Adv. & Appl. in Stat. **5**(2) (2005), 183-195

STATIONARY DISTRIBUTION ESTIMATION IN HIDDEN MARKOV MODELS

C. C. Y. DOREA, G. L. GILARDONI and C. R. GONÇALVES

Instituto de Ciências Exatas
Universidade de Brasília
70910-900 Brasília - DF, Brazil

Abstract

Let $\{X_n, Y_n\}$ describe a HMM with values on a denumerable space, being $\{Y_n\}$ the observable process. In this note, we present a class of kernel estimates for the stable distribution of $\{Y_n\}$. It is shown that the estimates are strongly consistent with exponential rate of convergence. Also, we exhibit situations where the stationary distribution of the non-observable process $\{X_n\}$ can be determined through the stable distribution of $\{Y_n\}$.

1. Introduction

Hidden Markov Models (HMM) are based on a non-observable Markov chain $\{X_n\}$ which describes the evolution of the state of a system. Associated with this chain we observe a sequence of conditionally independent random variables $\{Y_n\}$, with the distribution of each Y_n depending on the corresponding state X_n. HMM form a class of stochastic processes models that play an important role in a wide variety of

2000 Mathematics Subject Classification: Primary 62M09, 62G07.

Key words and phrases: hidden Markov model, kernel estimate, stationary distribution.

Partially supported by CAPES/PROCAD-Brazil.

Partially supported by CNPq, FINEP/PRONEX-Brazil.

Received August 11, 2003

applications (see, for example, Rabiner [5]). A classical example occurs in signal processing: a sequence of characters $\{X_n\}$ from a finite alphabet is transmitted, and a sequence $\{Y_n\}$ of corrupted signal, either by noise or by transmission distortion, is received.

A central problem in these models is that of finding properties of the chain $\{X_n\}$ based on a finite number of observations from the process $\{Y_n\}$. We will be concerned here with the problem of estimating the stationary distribution of the chain.

Let $\{X_n\}_{n\geq 0}$ be a Markov chain with state space S and transition matrix P. The observable process $\{Y_n\}_{n\geq 0}$ with values on S satisfies

$$P(Y_n = j_n \mid Y_0 = j_0, \dots, Y_{n-1} = j_{n-1}, X_0 = i_0, \dots, X_n = i_n)$$

$$= P(Y_n = j_n \mid X_n = i_n) = Q_{i_n j_n} \tag{1}$$

and

$$P(Y_0 = j_0, \dots, Y_n = j_n \mid X_0 = i_0, \dots, X_n = i_n) = Q_{i_0 j_0} \cdots Q_{i_n j_n}. \tag{2}$$

Assume that $\{X_n\}$ is ergodic and that it converges at a geometric rate, that is, there is a probability π on S and constants $\gamma > 0$ and $0 \leq \rho < 1$ such that

$$| P_{ij}^n - \pi(j) | \leq \gamma \rho^n, \quad \forall i \in S. \tag{3}$$

Clearly, the equilibrium distribution π coincides with the unique stationary distribution of the chain. Also, if S is finite the assumption of ergodicity suffices to guarantee (3). Now let

$$\nu(j) = \sum_{i\in S} \pi(i) Q_{ij}. \tag{4}$$

Then, in some sense, ν represents the distribution of Y_n when the process reaches some "stable" status. Though the knowledge of $\nu(\cdot)$ does not determine the stationary distribution π, in the Section 3 we exhibit situations under which this can be accomplished.

To estimate ν define for $i \in S$,

$$\nu_n(i) = \frac{1}{n} \sum_{k=1}^{n} W(h, i, Y_k), \tag{5}$$

where the window $h = h_n > 0$ and the weight kernel functions $W(h, i, \cdot)$ are suitably chosen. We can interpret (5) as weighted linear combination of relative frequencies

$$\nu_n(i) = \sum_{j \in S} W(h, i, j) \left[\frac{1}{n} \sum_{k=1}^{n} 1_{\{j\}}(Y_k) \right].$$

Also, as pointed out in Campos and Dorea [1], it can be viewed as a discrete version of the kernel estimate $\frac{1}{nh} \sum W(h, x, Y_k)$ used when $\nu(x)$ is a density function. Just regard h as the Lebesgue measure of $\left(x - \frac{h}{2}, x + \frac{h}{2} \right)$ and, in the discrete case, use counting measure around $\{i\}$. Our main results, Theorem 1 and Corollary 1, provide sufficient conditions for the strong consistency of $\nu_n(\cdot)$ as well as its rate of convergence. These results constitute a discrete version of Theorem 2 from Dorea and Zhao [4].

