$\lim_{t\to\infty} \hat{P}_e^{\epsilon,t} = \left(\sum_i p_s(i) \sum_{j\neq i} \frac{\lambda_{ij}}{a_{ji}}\right) \epsilon \log\left(\frac{1}{\epsilon}\right)(1 + o(1))\,, \quad \epsilon \to 0\,.$$

Theorem 1' was derived by Wonham (1965) in the case $d = 2$ with symmetric transition rates by considering the exact optimal filter. While the optimal filters are known, both in discrete and continuous time, for arbitrary $d < \infty$, their structure is rather complicated, and it is not clear how to use this structure to derive theorems 1 and 1'. In Khasminskii and Lazareva (1992), the general case $d = 2$ in continuous time is handled, by analyzing a sub-optimal filter to derive an upper bound on the filtering error. Our approach here is somewhat different, using information-theoretic arguments for a lower bound and (different) sub-optimal filters for the upper bound.

Since the proof of Theorem 1' is similar to that of Theorem 1, we concentrate in the sequel on the latter. Theorem 1' then follows by replacing throughout sums with integrals, and using the explicit form of the likelihood in the continuous time setup.

2 Proofs

Proof of Theorem 1

As mentioned above, the proof of the lower bound uses information theoretic arguments, whereas the upper bound consists of exhibiting a suboptimal (asymptotically optimal) filter. In both the upper and lower bounds, the case $d = 2$ offers simplifications, and it is useful at first reading to consider it. Thus, we have structured our proofs in such a way that the basic idea is first illustrated in this simple case.

We begin by deriving the lower bound. The key to the proof consists of the analysis of an auxiliary hypothesis testing problem. Let the null hypothesis, H_0, consist of $\{y_1, \ldots, y_T\}$ being i.i.d. random variables of law μ_0, and let H_1 consist of $\{y_1, \ldots, y_{\tau-1}\}$ being i.i.d. random variables of law μ_0, $\{y_\tau, \ldots, y_T\}$ i.i.d. random variables of law μ_1, independent of $\{y_1, \ldots, y_{\tau-1}\}$, where τ is a random variable, independent of $\{y_1, \ldots, y_{\tau-1}\}$ and $\{y_\tau, \ldots, y_T\}$, uniformly distributed in $\{1, \ldots, T\}$. We assume that the prior probabilities of H_0, H_1, denoted P_{ap}, satisfy $P_{\mathrm{ap}}(H_0) = (1 - \lambda\epsilon T + o(\epsilon))$, $P_{\mathrm{ap}}(H_1) = \lambda\epsilon T + o(\epsilon)$. Denote by $HT(\epsilon, T, \mu_0, \mu_1, \lambda)$ the *optimal* probability of error in this auxiliary hypothesis testing problem.

Lemma 1. *Let* $a_{10} = \int d\mu_1 \log \frac{d\mu_1}{d\mu_0}$*, and assume* $E_{\mu_1} \left|\log\left(\frac{d\mu_1}{d\mu_0}\right)\right|^\alpha < \infty$ *for some* $\alpha > 2$*. Then for any* $\delta > 0$ *and* $T = T(\epsilon) \underset{\epsilon \to 0}{\longrightarrow} \infty$ *such that* $-\frac{\log \epsilon}{T} \geq \delta + a_{10}$*,*

$$HT(\epsilon, T, \mu_o, \mu_1, \lambda) \geq \epsilon\lambda T(1 + o(1))\,.$$

Proof of Lemma 1. The optimal test is the likelihood test (c.f. Lehmann (1986)), that is the optimal test forms the functional

$$\Lambda = \frac{p(y_1, \ldots, y_T \,|\, H_1)}{p(y_1, \ldots, y_T \,|\, H_0)}$$

and decides H_1 if $\epsilon\lambda T\Lambda(1 + o(1)) \geq 1$, H_0 otherwise.

