DEPARTMENT OF STATISTICS

University of Wisconsin
1210 West Dayton Street
Madison, WI 53706-1693

TECHNICAL REPORT NO. 891

JULY 1992

LOWER BOUNDS ON EXPECTED REDUNDANCY FOR CLASSES OF CONTINUOUS MARKOV SOURCES

by

Bin Yu

Lower Bounds on Expected Redundancy for Classes of Continuous Markov Sources

Bin Yu

Department of Statistics

University of Wisconsin-Madison

Abstract

A nonasymptotic lower bound is derived for the per symbol expected redundancy based on n observations from a continuous dth order Markov source. The bound is minimax over a Lipschitz class of such sources. The constant in the lower bound is explicitly described in terms of d. By making d go to infinity with n at an appropriate rate, it is shown that no universal rate of expected redundancy exists for the class of Markov sources of all orders, and this provides an alternative and simpler derivation of a similar result by Shields. Similar results are obtained for the Kullback-Leibler estimation error for the joint density of d-tuples based on n observations from a continuous $(d-1)$st order Markov source.

Key Words and Phrases: code, redundancy, density estimation, Markov source, minimax.

AMS 1980 subject classifications. Primary 94A29, 62G05; secondary 60J27.

Invited paper for the Fifth Purdue Symposium on Statistical Decision Theory and Related Topics, June 14th-20th, 1992. Research supported in part by ARO Grant DAAL03-91-G-0107 and by NSF Grant DMS-8505550 to MSRI at Berkeley through a postdoc-fellowship.

1. Introduction

By means of Kraft's inequality, any probability distribution p on a finite set (alphabet) $\mathcal{A}$ of symbols corresponds to a binary prefix code, i.e., a map C_p from $\mathcal{A}$ to strings of 0's and 1's (codewords) with the property that no codeword is the prefix of another codeword. Then roughly speaking, for any $x \in \mathcal{A}$, $-\log_2 p(x)$ gives the code length of x, i.e. the count of 0's and 1's in $C_p(x)$. For example, let $\mathcal{A} = \{a, b, c\}$, $p(a) = 1/4, p(b) = 1/4, p(c) = 1/2$, then p corresponds to the prefix code C_p such that $C_p(a) = 00, C_p(b) = 01, C_p(c) = 1$ and $-\log_2 p(x) =$ length of $C_p(x)$ for all x in $\mathcal{A}$. When x is generated by a probability distribution p_0 on $\mathcal{A}$, it is not surprising that its corresponding prefix code is the best on average. This code assigns short codewords to frequent x's and long codewords to rare x's, and it follows from Jensen's inequality $E_{p_0}[-\log_2 p(x)] \geq E_{p_0}[-\log_2 p_0(x)]$, which is the entropy of p_0. The difference $E_{p_0}[-\log_2 p(x)] - E_{p_0}[-\log_2 p_0(x)]$ is called expected redundancy of the code (corresponding to) p if the symbol x is generated from p_0.

Similarly, for any positive joint density $s(x^n) = s(x_1, x_2, ..., x_n)$ on n-tuples of real numbers we may regard $-\log_2 s(x^n)$ as the code length of a binary prefix code, cf. Rissanen (1986). Its expected code length is bounded from below by the entropy of the joint density f that generates the sequence, that is, $H_n(f) = -\int f(x^n) \log_2 f(x^n) dx^n$.

Definition

$E_f(-\log_2 s(x^n)) - H_n(f) = E_f \log[f/s]$ is called **expected redundancy** of s, and $n^{-1}[E_f(-\log_2 s(x^n)) - H_n(f)] = n^{-1} E_f \log[f/s]$ is called **per symbol expected redundancy** of s.

Redundancy measures the loss in terms of expected code length when the true density f is not known. s is called a universal code if the per symbol expected redundancy of s goes to zero as n tends to infinity. Since $s(x^n)$ can always be factored into a product of

predictive or conditional densities $\prod_t s(x_t|x^{t-1})$, we have

$$E_f(-\log_2 s(x^n)) - H_n(f) = \sum_t \int \log(f(x_t|x^{t-1})/s(x_t|x^{t-1}) f(x^t) dx^t. \tag{0.1}$$

Then expected redundancy can be viewed as the accumulated prediction error measured by the Kullback-Leibler divergence. In the iid case, $f(x^n) = \prod_t f(x_t)$ and $H_n(f) = nH(f)$, and (0.1) simplifies to the accumulated density estimation error:

$$E_f(-\log_2 s(x^n)) - nH(f) = \sum_t E_{f^{t-1}} \int \log(f(x_t)/s(x_t|x^{t-1}) f(x_t) dx_t. \tag{0.2}$$