2. Preliminaries and Statement of the Results

Since $\nu_n(\cdot)$ estimates a probability, it is natural to require $\sum_{i \in S} \nu_n(i) = 1$, that is, the kernel $W \geq 0$ satisfies

$$\sum_{i \in S} W(h, i, j) = 1, \quad \forall j \in S \text{ and } h > 0. \tag{6}$$

Also, Lemma 1 below shows that

$$\lim_{n \to \infty} \left| \nu_n(i) - \nu(i) \sum_{j \in S} W(h, i, j) \right| = 0. \tag{7}$$

Thus, for asymptotic unbiasedness it is required that $W(h, i, \cdot)$ is a probability on S. For technical reasons, we further require that

$$0 < h = h_n \underset{n\to\infty}{\to} 0 \quad \text{and} \quad \lim_{n\to\infty} W(h, i, j) = 1_{\{i\}}(j). \tag{8}$$

The simplest example of a kernel W is provided by the relative frequencies estimate

$$\nu_n^0(i) = \frac{1}{n}\sum_{k=1}^{n} 1_{\{i\}}(Y_k). \tag{9}$$

Wang and Van Ryzin [7] consider several other kernels to estimate a discrete distribution on the integers under the independent and identically distributed setting, e.g., the uniform kernel function

$$W(h, i, j) = \begin{cases} h/2k & \text{if } |j-i| = 1, \ldots, k, \\ 1-h & \text{if } j = i, \\ 0 & \text{if } |j-i| > k, \end{cases}$$

and the geometric kernel function

$$W(h, i, j) = \begin{cases} (1-h)h^{|j-i|}/2 & \text{if } |j-i| \geq 1, \\ 1-h & \text{if } j = i. \end{cases}$$

Lemma 1. *For any probability* μ *on* S, *we have*

$$\lim_{n\to\infty} \sum_{i\in S} \left| \sum_{j\in S} W(h, i, j)\mu(j) - \mu(i) \right| = 0, \tag{10}$$

and $\nu_n(\cdot)$ *is asymptotically unbiased, i.e.,* $\lim_{n\to\infty} E(\nu_n(i)) = \nu(i)$.

Lemma 2. *Let* $\nu_n^0(\cdot)$ *be defined by* (9) *and assume that, given* $\varepsilon > 0$, *there exist constants* $b_1 = b_1(\varepsilon)$ *and* $b_2 = b_2(\varepsilon) > 0$ *such that*

$$P(|\nu_n^0(i) - \nu(i)| \geq \varepsilon) \leq b_1 e^{-b_2 n}, \quad \forall\, i \in S. \tag{11}$$

Then, there exist constants $c_1 = c_1(\varepsilon)$ *and* $c_2 = c_2(\varepsilon) > 0$ *such that*

$$P\left(\sum_{i\in S} |\nu_n(i) - \nu(i)| \geq \varepsilon \right) \leq c_1 e^{-c_2 n}. \tag{12}$$

Lemma 3 (Devroye [2]). *Let* $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots \subset \mathcal{G}_n$ *be a sequence of nested* σ*-algebras. Let* U *be a* $\mathcal{G}_n$*-measurable and integrable random variable, and define the Doob martingale* $U_k = E(U \mid \mathcal{G}_k)$*. Assume that there exist* $\mathcal{G}_{k-1}$*-measurable random variable* V_k *and constants* a_k *such that* $V_k \leq U_k \leq V_k + a_k$*. Then, given* $\varepsilon > 0$,

$$P(|U - EU| \geq \varepsilon) \leq 4 \exp\left\{-2\varepsilon^2 \middle/ \sum_{k=1}^{n} a_k^2\right\}.$$