Note that, by definition,

$$\Lambda = \frac{1}{T}\sum_{j=1}^{T} e^{\sum_{i=j}^{T} \log f(y_i)}\,,$$

where $f(X) = \frac{d\mu_1}{d\mu_0}(X)$. Hence,

$$\begin{aligned} P_{\mathrm{error}} &\geq P_{\mathrm{ap}}(H_1) P(\lambda\epsilon T\Lambda(1 + o(1)) < 1 \,|\, H_1) \\ &= \lambda\epsilon T(1 + o(1)) \left(\frac{1}{T}\sum_{k=1}^{T} P(\lambda\epsilon T\Lambda < (1 + o(1)) \,|\, H_1, \tau = k)\right)\,. \end{aligned} \tag{2}$$

But,

$$P(\lambda\epsilon T\Lambda < (1+o(1))\,|\,H_1,\tau=k) \;=\; P\left(\sum_{j=1}^{T} e^{\sum_{i=j}^{T}\log f(y_i)} < \frac{1+o(1)}{\lambda\epsilon}\,|\,H_1,\tau=k\right)$$
$$\geq\; 1-\sum_{j=1}^{T} P\left(e^{\sum_{i=j}^{T}\log f(y_i)} \geq \frac{1+o(1)}{\lambda\epsilon T}\,|\,H_1,\tau=k\right). \quad (3)$$

Note that for $k \leq j$,

$$P(e^{\sum_{i=j}^{T}\log f(y_i)} \geq \frac{1}{\lambda\epsilon T^3}|\,H_1,\tau=k)$$
$$=\mu_1^{\otimes(T-j+1)}\left(\sum_{i=j}^{T}\log f(y_i) \geq -\log\epsilon - \log(\lambda T^3)\right)$$
$$=\mu_1^{\otimes(T-j+1)}\left(\sum_{i=j}^{T}(\log f(y_i)-a_{10}) \geq -\log\epsilon - \log(\lambda T^3) - (T-j+1)a_{10}\right)$$
$$\leq \frac{CT^{\alpha/2}}{(-\log\epsilon - \log(\lambda T^3) - Ta_{10})^{\alpha}} < C_\delta T^{-(\alpha/2)} \quad (4)$$

for some $C, C_\delta > 0$, where we have used Chebycheff's inequality (for the function X^α), in the last step, that is, we used the fact that if $X_i = \log f(y_i) - a_{10}$ are i.i.d. of zero mean and finite α-s moment then, by a successive application of Chebycheff's inequality and the inequalities of Marchinkiewicz and Zygmund and Minkowski (see, e.g., Shiryayev (1984), pp. 469, 192), one concludes that for deterministic M_T,

$$P(\sum_{i=1}^{T} X_i \geq M_T) \leq \frac{E|\sum_{i=1}^{T} X_i|^\alpha}{M_T^\alpha} \leq \frac{c_\alpha T^{\alpha/2}E(|X_1|^\alpha)}{M_T^\alpha},$$

where c_α does not depend on T, M_T.

In particular, one deduces from (4) that, for $k \leq j$,

$$\sum_{j=k}^{T} P\left(e^{\sum_{i=j}^{T}\log f(y_i)} \geq \frac{1+o(1)}{\lambda\epsilon T}\,|\,H_1,\tau=k\right) \underset{T\to\infty}{\longrightarrow} 0\,. \quad (5)$$