In particular, when f belongs to a parametric family $\{f_\theta, \theta \in \Theta\}$, we may choose $s(.|x^{t-1})$ as the plug-in predictive density $f_{\hat{\theta}_{t-1}}(.)$ where $\hat{\theta}_{t-1}$ is the MLE estimator, say, of θ based on the first $t-1$ observations. If the parametric family satisfies certain regularity conditions, (0.2) becomes

$$\begin{aligned} & E_f(-\log s(x^n)) - nH(f) \\ = & \sum_t E_{f^{t-1}} \int \log[f_\theta(x_t)/f_{\hat{\theta}_{t-1}}(x_t)] f_\theta(x_t) dx_t \\ \approx & \sum_t O(E(\hat{\theta}_{t-1} - \theta)^2) = O(\sum_t 1/(t-1)) = O(\log n). \end{aligned} \tag{0.3}$$

That is, the rate n^{-1} of estimating θ based on n observations translates into the rate of $n^{-1} \log n$ in terms of universal average expected redundancy, cf. Rissanen (1986) and Clarke and Barron (1990). Both n^{-1} and $n^{-1} \log n$ are the best possible rates, in the sense that lower bounds of the same order are given by the Cramér-Rao inequality and by Rissanen (1986) respectively. On the other hand, in nonparametric density estimation based on iid observations, if we restrict our class to smooth densities of a certain degree, the minimax rate for global deviation measures like L^2 and Hellinger distance is $n^{-2\alpha}$ with $\alpha < 1/2$. In contrast with the parametric case, the per symbol expected redundancy has the same

minimax rate in this case, as shown in Yu and Speed (1992). This is because instead of t^{-1} in (0.3), we have $t^{-2\alpha}$ and $n^{-1}\sum_t^n t^{-2\alpha} = O(n^{-2\alpha})$.

It is well-known, however, that no rate exists if the class is sufficiently large. There are at least two ways to enlarge the class: a) Devroye (1983, 1987) fixed the iid structure between observations and made the class large by including enough marginal densities of x_1 in the class to show the nonexisitence of any density estimation rate, hence nonexistence of a universal redundancy rate. b) Recently, Shields (1991) used the technique of cutting and stacking from ergodic theory to show the nonexistence of a universal per symbol expected redundancy rate for the class of ergodic sources on a finite alphabet $\mathcal{A}$. He enriched the class by allowing dependence structure between observations while holding the class of marginal densities of x_1 parametric. It is difficult, however, to figure out what type of dependence Shields uses since his argument is closer to an existence proof than a construction proof. He showed that for any density $s()$ on $\mathcal{A}^n$ (or equivalently a prefix code on $\mathcal{A}^n$), there exists an ergodic source for which the per symbol expected redundancy rate of s is not faster than a predescribed rate $\rho(n) = o(1)$. On the other hand, we believe that the class of Markov sources of all orders is large enough. This is because a Markov source of order $|\mathcal{A}|^d$ has number of parameters of order J^d and $|\mathcal{A}|^d$ appears in Rissanen's lower bound in front of the rate $n^{-1}\log n$. Formally taking $d = d_n = \log n$ in the bound gives the nonexistence of the universal rate for class of Markov sources of all orders; hence no rate exists for the class of ergodic sources. Of course, this is not legitimate since other terms negligible for a fixed d might become dominant when taking $d = \log n$.

In this paper, aiming at the nonexistence of universal per symbol expected redundancy rate for the class of Lipschitz Markov sources of all orders, we first provide a nonasymptotic minimax lower bound on the expected redundancy over a Lipschitz class of $(d-1)$st order

Markov chains. In Section 2, we extend from the iid case to the Markov case, the Assouad-type construction of a hypercube subclass within the Lipschitz class, cf. Assouad (1983), Birgé (1985), Devroye (1987), Donoho, McGibbon and Liu (1990). The construction is intuitive and geometric. The lower bound is given in Section 3 where the constant in the lower bound is found explicitly in d and the Lipschitz constant c_d of the class. When we fix c_d while letting d go to infinity at the rate $\log n$, the lower bound gives a rate of $(\log n)^{-5}$, namly, the per symbol expected redundancy rate cannot be faster than $(\log n)^{-5}$. Modifying the hypercube class slightly and letting $c_d = (\log n)^{5/2}$ gives the nonexistence of of the per symbol expected redundancy rate over the class of Markov sources of all orders. Similar results are then obtained for the Kullback-Leibler density estimation error for the joint density of d-tuples. We end the paper with a brief discussion (section 4) on the appropriateness of a class over which a minimax result should be sought.