Theorem 1. *Given* $\varepsilon > 0$ *there exist constants* $c_1 = c_1(\varepsilon)$ *and* $c_2 = c_2(\varepsilon) > 0$ *such that*

$$P\left(\sum_{i \in S} |\nu_n(i) - \nu(i)| \geq \varepsilon\right) \leq c_1 e^{-c_2 n}. \tag{13}$$

Note that (13) holds regardless of the initial distribution. Hence we are not assuming here strict stationarity of the chain. Since $\sum_{n \geq 1} c_1 e^{-c_2 n} < \infty$, an application of the Borel-Cantelli lemma gives the corollary below.

Corollary 1. *For any initial distribution the estimator* $\nu_n(\cdot)$ *is strongly consistent,*

$$P(\lim_{n \to \infty} \nu_n(i) = \nu(i)) = 1.$$

Moreover,

$$P\left(\lim_{n \to \infty} \sum_{i \in S} |\nu_n(i) - \nu(i)| = 0\right) = 1. \tag{14}$$

From Doob [3, Chapter 5], for finite S the necessary and sufficient condition for the ergodicity of the chain, hence for its geometric ergodicity (S), is given by the simple condition: there exist $j_0 \in S$, $\delta > 0$ and $n_0 \geq 1$ such that

$$\min_{i \in S} P_{ij_0}^{n_0} \geq \delta. \tag{15}$$

Corollary 2. *Let* $\{X_n\}$ *be any Markov chain with finite state space* S *that satisfies* (15). *Then, we have* (14).

Corollary 3. *Assume that the state space* S *is finite, the matrix* $Q = (Q_{ij})$ *is non-singular and the stable distribution* π *satisfies that* $\pi(i) > 0$ *for every* $i \in S$. *Then*

$$P(\lim_{n\to\infty} \pi_n(i) = \pi(i)) = 1,$$

where $\pi_n = \nu_n Q^{-1}$.

3. Estimation of π

Quite often the main objective in this problem would be to use the estimate $\nu_n(\cdot)$ of $\nu(\cdot)$ in order to produce an estimate $\pi_n(\cdot)$ for $\pi(\cdot)$. In this section we will discuss some aspects of this problem. Assume for the rest of the section that $S = \{0, 1, \ldots, N\}$ is finite and let Q be the matrix with entries $Q_{ij} = P(Y_n = j \mid X_n = i)$, so that (4) can be written as $\nu = \pi Q$, where $\nu = (\nu(1), \ldots, \nu(N))$ and similarly for π. Further, we will assume that all entries in π are strictly positive and of course they must sum to 1. Since the matrix Q is stochastic $\left(\sum_j Q_{ij} = 1\right)$, it follows from $\nu = \pi Q$ that $\sum_j \nu(j) = \sum_i \pi(i) = 1$ and $\nu(j) \geq 0$ for every j.

Of course, if $\nu = \pi Q$ and $\det(Q) \neq 0$, then there is a unique $\pi = \nu Q^{-1}$ that satisfies (4). Also in this case, since each column of Q must have at least one positive entry, strict positivity of π also implies strict positivity of ν. However, in our problem we want to define $\pi_n = \nu_n Q^{-1}$, and although from the argument above it follows that $\sum_i \pi_n(i) = \sum_j \nu_n(j) = 1$, we have no guarantee that the entries of π_n would be non-negative. In short, this problem is caused by the fact that if $\mathcal{P}$ is the set of all probabilities on S with strictly positive entries, then it is possible that the set $\mathcal{Q} = \{\nu = \pi Q : \pi \in \mathcal{P}\}$ is a proper subset of $\mathcal{P}$, and even when ν belongs to $\mathcal{P}$, it may be that ν_n as defined in (5) does not.

Although the situation described above may happen for finite n, Corollaries 2 and 3 assure that as $n \to \infty$, with probability 1 we will have that $\pi_n = \nu_n Q^{-1}$ will be a probability distribution on S with strictly positive entries. To see this, observe that the mapping $\nu \mapsto \pi = \nu Q^{-1}$ is continuous, so that if for ν as in (4), we have that $\pi = \nu Q^{-1}$ have positive entries, the same must happen for all ν^* in a neighborhood of ν. Hence Corollary 3 follows now from (14).