On the other hand, for $k > j$,

$$P\left(e^{\sum_{i=j}^{T}\log f(y_i)} \geq \frac{1+o(1)}{\lambda\epsilon T}\,|\,H_1,\tau=k\right)$$

$$
\begin{aligned}
&\leq \mu_0^{\otimes(k-j)}\left(e^{\sum_{i=j}^{k-1}\log f(y_i)} \geq T^2 \,|\, H_1, \tau = k\right) + \mu_1^{\otimes(T-k+1)}\left(e^{\sum_{i=k}^{T}\log f(y_i)} \geq \frac{1}{\lambda\epsilon T^3} \,|\, H_1, \tau = k\right) \\
&\leq T^{-2} + C_\delta T^{-(\alpha/2)}, \qquad (6)
\end{aligned}
$$

where we have used the fact that $E_{\mu_0}(e^{\log f(y_i)}) = 1$ and Chebycheff's inequality, coupled with (4), to derive the last inequality. Combining (3), (5) and (6), we arrive at

$$
P_{\text{error}} \geq \lambda\epsilon T(1 + o(1))\,.
$$

□

We return to the proof of the lower bound. Our technique consists roughly of extending the conditioning σ-field $\mathcal{F}_{y,n}$ with the value of the state before the last jump, and then applying Lemma 1. Let $\delta > 0$ small enough be given, let $T_{ij} = \left(\frac{1}{a_{ji}} - \delta\right)\log\left(\frac{1}{\epsilon}\right) > 0$, define $\overline{T}_i = \max_{j\neq i} T_{ij}$ and $\overline{T} = \max_i \overline{T}_i$. Note next that X_n^ϵ may be constructed using independent Bernoulli random variables $N_n(i)$ of parameter $\epsilon\lambda_{ii}$, independent random variables $I_{n,i}$ with $P(I_{n,i} = j) = \lambda_{ij}/\lambda_{ii}$, and setting $X_{n+1}^\epsilon = \mathbf{1}_{N_n(X_n^\epsilon)=0} X_n^\epsilon + \mathbf{1}_{N_n(X_n^\epsilon)=1} I_{n,X_n^\epsilon}$. For any $t > 0$, we denote by I_t the value of I_{n,X_n^ϵ} *at the first jump of* X_n^ϵ *after* t. Let $\overline{I} = I_{n-\overline{T}}$. Let $\sigma(i)$ denote a bijection of $\{1,\ldots,d-1\} \to \{1,\ldots,d\}\setminus\overline{I}$ such that

$$
T_{\sigma(i),\overline{I}} \geq T_{\sigma(i+1),\overline{I}}, \qquad i = 1,\ldots,d-1
$$

and let $\overline{\imath} = \sigma(\min\{i : X^\epsilon_{n-T_{\sigma(i),\overline{I}}} = (\sigma(i) \text{ or } \overline{I})\})$, with $\overline{\imath} = 0$ if none of the equalities hold.

Note that

$$
P(\overline{\imath} = 0) \leq P(\text{two jumps or more occurred in } [n - \overline{T}, n]) = O(\epsilon^2 \log^2 \epsilon)\,,
$$

while for $\bar{i} \neq \bar{I}$, $\bar{i} \neq 0$, the pair $(\bar{i}, \bar{I})$ denotes the two states between which the first jump after $n - \bar{T}$ occurs. Finally, let $\overline{Y}_n = X^\epsilon_{n-T_{\bar{\imath},\overline{I}}}$ if $\bar{i} \neq \bar{I}$ and $\bar{Y}_n = X_n^\epsilon$ otherwise, and let $\tilde{X}_n^\epsilon$ denote the optimal filter given the information $\{\overline{Y}_n, \overline{\imath}, \overline{I}, \mathcal{F}_{y,n}\}$. Clearly, with

$$
OJ = \{\text{at most one jump occurred in } [n - \overline{T}, n]\}\,,
$$

$$
\hat{P}_e^{\epsilon,n} \quad \geq \quad E\left(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n^\epsilon}\right)
$$

$$\geq \sum_{i,j=1,i\neq j}^{d} P(\overline{Y}_n = i, \overline{I} = j\,, OJ) E\left(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n^\epsilon} | \overline{Y}_n = i, \overline{I} = j,\, OJ\right)$$

$$\geq \sum_{i,j=1,i\neq j}^{d} (1+o(1)) P(\bar{Y}_n = i, I_n = j) E\left(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n^\epsilon} | \overline{Y}_n = i, \overline{I} = j,\, OJ\right) - O(\epsilon^2 \log^2 \epsilon)\,.$$