2. A Lipschitz Markov Class and Its Hypercube Subclass

In this section, we introduce a continuous Lipschitz class of Markov sources of order $(d-1)$ and its hypercube subclass. Let us start with some notations. For any integer $d \geq 2$, let $x^d = (x_1, ..., x_d)$ and denote $[-1/2, 1/2]$ by J. Write

$$\mathcal{D}_d := \{f \geq 0 : \int_{J^d} f(x^d) = 1, \ |f(x^d) - f(y^d)| \leq c_d ||x^d - y^d||, \ \forall \, x^d, y^d \in J^d\}.$$

$\mathcal{D}_d$ is a Lipschitz class of joint densities of d-tuples on J^d with a Lipschitz constant c_d. In order to generate a unique $(d-1)$st order stationary Markov source from a density f in $\mathcal{D}_d$, f has to further satisfy, for any $k \neq l$, $t \leq \min(d-1, d-k, d-l)$, and any $y^t = (y_1, y_2, ..., y_t) \in J^t$,

$$\int_{J^{d-t}} f(x_1, ... \, x_{k-1}, y^t, x_{k+t}, ... \, x_d) dx_1 ... dx_{k-1} dx_{k+t} ... dx_d$$
$$\equiv \int_{J^{d-t}} f(x_1, ... \, x_{l-1}, y^t, x_{l+t}, ... \, x_d) dx_1 ... dx_{l-1} dx_{l+t} ... dx_d,$$

that is, all tth order marginals of f have to coincide. Denote the set of all such f's by $\mathcal{M}_d$. This is the Lipschitz Markov class over which minimax results will be sought. From now on we use the same f to indicate a density in $\mathcal{M}_d$ and the joint density of the Markov chain it generates.

The hypercube subclass $\mathcal{F}_{r,d}$ of $\mathcal{M}_d$ is a collection of densities in $\mathcal{M}_d$ which are suitably perturbed uniform densities on $J^d = [-1/2, 1/2]^d$. The construction of $\mathcal{F}_{r,d}$ can be divided into three steps.

1. **Off-diagonal cells or cubes of size h^d in the positive quadrant**

For h small let $r_0 = (2h)^{-1}$ and $I_i := [(i-1)h, ih]$ for $i = 1, 2, ..., r_0$. Then divide $[0, 1/2]^d$ into ${r_0}^d = (2^d h^d)^{-1}$ cubes of size h^d: $\{I_{i_1} \times I_{i_2} \times ... \times I_{i_d} : \quad 1 \leq i_1, i_2, ..., i_d \leq r_0\}$. For technical reasons which will be explained later, we only take the cubes H such that $H = I_{i_1} \times I_{i_2} \times ... \times I_{i_d}$ where $i_1, i_2, ..., i_d$ are all different. Denote the collection of these H's by $R := \{A_i, i = 1, 2, .., r\}$ with $r := |R| = r_0 \times (r_0 - 1) \times ... \times (r_0 - d + 1)$, and denote the centers of A_i by a_i.

2. **The pyramid perturbation**

The following pyramid function q is the basic "perturbation" added to the uniform density. It sits on the base $[-h/2, h/2]^d$ with a height $hc_d/2$ ($hc_d/2 < 1$); or more precisely, for $0^d := (0, 0, ...0) \in R^d$, $E := \{z_d = (0, ..., z, ...0) \in R^d, z_d = h/2 \text{ or } -h/2\}$= facet set of the hypercube $[-h/2, h/2]^d$, define $q(0^d) = hc_d/2$, and $q(z_d) = 0$ for all $z_d \in E$. For any $x^d \in [-h/2, h/2]^d$, there exists a $z_d \in E$ such that z_d is the intersection of the line connecting 0^d and x^d with the facet set E. Since the line segment between 0^d and any $z_d \in E$ belongs to the hypercube and any point in the hypercube belongs to such a line segment, it is sufficient to define q on this line segment as the linear function connecting $(0^d, hc_d/2)$ and $(z_d, 0)$.

For any $A_i \in R$ with a center a_i, define the pyramid function sitting on the base A_i as

$q_i(x^d) = q(x^d - a_i)$, and define the hypercube subclass on $[0,1/2]^d$ as

$$\mathcal{F}_{r,d} = \{f_\theta(x^d) := 1 + \sum_{i=1}^{r} I_{A_i}\theta_i q(x^d - a_i) : x^d \in [0,1/2]^d,\ \theta = (\theta_1, \theta_2, ..., \theta_r),\ \theta_i = \pm 1\}$$

3. **Extension from $[0,1/2]^d$ to $J^d = [-1/2,1/2]^d$**

Observe that for any $x^d = (x_1, x_2, \dots, x_d) \in J^d$, there exist $1 \le k_1 < k_2 < ... < k_j \le d$ such that $x^{*d} := (x_1, ..., -x_{k_1}, ..., -x_{k_j}, ..., x_d) \in [0,1/2]^d$. Define $Y : J^d \to [0,1/2]^d$ such that $Y(x^d) = x^{*d}$. For any set $A \subset [0,1/2]^d$ define $A^* := \{x : Y(x) \in A\}$. Any function g defined on $A \subset [0,1/2]^d$ can be extended to A^* as follows:

If j in the definition of x'^d is odd, define $g(x^d) = -g(Y(x^d))$; If j is even, define $g(x^d) = g(Y(x^d))$. Every function in $\mathcal{F}_{r,d}$ is now extended to J^d by the extension rule just described.