Below we discuss a few examples using kernels that concentrate weight on a neighborhood of the target point i to illustrate these ideas.

Example 1 (Uniform link)**.** Assume that when the signal i is transmitted it is correctly interpreted with probability $Q_{ii} = \alpha > 1/2$, while it will read as one of the remaining characters with equal probability $Q_{ij} = \beta = \frac{1-\alpha}{N-1}$, $j \neq i$. Hence $\nu(i) = \alpha\pi(i) + \beta\sum_{j\neq i}\pi(j) = \alpha\pi(i)$ $+ \beta(1 - \pi(i))$. Hence $\pi(i) = \frac{(N-1)\nu(i) - (1-\alpha)}{(N-1)\alpha - (1-\alpha)}$. Note that for π to have positive entries we must have that $\nu(i) > (1-\alpha)/(N-1)$.

Example 2. Let $Q_{ii} = \alpha > \frac{1}{2}$ and $Q_{ij} = \beta = \frac{1-\alpha}{2}$ for $j = i-1$ or $i+1 (\mathrm{mod}\ N)$. Here we have $\nu(i) = \alpha\pi(i) + \beta[\pi(i-1) + \pi(i+1)]$. Since Q is non-singular we have the solution $\pi = Q^{-1}\nu$. For instance, for $N = 4$,

$$\pi(1) = \pi(3)$$

$$= \frac{1}{2(2\alpha-1)}\left(\frac{(\alpha^2 + 2\alpha - 1)}{\alpha}\nu(0) + (\alpha-1)\nu(1) + \frac{(\alpha-1)^2}{\alpha}\nu(2) + (\alpha-1)\nu(3)\right)$$

and

$$\pi(2) = \pi(4) = \frac{1}{2(2\alpha-1)}((\alpha-1)(\nu(0)+\nu(3)) + (\alpha-1)^2\nu(1) + (\alpha^2 - 1 + 2\alpha)\nu(3)).$$

Example 3. Let $Q_{ii} = 1/2$ and $Q_{ij} = 1/4$ for $j = i-1$ or $i+1 (\mathrm{mod}\ N)$. Then we have $\nu(i) = \frac{1}{2}\pi(i) + \frac{1}{4}[\pi(i-1) + \pi(i+1)]$. If $N \geq 4$ is even, then

$\det(Q) = 0$ and we can usually find many solutions. For instance, for $N = 4$ the system will have a solution only when $\nu(1) - \nu(2) + \nu(3) - \nu(4) = 0,$ when we obtain for a given $\pi(4)$ that $\pi(1) = \nu(2) - 2\nu(3) + 3\nu(4) - \pi(4),$ $\pi(2) = 2\nu(2) - 2\nu(4) + \pi(4)$ and $\pi(3) = -\nu(2) + 2\nu(3) + \nu(4) - \pi(4).$ Observe here that the condition $\det(Q) \neq 0$ is not necessary in order to have uniqueness of π for a given ν. Indeed, when $N = 4$ and for $\nu(1) = \nu(2) = \frac{3}{8}$ and $\nu(3) = \nu(4) = \frac{1}{8},$ we have the unique solution $\pi(1) = \pi(2) = \frac{1}{2}$ and $\pi(3) = \pi(4) = 0$ for π.