Let $\tilde{X}_n^{\epsilon,i,j}$ denote the optimal filter given by $\{\overline{Y}_n = i, \overline{I} = j, \bar{\imath}, OJ, \mathcal{F}_{y,n}\}$. Then, by the Markov structure of the pair $(X_n^\epsilon, y_n^\epsilon)$,

$$\begin{aligned} E\left(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n^\epsilon} | \overline{Y}_n = i, \overline{I} = j, OJ\right) &\geq E\left(\mathbf{1}_{X_n^\epsilon \neq \tilde{X}_n^{\epsilon,i,j}} | \overline{Y}_n = i, \overline{I} = j, OJ\right) \\ &= HT(\epsilon, T_{ij}, \mu_i, \mu_j, \lambda_{ii})\,. \end{aligned}$$

(notations as in Lemma 1). Hence,

$$\begin{aligned} \hat{P}_e^{\epsilon,n} &\geq (1+o(1)) \sum_{i,j=1,i\neq j}^{d} p_s(i) \frac{\lambda_{ij}}{\lambda_{ii}} \cdot HT(\epsilon, T_{ij}, \mu_i, \mu_j, \lambda_{ii}) - O(\epsilon^2 \log^2 \epsilon) \\ &= (1+o(1)) \sum_{i,j=1,i\neq j}^{d} p_s(i) \frac{\lambda_{ij}}{\lambda_{ii}} \cdot \lambda_{ii} \left(\frac{1}{a_{ji}} - \delta\right) \epsilon \log \frac{1}{\epsilon}\,. \end{aligned}$$

Since δ is arbitrary, the required lower bound follows.

We next turn to the proof of the upper bound. We do this by proposing a sub-optimal filter, denoted $\tilde{X}_n^\epsilon$. Since the case $d = 2$ with $a_{12} = a_{21} = a$ is particularly transparent, we begin by giving a quick proof of the theorem in that case, which illustrates our basic approach. For $\eta \in \mathbb{R}$, define $\Lambda(\eta) = \log E_{\mu_1}((d\mu_2/d\mu_1)^\eta)$, and $0 < \Lambda^* = \sup_{\eta\in\mathbb{R}} -\Lambda(\eta) < \infty$ (see Dembo and Zeitouni (1993), Chapter 2.2, for useful facts concerning Λ^*.) In particular, let $\delta > 0$ be given, define $\bar{T} = (a^{-1} + \delta)\log(1/\epsilon)$ and for $T = \max((1+\delta)\log(1/\epsilon)/\Lambda^*, \bar{T})$, let $Z_n = \sum_{j=n-T+1}^{n} \log \frac{d\mu_1}{d\mu_2}(y_j^\epsilon)$. Then from Dembo and Zeitouni (1993), Theorem 3.4.3, one knows that for ϵ small enough it holds that

$$\mu_1^{\otimes T}(Z_n < 0) \leq e^{-T\Lambda^*(1+o(1))} \leq \epsilon^{1+3\delta/4}\,. \tag{7}$$

Next, let $\bar{Z}_n = \sum_{j=n-\bar{T}+1}^{n} \log \frac{d\mu_1}{d\mu_2}(y_j^\epsilon)$. Let $I_n = 1$ if $Z_n > 0$ and $I_n = 2$ otherwise. I_n is the first, "coarse", stage of our suboptimal filter. Indeed, (7) ensures that if no jump has occured in $(n - T + 1, n)$, I_n is a good estimate of X_n^ϵ. However, the probability of a jump in this interval is too large, and thus a good filter must refine the information given by I_n. To this end, the proposed suboptimal filter is taken as $\tilde{X}_n^\epsilon = 1$ if $I_n = 1, \bar{Z}_n > (1+a\delta/2)\log(\epsilon)$ or $I_n = 2, \bar{Z}_n > -(1+a\delta/2)\log(\epsilon)$, and $\tilde{X}_n^\epsilon = 2$ otherwise. To evaluate the performance, note that, for some constants $C > 0$ independent of ϵ,