By the symmetry of the pyramid and by the symmetry of the functions in $\mathcal{F}_{r,d}$ introduced by the extension rule, the tth order ($\forall\ t \le d-1$) marginal density of f_i is the uniform density:

$$f_\theta(x_k, ..., x_{k+t-1}) = \int_{J^{d-t}} f_\theta(x_1, ..., x_k, ..., x_d) dx_1 ... dx_{k-1} dx_{k+t} ... dx_d \equiv 1.$$

Hence $\mathcal{F}_{r,d} \subset \mathcal{M}_d$. Note that f_θ's in $\mathcal{F}_{r,d}$ are flat ($\equiv 1$) on all the cubes not in R and pyramid perturbations are added only to those cubes in R. In the case of $d = 2$, $f_\theta \equiv 1$ on diagonal cubes $I_i \times I_i$ ($i = 1, .., r_0$). This construction makes f_θ closer to the uniform density than what we would have obtained by adding perturbations on all cubes. It also makes possible the calculations of the lower bounds in the next section.

3. Nonasymptotic Minimax Lower Bounds on Redundancy: the Continuous Case

In this section, we derive a minimax lower bound on the expected redundancy of any joint density (or prefix code) s over the continuous Lipschitz Markov class $\mathcal{M}_d$. We bound this minimax redundany from below by the minimax redundancy over the hypercube subclass $\mathcal{F}_{r,d}$ and then mimick the Assouad argument to obtain a lower bound involving two char-

acteristic quantities of the subclass $\mathcal{F}_{r,d}$. and the calculation of one of them is rather tricky in the Markov case.

For any joint density $s(x^n)$ on $J^d = [-1/2, 1/2]^n$,

$$\begin{aligned}
& \max_{f\in\mathcal{M}_d} \int_{J^n} f(x^n)\log(f(x^n)/s(x^n))dx^n \\
\geq\ & \max_{f\in\mathcal{F}_{r,d}} \int_{J^n} f(x^n)\log(f(x^n)/s(x^n))dx^n \\
=\ & \max_{f\in\mathcal{F}_{r,d}} \sum_t E_f \log(f(x_t|x^{t-1})/s(x_t|x^{t-1})) \\
\geq\ & 2^{-r}\sum_t\sum_\theta E_{f_\theta^{t-1}} \int_{-1/2}^{1/2} f_\theta(x_t|x^{t-1})\log(f_\theta(x_t|x^{t-1})/s(x_t|x^{t-1}))dx_t \\
\geq\ & 2^{-r}\sum_t\sum_\theta E_{f_\theta^{t-1}} \int_{-1/2}^{1/2} \{\sqrt{f_\theta(x_t|x^{t-1})} - \sqrt{s(x_t|x^{t-1})}\}^2 dx_t.
\end{aligned}$$

The validity of the last step is due to the following inequality (cf. Devroye, 1987, p. 16): for any two densities f and g, $\int f\log(f/g) \geq \int(\sqrt{f}-\sqrt{g})^2$. Also recall (1.2) and note that in this Markov dependent case, although the expected redundancy still equals the accumulated prediction error, but it is no longer the accumulated density estimation error.

For simplicity, we take out for analysis the tth term in the above expression, which is the prediction error in Helling distance (or the estimation error for the predictive density). We use the following notation $x_i^j := (x_i, x_{i+1}, ..., x_j)$, $x^j := x_1^j = (x_1, ..., x_j)$, and θ_{i+} and θ_{i-} are two vectors which differ only at the ith coordinate. The following arguments are essentially the same as Assouad's in the iid case except for the step where Lemma A has to be called for to deal with the dependence.

$$\begin{aligned}
& 2^{-r}\sum_\theta E_{f_\theta^{t-1}} \int_{-1/2}^{1/2} (\sqrt{f_\theta(x_t|x^{t-1})} - \sqrt{s(x_t|x^{t-1})})^2 dx_t \\
=\ & 2^{-r}\sum_\theta \int_{J^{t-1}} f_\theta(x^{t-1}) \int_{-1/2}^{1/2} (\sqrt{f_\theta(x_t|x^{t-1})} - \sqrt{s(x_t|x^{t-1})})^2 dx_t dx^{t-1} \\
=\ & 2^{-r}\sum_\theta \int_{J^{t-d}} \sum_{i=1}^{r} \int_{A_i^*} f_\theta(x^{t-1})\{\sqrt{f_\theta(x_t|x^{t-1})} - \sqrt{s(x_t|x^{t-1})}\}^2 dx_{t-d+1}^t dx^{t-d}
\end{aligned}$$