4. Proof of the Results

Proof of Lemma 1. (i) Since $W(h, i, \cdot) \leq 1,$ we have by (8) and using dominated convergence that

$$\lim_{n\to\infty} \sum\nolimits_{j\in S} W(h, i, j)\mu(j) = \mu(i).$$

Given $\varepsilon > 0$ let S_ε be a finite subset of S such that $\sum_{i\in S_\varepsilon} \mu(i) \geq 1 - \varepsilon/4.$ Let N_ε be such that for $n \geq N_\varepsilon,$

$$A = \sum_{i\in S_\varepsilon} \left| \sum_{j\in S} W(h, i, j)\mu(j) - \mu(i) \right| \leq \varepsilon/4.$$

Then,

$$B = \sum_{i\in S_\varepsilon} \sum_{j\in S} W(h, i, j)\mu(j) \geq \sum_{i\in S_\varepsilon} \mu(i) - \varepsilon/4 \geq 1 - \varepsilon/2$$

and

$$\sum_{i\in S} \left| \sum_{j\in S} W(h, i, j)\mu(j) - \mu(i) \right| \leq A + B + \sum_{i\in S_\varepsilon} \mu(i) \leq \varepsilon.$$

(ii) Assume that the chain is strictly stationary, that is, the initial distribution is π. Then X_k has distribution π and by (1) and (4),

$$P(Y_k = j) = \sum_r P(X_k = r)Q_{rj} = \sum_r \pi(r)Q_{rj} = \nu(j).$$

From (5) we have

$$E(\nu_n(i)) = E(W(h, i, Y_k)) = \sum_j W(h, i, j)\nu(j).$$

And from (10) we have the asymptotic unbiasedness.

Now, assume that the initial distribution is π_0. Since (3) holds, it follows from Roussas and Ioannides [6, Proposition 3.1] that there exists a constant $\gamma' > 0$ such that

$$\sum_{j \in S} |P_{ij}^k - \pi(j)| \le \gamma' \rho^k. \tag{16}$$

It follows that

$$\left| E(W(h, i, Y_k)) - \sum_j W(h, i, j)\nu(j) \right| \le \sum_j W(h, i, j) \sum_r \pi_0(r) \sum_s |P_{rs}^k - \pi(s)| Q_{sj}$$

$$\le \sum_j W(h, i, j) Q_{sj} \gamma' \rho^k \le \gamma' \rho^k.$$

Then

$$\left| E(\nu_n(i)) - \sum_j W(h, i, j)\nu(j) \right| \le \frac{1}{n} \sum_{k=1}^n \gamma' \rho^k \underset{n \to \infty}{\to} 0$$

and the desired result follows from (10).

Proof of Lemma 2. (i) Given $\varepsilon > 0$ let S_ε be finite and such that

$$\sum_{i \in S_\varepsilon} \nu(i) \ge 1 - \varepsilon/8. \tag{17}$$

From (10) there exists N_ε satisfying for $n \ge N_\varepsilon$,

$$\sum_{i \in S} | \lambda_n(i) - \nu(i) | \le \frac{\varepsilon}{8}, \tag{18}$$

where $\lambda_n(i) = \sum_{j \in S} W(h, i, j)\nu(j)$. From (11) there exist b_1' and $b_2' > 0$ such that

$$P\left(\sum_{i \in S_\varepsilon} |\nu_n^0(i) - \nu(i)| \geq \frac{\varepsilon}{8}\right) \leq b_1' e^{-b_2' n}. \tag{19}$$

(ii) We can write

$$\nu_n(i) = \frac{1}{n} \sum_{j \in S} W(h, i, j) \sum_{k=1}^{n} 1_{\{j\}}(Y_k) = \sum_{j \in S} W(h, i, j)\nu_n^0(j).$$

Define $\nu_n^*(i) = \sum_{j \in S_\varepsilon} W(h, i, j)\nu_n^0(j)$. By (6), we have

$$\sum_{i \in S} |\nu_n(i) - \nu_n^*(i)| = \sum_{i \in S} \sum_{j \in S_\varepsilon^c} W(h, i, j)\nu_n^0(j) = \sum_{j \in S_\varepsilon^c} \nu_n^0(j).$$

From (19), we have

$$P\left(\sum_{i \in S_\varepsilon} \nu_n^0(i) \leq \sum_{i \in S_\varepsilon} \nu(i) - \varepsilon/8\right) \leq b_1' e^{-b_2' n}$$

and from (17),

$$P\left(\sum_{i \in S_\varepsilon} \nu_n^0(i) \leq 1 - \varepsilon/4\right) \leq b_1' e^{-b_2' n}.$$