$$
\begin{aligned}
P(\text{error} \,|\, X_n^\epsilon = 1) \;\le\;& P(\text{two jumps or more occurred in } [n-T+1, n] \;|X_n^\epsilon = 1) \\
&+P(\text{one jump occurred in } [n-\bar{T}+1, n] \;|X_n^\epsilon = 1) \\
&+P(\tilde{X}_n^\epsilon = 2 | X_i^\epsilon = 1, i \in [n-T+1, n]) \\
&+P(\tilde{X}_n^\epsilon = 2 | X_i^\epsilon = 1, i \in [n-\bar{T}+1, n], \text{ one jump occurred in } [n-T+1, n-\bar{T}]) \\
&\cdot P(\text{one jump occurred in } [n-T+1, n-\bar{T}] | X_{n-\bar{T}+1}^\epsilon = 1) \\
\le\;& C\epsilon^2 T^2 + P(\text{one jump occurred in } [n-\bar{T}+1, n] \;|X_n^\epsilon = 1) + \mu_1^{\otimes T}(Z_n \le 0) \\
&+\mu_1^{\otimes \bar{T}}(\bar{Z}_n \le (1+\frac{a\delta}{2})\log(\epsilon)) + C\epsilon\log(1/\epsilon)\mu_1^{\otimes \bar{T}}(\bar{Z}_n \le -(1+\frac{a\delta}{2})\log(\epsilon)),
\end{aligned} \tag{8}
$$

where the last three terms in (8) are due to the fact that $\tilde{X}_n^\epsilon = 2$ implies always that $\bar{Z}_n \le -(1+\frac{a\delta}{2})\log(\epsilon)$, while, if no jump occured in $[n-T+1, n]$, $\tilde{X}_n^\epsilon = 2$ implies that either $Z_n \le 0$ or $\bar{Z}_n \le (1+\frac{a\delta}{2})\log(\epsilon)$.

Using Chebycheff's inequality and the fact that $E_{\mu_1}\exp(\log d\mu_2/d\mu_1)) = 1$, one obtains the bounds

$$
\mu_1^{\otimes \bar{T}}(\bar{Z}_n \le (1+\frac{a\delta}{2})\log(\epsilon)) \le \exp((1+a\delta/2)\log(\epsilon)) = \epsilon^{1+a\delta/2},
$$

and, letting $\xi_i = \log(d\mu_1/d\mu_2)(y_{n-i}^\epsilon) - a$,

$$
\mu_1^{\otimes \bar{T}}(\bar{Z}_n \le -(1+\frac{a\delta}{2})\log(\epsilon)) \le \mu_1^{\otimes \bar{T}}(\frac{1}{\bar{T}}\sum_{i=1}^{\bar{T}}\xi_i \le \frac{-a\delta}{2(\delta+a^{-1})}) = o(1). \tag{9}
$$

(Actually, using the argument in (4), the right hand side in the last inequality can be bounded by $(\log(1/\epsilon))^{-\alpha/2}$). Combining the last two bounds with (7) and (8), one gets that for some $\delta' > 0$,

$$
P(\text{error} \,|\, X_n^\epsilon = 1) \le P(\text{one jump occurred in } [n-\bar{T}+1, n] \;|X_n^\epsilon = 1) + \epsilon^{(1+\delta')} + \epsilon\log(1/\epsilon)o(1),
$$

and similarly

$$
P(\text{error} \,|\, X_n^\epsilon = 2) \le P(\text{one jump occurred in } [n-\bar{T}+1, n] \;|X_n^\epsilon = 2) + \epsilon^{(1+\delta')} + \epsilon\log(1/\epsilon)o(1).
$$