$$
\begin{aligned}
&= 2^{-r-1}\sum_{\theta}\int_{J^{t-d}}\sum_{i=1}^{r}[\int_{A_i^*} f_{\theta_{i+}}(x^{t-1})\{(\sqrt{f_{\theta_{i+}}(x_t|x^{t-1})}-\sqrt{s(x_t|x^{t-1})})^2 \\
&\quad +\int_{A_i^*} f_{\theta_{i-}}(x^{t-1})(\sqrt{f_{\theta_{k-}}(x_t|x^{t-1})}-\sqrt{s(x_t|x^{t-1})})^2\}]dx_{t-d+1}^t dx^{t-d} \\
&\geq 2^{-r-2}\sum_{\theta}\int_{J^{t-d}}\{\sum_{i=1}^{r}\int_{A_i^*}(\sqrt{f_{\theta_{i+}}(x_t|x^{t-1})}-\sqrt{f_{\theta_{i-}}(x_t|x^{t-1})})^2\times \\
&\quad \min(f_{\theta_{i+}}(x^{t-1}),f_{\theta_{i-}}(x^{t-1}))\}dx_{t-d+1}^t dx^{t-d} \\
&= 2^{-r-2}\cdot 2^d\sum_{\theta}\{\sum_{i=1}^{r}\int_{[-h/2,h/2]^d} q^2(x_{t-d+1}^t)dx_{t-d+1}^t\times \\
&\quad \int_{J^{t-d}}\min(f_{\theta_{i+}}(x^{t-d}),f_{\theta_{i-}}(x^{t-d}))dx^{t-d} \quad \text{by Lemma A in the Appendix} \\
&= 2^{-r-2}\cdot 2^d\cdot 2^r\cdot r\cdot\int_{[-h/2,h/2]^d} q^2(x_d)dx^d\times \\
&\quad \inf_{\theta,i}\int_{[-1/2,1/2]^{t-d}}\min(f_{\theta_{i+}}(x^{t-d}),f_{\theta_{i-}}(x^{t-d}))dx^{t-d} \\
&= 2^d\cdot(r/4)\cdot\int_{[-h/2,h/2]^d} q^2(x^d)dx^d\cdot\inf_{\theta,i}\int_{[-1/2,1/2]^{t-d}}\min(f_{\theta_{i+}}(x^{t-d}),f_{\theta_{i-}}(x^{t-d}))dx^{t-d}.
\end{aligned}
$$

Recall that $r = r_0^d\prod_{j=1}^{d-1}(1-j/r_0) = (2^d h^d)^{-1}\prod_{j=1}^{d-1}(1-j/r_0)$, and let $\alpha_h := \int_{[-h/2,h/2]^d} q^2(x^d)dx^d$ and assume for any θ and i, $\int_{J^{t-d}}\min(f_{\theta_{i+}}(x^t),f_{\theta_{i-}}(x^t))dx^t \geq \beta_h(t)$. Then the last expression is bounded below by $(4h^d)^{-1}\ \alpha_h\ \beta_h(t-d)\ \prod_{j=1}^{d-1}(1-j/r_0)$. To deal with β_h, we note that for any fixed θ and i

$$
\int_{[-1/2,1/2]^{t-d}}\min(f_{\theta_{i+}}^{t-d},f_{\theta_{i-}}^{t-d})dx^{t-d}\geq 1-\sqrt{2-2\int_{-1/2,1/2]^{t-d}}\sqrt{f_{\theta_{i+}}^{t-d}f_{\theta_{i-}}^{t-d}}dx^{t-d}}. \qquad (0.1)
$$

Define $Q_t := \int_{[-1/2,1/2]^t}\sqrt{f_{\theta_{i+}}^t f_{\theta_{i-}}^t}dx^t = \int_{[-1/2,1/2]^t}\sqrt{f_{\theta_{i+}}(x^t)f_{\theta_{i-}}(x^t)}dx^t$. Then $Q_t\equiv 1$ for $t=1,2,...,d-1$, and for $t\geq d$,

$$
\begin{aligned}
Q_t &= \int_{[-1/2,1/2]^t}\sqrt{\prod_{j=d}^{t}(f_{\theta_{i+}}(x_t|x_{t-d+1}^{t-1})f_{\theta_{i-}}(x_t|x_{t-d+1}^{t-1})}dx^t \ \text{(by (1) and (2))} \\
&= \int_{[-1/2,1/2]^t}\sqrt{\prod_{j=d}^{t}(f_{\theta_{i+}}(x_{t-d+1},...,x_t)f_{\theta_{i-}}(x_{t-d+1},...,x_t)}dx^t, \ \text{(by (3))}
\end{aligned}
$$

because (1) the sequences corresponding to $f_{\theta_{i+}}$ and $f_{\theta_{i-}}$ are $(d-1)$st order Markov, (2) by our construction of f_θs $f_{\theta_{i+}}(x_1,...,x_{d-1})\equiv f_{\theta_{i-}}(x_1,...,x_{d-1})\equiv 1$, and hence (3) $f_{\theta_{i+}}(x_t|x_{t-d+1}^{t-1})\equiv f_{\theta_{i+}}(x_{t-d+1},...,x_t)$, $f_{\theta_{i-}}(x_t|x_{t-d+1}^{t-1})\equiv f_{\theta_{i-}}(x_{t-d+1},...,x_t)$.