Hence,

$$P\left(\sum_{i \in S} |\nu_n(i) - \nu_n^*(i)| \geq \frac{\varepsilon}{3}\right) \leq P\left(\sum_{i \in S_\varepsilon^c} \nu_n^0(i) \geq \frac{\varepsilon}{4}\right) \leq b_1' e^{-b_2' n}. \tag{20}$$

(iii) From (18), we have

$$\sum_{i \in S} |\nu_n^*(i) - \lambda_n(i)| \leq \sum_{i \in S} \left| \sum_{j \in S_\varepsilon} W(h, i, j)[\nu_n^0(j) - \nu(j)] \right| + \sum_{i \in S} \sum_{j \in S_\varepsilon^c} W(h, i, j)\nu(j)$$

$$\leq \sum_{j \in S_\varepsilon} |\nu_n^0(j) - \nu(j)| + \sum_{j \in S_\varepsilon^c} \nu(j).$$

Then by (17) and (19),

$$P\left(\sum_{i\in S}|\nu_n^*(i)-\lambda_n(i)|\geq\frac{\varepsilon}{3}\right)\leq P\left(\sum_{j\in S_\varepsilon}|\nu_n^0(j)-\nu(j)|\geq\frac{\varepsilon}{6}\right)+P\left(\sum_{j\in S_\varepsilon^c}\nu(j)\geq\frac{\varepsilon}{6}\right)$$

$$\leq b_1'e^{-b_2'n}. \tag{21}$$

By (18) we have for $n\geq N_\varepsilon$,

$$P\left(\sum_{i\in S}|\lambda_n(i)-\nu(i)|\geq\frac{\varepsilon}{3}\right)=0. \tag{22}$$

Finally, from (20), (21) and (22) we have (12) with $c_1=2b_1'$ and $c_2=b_2'$.

Proof of Theorem 1. (i) By Lemma 2 it is enough to show that (11) holds. Hence, we must prove that given $\varepsilon>0$ there exist $b_1=b_1(\varepsilon)$ and $b_2=b_2(\varepsilon)>0$ such that

$$P\left(\left|\nu_n^0(i)-\frac{1}{n}\sum_{j=1}^n Q_{X_ji}\right|\geq\varepsilon\right)\leq b_1e^{-b_2n} \tag{23}$$

and

$$P\left(\left|\nu(i)-\frac{1}{n}\sum_{j=1}^n Q_{X_ji}\right|\geq\varepsilon\right)\leq b_1e^{-b_2n}. \tag{24}$$

Proceeding as in the proof of Lemma 1 it suffices to show (23) and (24) when the initial distribution is the stationary distribution π. It will be carried out using Lemma 3. Define $\mathcal{F}_n=\sigma(X_0, X_1, \ldots, X_n)$.

(ii) Let $U=\sum_{j=1}^n[1_{(Y_{j=i})}-Q_{X_ji}]$. Then

$$U=\sum_{j=1}^n[1_{(Y_{j=i})}-E(1_{(Y_{j=i})}\,|\,\mathcal{F}_j)]. \tag{25}$$

Clearly, U is $\mathcal{F}_n$-measurable and $EU=0$. Note that for $j\leq k$ we have $E(1_{(Y_{j=i})}\,|\,\mathcal{F}_k)=1_{(Y_{j=i})}$ and $E(Q_{X_ji}\,|\,\mathcal{F}_k)=Q_{X_ji}$, while for $j>k$,

$$E(1_{(Y_j=i)} - Q_{X_j i} \mid \mathcal{F}_k) = \sum_{\ell} P^{k-j}_{X_k \ell} Q_{\ell i} - \sum_{\ell} P^{k-j}_{X_k \ell} Q_{\ell i} = 0. \tag{26}$$

It follows that

$$U_k = E(U \mid \mathcal{F}_k) = \sum_{j=i}^{k} [1_{(Y_j=i)} - Q_{X_j i}].$$

Hence for $V_k = U_{k-1} - 1$ and $a_k = 2$ we have the hypotheses of Lemma 3 satisfied. Using (25) we have (23) with $b_1 = 4$ and $b_2 = \varepsilon^2/2$.