Hence, using the fact that the first term is of order $\epsilon \log(1/\epsilon)$,

$$\begin{aligned} P(\text{error}) &\leq P(\text{one jump occurred in } [n-\bar{T}+1, n]\ )(1+o(1)) \\ &= p_s(1)\epsilon\lambda_{12}\bar{T}(1+o(1)) + p_s(2)\epsilon\lambda_{21}\bar{T}(1+o(1)) , \end{aligned}$$

which completes the proof of the upper bound in this case.

Turning to the general case, we assume throughout that all the a_{ij} differ, the general case following by continuity. Unfortunately, the situation here is more complex because there are more than two simple hypotheses to test. In particular, it is not clear a-priori what number should play the role of $\bar{T}$ in the previous situation. This results with a considerably more cumbersome suboptimal filter. Thus, let $\delta > 0$ be given, let

$$\Lambda_{ij} = \log E_{\mu_i}(d\mu_j/d\mu_i)^\eta ,$$

$0 < \Lambda_{ij}^* = \sup_{\eta\in\mathbb{R}} -\Lambda_{ij}(\eta) < \infty$. Next, define $T_{ij} = (1/a_{ji} + \delta)\log(1/\epsilon)$, $\bar{T} = \max_{i\neq j}(T_{ij}, (1+\delta)\log(1/\epsilon)/\Lambda_{ij}^*)$. Define next

$$Z_{n,ij}^k = \sum_{t=n-kT+1}^{n-(k-1)T} \log \frac{d\mu_i}{d\mu_j}(y_t^\epsilon) \quad k = 1,2,3 ,$$

$$D^k = \{i : Z_{n,ij}^k > 0 \quad \forall j \neq i\} ,$$

with $D^k = 1$ if there is no i satisfying the above constraints. Note that D^k is uniquely determined. Next, if $D^1 = D^2 = i$ then $I = i$. If $D^1 \neq D^2$ then $I = D^3$. Let $\sigma(\cdot) = \sigma^I(\cdot)$ denote the bijection $\{1, \ldots, d-1\} \to \{1, \ldots d\} \setminus \{i\}$ such that $T_{I,\sigma(j)} \leq T_{I,\sigma(j+1)}$. Define $\rho_j(n) = \sum_{t=n-T_{I,\sigma(j)}}^n \log \frac{d\mu_{\sigma(j)}}{d\mu_I}(y_t^\epsilon)$, and let $\delta' = \delta \min_{i\neq j} a_{ij}/2$. Then, $\tilde{X}_n^\epsilon = \sigma(j)$ if, for all $j' < j < d-1$, $\rho_{j'}(n)/\log(1/\epsilon) < (1+\delta')$ but $\rho_j(n)/\log(1/\epsilon) \geq (1+\delta')$ (with $\tilde{X}_n^\epsilon = D^1$ if none of the above conditions hold). Heuristically, error in the decision D^k occurs with meaningful probability only if a jump occurred in the interval $N_k = [n-kT+1, n-(k-1)T]$. Since the probability of two such jumps is negligible, the decision after the first stage of the algorithm is, up to a negligible probability, either on a correct estimate, or a jump occurred from state i to some state j and one needs only try to detect where has the jump occurred. The latter is achieved by a likelihood test adapted to the appropriate divergence.

Turning to the proof, let

$$J_0 = \{\text{there was a jump in more than one of the intervals } N_k\} ,$$

$$J_k = \{\text{there was a jump in the interval } N_k \text{, and none in } N_{k'}, k \neq k'\}, k = 1,2,3,$$

$$J_4 = \{\text{there was no jump in any of the intervals } N_k\}.$$

Then,

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon) \leq P(J_0) + \sum_{k=1}^4 P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_k) \leq O(\epsilon^2 \log^2(1/\epsilon)) + \sum_{k=1}^4 P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_k). \tag{10}$$

We make use of the following lemma, whose proof is identical to the proof of (4), and is therefore omitted.