Without loss of generality, for any $A_i \in R$ we may assume $A_i = I_d \times I_{d-1} \times ... \times I_1$. $A_i = I_{i_1} \times I_{i_2} \times ... \times I_{i_d}$. Observe that $f_{\theta_{i+}}(x^d) \equiv f_{\theta_{i-}}(x^d)$ on $A_i^{*c} := (I_d^* \times I_{d-1}^* \times ... \times I_1^*)^c$ and $(f_{\theta_{i+}}(x^d) - 1) = -(f_{\theta_{i-}}(x^d) - 1)$ on A_i^*. For simplicity denote $u(x^d) := \sqrt{f_{\theta_{i+}}(x^d) f_{\theta_{i-}}(x^d)}$. The next few lemmas calculate α_h and β_h. Their proofs can be found in the appendix.

Lemma 1 (i) $\alpha_h = \int_{[-h/2,h/2]^d} q^2(x^d)dx^d = c_d^2 h^{d+2}/(6d)$.

(ii) $Q_d = \int_{[-1/2,1/2]^d} u(x^d)dx^d = 1 - c_h$, where $c_h := (2h)^d - \int_{A_i^*} u(x^d)dx^d \le 2^d c_d^2 h^{d+2}/(6d)$.

(iii) $\int_{[-1/2,1/2]^{t-d}} \int_{I_d^*} ... \int_{I_1^*} u(x^t)dx_t...dx_1 = Q_d \cdot Q_{t-d}$.

Lemma 2 For $t \ge d$, $Q_t = Q_{t-1} - c_h Q_{t-d}$. .

Lemma 3 For $t \ge d$, $Q_t \ge 1 - (t - d + 1)c_h$.

For $t \ge 2d$, by (0.1) and Lemma 3,

$$\int_{[-1/2,1/2]^t} \min(f_{\theta_{i+}}^t, f_{\theta_{i-}}^t) \ge 1 - c_d(2h)^{1+d/2}\sqrt{(t-d+1)}/\sqrt{6d}.$$

Hence

Lemma 4 For $t \ge 2d$, $\beta_h(t) \ge 1 - c_d(2h)^{1+d/2}\sqrt{(t-d+1)}/\sqrt{6d}$. □

Plugging Lemmas 1(i) and 4 back into (3.1), we obtain the following minimax lower bound on the per symbol redundancy over Markov sources induced by $\mathcal{F}_{r,d}$ and hence over $\mathcal{M}_d$:

$$n^{-1} \sum_{t=2d}^{n} \frac{4^{-1}h^{-d}c_d^2 h^{d+2}}{4d} (1 - c_d(2h)^{1+d/2}\sqrt{(t-2d+1)}/\sqrt{6d}) \prod_{j=1}^{d-1}(1 - j/r_0)$$

$$\ge 4^{-1}c_d^2 h^2/(4d)\,(1 - 2d/n - (2/3)\sqrt{n}c_d(2h)^{1+d/2}/\sqrt{6d}) \prod_{j=1}^{d-1}(1 - j/r_0) \qquad (0.2)$$

where $r_0 = (2h)^{-1}$. Choosing h such that $(2/3)\sqrt{n}c_d(2h)^{1+d/2}/\sqrt{6d} = 1/2$, the redundancy lower bound is

$$c_d^2/(48d)\, n^{-2/(d+2)}\, (1 - 4d/n) \prod_{j=1}^{d-1}(1 - j/r_0),$$

where $r_0 = 18 \cdot (c_d^2/d)^{1/(d+2)} \cdot n^{1/(d+2)}$.

Similar arguments give a minmax lower bound on the Kullback-Leibler density estimation error for the joint density of d-tuples over the class $\mathcal{M}_d$. Hence we have

Theorem 1 There is an $n_0 > 0$ such that for $n > n_0$, any prefix code s on x^n and any density estimator g based on x^n,

$$\max_{f\in\mathcal{M}_d} \int_{[-1/2,1/2]^n} f(x^n)\log(f(x^n)/s(x^n))dx^n \geq A_d \cdot (1-4d/n)n^{-2/(d+2)}, \text{ and}$$

$$\max_{f\in\mathcal{M}_d} \int_{[-1/2,1/2]^n} f(x^n) \int_{[-1/2,1/2]^d} f(y^d)\log(f(y^d)/g(y^d|x^n))dy^d dx^n \geq (A_d/2)\cdot n^{-2/(d+2)},$$

where $A_d := c_d^2/(32d)\prod_{j=1}^{d-1}(1-j/r_0)$ and $r_0 = ((8/27)c_d^2/d)^{1/(d+2)} \cdot n^{1/(d+2)}$. □

Remark: We may choose n_0 to be the smallest n larger than $4d$ such that $c_d h/2 = [(27/8)d/(c_d^2)n^{-1}]^{1/(d+2)} < 1$. It is believed that the rates in the theorem are optimal and can be achieved by a histogram estimator g and a code s based on histogram predictive densities, cf. Yu and Speed (1992).