(iii) A key point in the proof of (23) is the identity (26). But for (24) we cannot guarantee $E\{(Q_{X_j i} - \nu(i)) \mid \mathcal{F}_k\} = 0$ for $j > k$. This difficulty can be handled by defining

$$\varphi(X_j) = Q_{X_j i} - \nu(i) + \hat{\varphi}(X_j), \quad \hat{\varphi}(X_j) = \sum_{r \geq 1} [E(Q_{X_{j+r} i} \mid \mathcal{F}_j) - \nu(i)]. \tag{27}$$

We have φ well-defined since by (4) and (16) we have

$$|\hat{\varphi}(X_j)| = \left| \sum_{r \geq 1} \sum_{\ell} (P^r_{X_j \ell} - \pi(\ell)) Q_{\ell i} \right| \leq \sum_{r \geq 1} \gamma' \rho^r < \infty. \tag{28}$$

From (27) we can write

$$\sum_{j=1}^{n} [Q_{X_j i} - \nu(i)] = \varphi(X_1) - \hat{\varphi}(X_n) + \sum_{j=2}^{n} [\varphi(X_j) - \hat{\varphi}(X_{j-1})].$$

By (28) we have $|\varphi(X_1) - \hat{\varphi}(X_n)|$ bounded and for n large

$$P(|\varphi(X_1) - \hat{\varphi}(X_n)| \geq n\varepsilon) = 0. \tag{29}$$

Now define $U = \sum_{j=2}^{n} [\varphi(X_j) - \hat{\varphi}(X_{j-1})]$ so that we will have (24) if

$$P\left(|U| \geq \frac{n\varepsilon}{2}\right) \leq b_1 e^{-b_2 n}. \tag{30}$$

This would follow if the hypotheses of Lemma 3 are satisfied. First, note that

$$E(\varphi(X_j) \mid \mathcal{F}_{j-1}) = E(Q_{X_j i} - \nu(i) \mid \mathcal{F}_{j-1}) + \sum_{r \geq 1} [E(Q_{X_{j+r} i} \mid \mathcal{F}_{j-1}) - \nu(i)]$$

$$= \sum_{s \geq 1} [E(Q_{X_{j-1+s} i} \mid \mathcal{F}_{j-1}) - \nu(i)] = \hat{\varphi}(X_{j-1}).$$

Hence $EU = 0$. Also, for $j > k$,

$$E\{(\varphi(X_j) - \hat{\varphi}(X_{j-1})) \mid \mathcal{F}_k\} = E\{E(\varphi(X_j) - \hat{\varphi}(X_{j-1}) \mid \mathcal{F}_{j-1}\} = 0.$$

Then $U_k = E(U \mid \mathcal{E}_k) = \sum_{j=2}^{k} [\varphi(X_j) - \hat{\varphi}(X_{j-1})]$. Just take $V_k = U_{k-1} - M$ and $a_k = 2M$ with M satisfying $|\varphi(\cdot) - \hat{\varphi}(\cdot)| \leq M$. And we have (29) with $b_1 = 4$ and $b_2 = \varepsilon^2/8M^2$.

References

[1] V. S. M. Campos and C. C. Y. Dorea, Kernel density estimation: the general case, Statist. Probab. Lett. 55 (2001), 173-180.

[2] L. Devroye, Exponential inequalities in nonparametric estimation, Nonparametric Functional Estimation and Related Topics, G. G. Roussas ed., pp. 31-44, Kluwer Academic Publishers, 1991.

[3] J. L. Doob, Stochastic Processes, John Wiley & Sons, New York, 1953.

[4] C. C. Y. Dorea and L. C. Zhao, Nonparametric density estimation in hidden Markov models, Stat. Infer. Stoch. Process. 178 (2001), 1-10.

[5] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989), 257-284.

[6] G. G. Roussas and D. Ioannides, Moment inequalities for mixing sequences of random variables, Stochastic Anal. Appl. 5(1) (1987), 61-120.

[7] M. C. Wang and Van Ryzin, A class of estimators for discrete distributions, Biometrika 68 (1981), 301-309.

■