Lemma 2. *Assume (A). Then,*

$$P(|\sum_{t=n-m+1}^{n} \log \frac{d\mu_i}{d\mu_j}(y_t^\epsilon) - m(a_{kj} - a_{ki})| \geq \delta m | X_t^\epsilon = k, t = n-m+1, \ldots, n) \leq c(\delta, \{a_{ij}\}) m^{-\alpha/2}.$$

Note next that if J_3 occured, an error may occur only if either $D^1 \neq X_n^\epsilon$, $D^2 \neq X_n^\epsilon$, or an error occured in the likelihood tests involving $\rho_j(n)$. Hence,

$$\begin{aligned}
P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_3) &\leq 2P(J_3) \max_i P(D^1 \neq i | X_t^\epsilon = i, t \in N_1) + \\
&\qquad d^2 P(J_3) \max_{j \neq i} P(\sum_{t=n-T_{ij}}^{n} \log \frac{d\mu_j}{d\mu_i}(y_t^\epsilon) > (1+\delta') \log(1/\epsilon) | X_t^\epsilon = i, J_3) \\
&\leq C(\epsilon \log(1/\epsilon))(\log(1/\epsilon))^{-\alpha/2},
\end{aligned} \tag{11}$$

where we have used Lemma 2 in the last inequality.

Next, denoting by $\sigma^i(\cdot)$ the bijection $\sigma(\cdot)$ corresponding to $I = i$, and $\rho_j^i(n) = \sum_{n-T_{i,\sigma^i(j)}}^n \log \frac{d\mu_{\sigma^i(j)}}{d\mu_i}(y_t^\epsilon)$,

$$\begin{aligned}
P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_2) &\leq d^2 P(J_2) \max_{k,i} P(\tilde{X}_n^\epsilon \neq \sigma^i(k) | X_t^\epsilon = \sigma^i(k), t = n-T, \ldots, n, I = i) \\
&\leq d^3 P(J_2) \max_{j<k,i} P(\rho_j^i > (1+\delta') \log(1/\epsilon) | X_t^\epsilon = \sigma^i(k), t = n - T_{i,\sigma^i(j)}, \ldots, T) \\
&\quad + d^2 P(J_2) \max_{k,i} P(\rho_k^i < (1+\delta') \log(1/\epsilon) | X_t^\epsilon = \sigma^i(k), t = n - T_{i,\sigma^i(j)}, \ldots, T).
\end{aligned}$$

But, since

$$\frac{T_{i,\sigma^i(j)}}{\log(1/\epsilon)} E_{\sigma^i(k)} \left(\log \frac{d\mu_{\sigma^i(j)}}{d\mu_i} \right) = (a_{\sigma^i(k)i} - a_{\sigma^i(k)\sigma^i(j)})(\frac{1}{a_{\sigma^i(j),i}} + \delta) \leq 1 - \frac{a_{\sigma^i(k),i}}{a_{\sigma^i(j),i}} + C\delta < 1 + \delta'$$

for all $j < k$ and δ small enough, whereas

$$\frac{T_{i\sigma^i(j)}}{\log(1/\epsilon)} E_{\sigma^i(j)} \left(\log \frac{d\mu_{\sigma^i(j)}}{d\mu_i} \right) = 1 + \delta a_{\sigma^i(k)i} \geq 1 + 2\delta' ,$$

an application of Lemma 2 yields that, for δ small enough,

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_2) = 0(\epsilon \log(1/\epsilon)^{1-\alpha/2}) . \tag{12}$$