If our class is so large to include Markov sources of any order, it would also include $\mathcal{M}_d$ with $d = d_n := \log n$. Suppose that the Lipschitz constant c_d is independent of n. Note that $\prod_{j=1}^{d-1}(1-j/r_0) \geq (1-d_n r_0)^{d_n} = (1-2h_n d_n)^{d_n}$ and $\lim_{n\to\infty} n^{-2/(d_n+2)} > 0$. In Theorem 1 we take $h_n = 1/d_n^2 = 1/(\log n)^2$ to obtain:

Corollary 1 For $d_n = \log n$, and $c_{d_n} \equiv c_1 < \infty$, $\exists\ c_0 > 0$, and $n_0 > 0$, for any $n > n_0$, any code s on x^n and any density estimator g

$$\max_{f\in\mathcal{M}_{d_n}} \int_{[-1/2,1/2]^n} f(x^n)\log(f(x^n)/s(x^n))dx^n \geq c_0(\log n)^{-5}, \quad \text{and}$$

$$\max_{f\in\mathcal{M}_{d_n}} \int_{[-1/2,1/2]^n} f(x^n) \int_{[-1/2,1/2]^d} f(y^d)\log(f(y^d)/g(y^d|x^n))dy^d dx^n \geq (c_0/2)(\log n)^{-5}. \ \square$$

The above result says that there is no universal redundancy or density estimation rate faster than $(\log n)^{-5}$ over the class $\mathcal{M}_{d_n}$. We now change the classes $\mathcal{M}_{d_n}$ and $\mathcal{F}_{r,d_n}$ slightly by relaxing the finiteness of c_d to $c_d = c_{d_n} := (\log n)^{5/2}$ and replace q by $\min(q, a_0)$ where

$a_0 < 1$ in order to avoid f's in $\mathcal{F}_{r,d_n}$ being negative. Note that $c_{d_n}^2 h_n^2 = d_n$ and

$$\begin{aligned}\alpha_h &= \int_{[-h/2,h/2]^d} \min(q^2, a_0^2) dx^d = (2/3) h_n^d c_{d_n}^2/(4d_n) h_n^2 [1 - (1 - 4a_0^2/(c_{d_n}^2 h_n^2))^{d_n}] \\ &= (2/3) h_n^d c_{d_n}^2/(4d_n) h_n^2 [1 - (1 - 4a_0^2/d_n)^{d_n}] \approx h_n^d [1 - e^{-4a_0^2}].\end{aligned}$$

Similar arguments as used to obtain Theorem 1 give

Corollary 2 For $d_n = \log n$, and $c_{d_n} = (\log n)^{5/2}$, $\exists\ c_0 > 0$, and $n_0 > 0$, for any $n > n_0$, any prefix code s on x^n and any density estimator g.

$$\max_{f \in \mathcal{M}_{d_n, c_{d_n}}} \int_{[-1/2,1/2]^n} f(x^n) \log(f(x^n)/s(x^n)) dx^n \geq c_0, \quad \text{and}$$

$$\max_{f \in \mathcal{M}_{d_n, c_{d_n}}} \int_{[-1/2,1/2]^n} f(x^n) \int_{[-1/2,1/2]^d} f(y^d) \log(f(y^d)/g(y^d|x^n)) dy^d dx^n \geq c_0/2. \ \square$$

4. Discussion

All the Markov sources corresponding to densities in class $\mathcal{M}_{d_n}$ with $c_{d_n} = (\log n)^{5/2}$ have bounded transition kernels and hence are ϕ-mixing, Doob (pp. 215, 1953). Thus the class $\mathcal{M} := \cup_{d=1}^{\infty} \mathcal{M}_d$ is a smooth subset of continuous Markov sources of all orders. Not only the ergodic theorem holds for this class, but also the CLT. Therefore we reach a similar conclusion in a nonparametric setting as by Shields (1991) in the finite alphabet parametric setting. (In fact, the same technique can be used to obtain a similar result even in the finite alphabet case, cf. Yu, 1992). Our proof is more direct and the explicit construction of the hypercube subclass $\mathcal{F}_{r,d}$ provides something more concrete than a proof of non-existence of the universal redundancy rate. For fixed r, it is those well-separated elements in $\mathcal{F}_{r,d}$ that make the universal coding task so hard: no code can be good simultaneously for all of them. For any fixed sample size n, r has to be chosen accordingly to make sure that the complexity of $\mathcal{F}_{r,d}$ resembles $\mathcal{M}_d$'s.