Turning to J_1, note that

$$P(I \neq X_{n-T}^\epsilon | J_1) \leq 2P(D^3 \neq X_{n-T}^\epsilon | J_1) \leq C \log(1/\epsilon)^{-\alpha/2} ,$$

and also

$$\begin{aligned} P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_1) \quad \leq \quad & P(J_1)P(I \neq X_{n-T}^\epsilon | J_1) \\ & + P(J_{11})P(X_n^\epsilon \neq \tilde{X}_n^\epsilon | J_{11}, I = X_{n-T}^\epsilon) + P(J_{12})P(X_n^\epsilon \neq \tilde{X}_n^\epsilon | J_{12}, I = X_{n-T}^\epsilon) , \end{aligned} \tag{13}$$

where

$$J_{11} = \{\text{there was a jump in the interval } [n - T, n - \max_{j \neq I} T_{Ij}]\} , J_{12} = J_1 \setminus J_{11} .$$

Hence, by the same proof as for J_2,

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_1) \leq C\epsilon \log(1/\epsilon)^{1-\alpha/2} + P(J_{12})P(X_n^\epsilon \neq \tilde{X}_n^\epsilon | J_{12}, I = X_{n-T}^\epsilon) .$$

Concerning the last term, let

$$J_{12}^{j,\tau} = J_{12} \cap \{X_t^\epsilon = j, t = n - \tau, \ldots, n \ , X_t^\epsilon = I, t = n - T, \ldots, n - \tau - 1\} .$$

Note that $P(\cup_{\tau=0}^{T_{I,j}} J_{12}^{j,\tau}) = \epsilon\lambda_{Ij} T_{I,j}(1 + o(1))$, while by the same argument as for J_2, for $\tau > T_{I,j}$,

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_{12}^{j,\tau} | I = X_{n-T}^\epsilon) \leq C\epsilon \log(1/\epsilon)^{-\alpha/2} .$$

We conclude that

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_1; X_{n-T}^\epsilon = i) \leq C\epsilon \log(1/\epsilon)^{1-\alpha/2} + \sum_{j \neq i} \epsilon\lambda_{ij} T_{ij}(1 + o(1)) ,$$

and hence

$$P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_1) \leq C\epsilon \log(1/\epsilon)^{1-\alpha/2} + \sum_i p_s(i) \sum_{j \neq i} \epsilon\lambda_{ij} T_{ij}(1 + o(1)) . \tag{14}$$

Finally, using the relation $E_i(e^{\log d\mu_j/d\mu_i(x)}) = 1$ and Chebycheff's inequality, together with the argument leading to (9),

$$\begin{aligned} P(X_n^\epsilon \neq \tilde{X}_n^\epsilon; J_4) &\leq d^2 \max_{i\neq j} P(\sum_{n-T_{ij}}^{n} \log d\mu_j/d\mu_i(y_t^\epsilon) \geq (1+\delta')\log(1/\epsilon)|X_t^\epsilon = i, t = n - T_{ij}, \dots, n) \\ &\quad + P(D^1 \neq i|X_t^\epsilon = i, t \in N_1 \cup N_2) \\ &\leq d^2 e^{-(1+\delta')\log(1/\epsilon)} + \epsilon \log(1/\epsilon) o(1) = \epsilon \log(1/\epsilon) o(1)\,. \end{aligned} \tag{15}$$

Combining (10), (11), (12), (14) and (15) yields the upper bound, and hence the theorem. □

References

[1] A. Dembo and O. Zeitouni, Large deviations techniques and applications, Jones and Bartlett, Boston, 1993.

[2] R. Z. Khasminskii and B. V. Lazareva, *On some filtration procedure for jump Markov process observed in white Gaussian noise,* Ann. Statistics, 20, 1992, pp. 2153–2160.

[3] E. L. Lehmann, Testing statistical hypotheses, Wiley, New York, 1986.

[4] A. N. Shiryayev, Probability, Springer, New York, 1984.

[5] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering,* SIAM J. Control, 2, 1965, pp. 347–368.