Corollaries 1 and 2 also illustrate a well-known fact or drawback of minimax results, that is, although in many nonparametric estimation problems, minimax lower bounds are the only existing criteria against which we may compare estimators, they depend heavily on the classes chosen a priori. If we may assume that researchers agree that the smoothness conditions on the density in iid case are reasonable, then how about the dependent case? Which of the two classes in Corollaries 1 and 2 is more appropriate? Should we relax the uniform boundedness condition on the c_{d_n}? Intuitively, the more homogeneous the class, the more meaningful the minimax result. In other words, if all densities in that class are close to be equally likely, the worst case becomes an average case. In fact, the minimax redundancy (or density estimation error) was bounded below by the average redundancy over a homogeneous class $\mathcal{F}_{r,d}$ or the Bayes redundancy with the uniform prior on $\mathcal{F}_{r,d}$. The other hope is to rely on the consensus on smoothness conditions. In the case of a Gaussian process, dependence measured by a certain (β) mixing condition is equivalent to smoothness conditions of the spectral density. The smoother the spectral density, the less dependent the process. And this connection is true in general at an intuitive level. Hence we may compare the two classes in terms of the smoothness of the their spectral densities. Roughly speaking, for the class $\mathcal{F}_{r,d}$, the smaller c_d is, the closer the density is to the uniform density so the Markov sequences generated by the densities should be closer to the uniform independent sequence. How much more dependence is introduced when we increase c_d to $(\log n)^{5/2}$? Will this extra dependence alter the smoothness of the spectral density? It turns out, however, that it is no easy task to calculate the mixing coefficient even in the Markov case and even when the explicit formula of the transition kernel is known. Further investigation is needed.

Appendix

Proof of Lemma 1: (i) Observe that the surface of q^2 is a curved (inward) pyramid with

the height $c_d^2 h^2/4$ and we may write any point on the facets of our hypercube $[-h/2.h/2]^d$ in terms of the polar coordinate system as (ϕ, l_ϕ) for some $\phi \in \Phi$. (Note that $\Phi = (0, 2\pi)$ when $d = 2$ and ϕ is a vector for $d > 2$.) Then by Fubini's theorem

$$\begin{aligned} V_q &:= \text{volume of the curved pyramid} \\ &= \int_{[-h/2,h/2]^d} q^2(x^d)dx^d = \int_\Phi \int_0^{l_\phi} b_\phi z^2 dz d\phi = (1/3)\int_\Phi b_\phi l_\phi^3 d\phi; \\ V_p &:= \text{volume of the pyramid sitting on} [-h/2, h/2]^d \text{ with the height } c_d^2 h^2/4 \\ &= \int_\Phi \int_0^{l_\phi} b'_\phi z dz d\phi = (1/2)\int_\Phi b'_\phi l_\phi^2 d\phi, \end{aligned}$$

where, to match the height of the pyramid and the curved pyramid, $b_\phi l_\phi^2 = b'_\phi l_\phi$. Thus

$$V_p = (1/2)\int_\Phi b'_\phi l_\phi^2 d\phi \text{ and } V_q = (1/2)\int_\Phi b_\phi l_\phi^3 d\phi,$$

which gives, using $V_p = c_d^2 h^{d+2}/(4d)$, $V_q = (2/3)V_p = c_d^2 h^{d+2}/(6d)$.

$$\text{(ii)} \quad c_h = (2h)^d - \int_{A^*} u(x^d)dx^d = (2h)^d - 2^d \int_A u(x^d)dx^d.$$

On A_i^*, $u(x^d) = \sqrt{1-q^2} \geq 1 - q^2$. Hence

$$c_h \leq (2h)^d - 2^d \int_{[-h/2,h/2]^d} (1 - q^2(x^d))dx^d = 2^d \int_{[-h/2,h/2]^d} q^2(x^d)dx^d.$$

(ii) is proved by plugging (i) into the last expression.

(iii) For $1 \leq k \leq d$, let $v_k(x_{t-d}, ..., x_{t-k}) := \int_{I_k^*} ... \int_{I_1^*} u(x_{t-d+1}, ..., x_t)dx_t ... dx_{t-k+1}$. Obviously, $v_d = Q_d$. Since $u(x_{t-d+1}, ..., -x_{t-k}, ..., -x_t) = u(x_{t-d+1}, ..., x_{t-k}, ..., x_t)$ on $I_k^* \times ... \times I_1^*$, integrating out $x_t, ..., x_{t-k+1}$ and using symmetry of the integration domain $I_k^* \times ... \times I_1^*$, we obtain that, for any fixed $(x_{t-d}, ..., x_{t-k-1})$, v_k is an even function of x_{t-k} on I_{k+1}^*, namely, $v_k(x_{t-d}, ..., -x_{t-k}) = v_k(x_{t-d}, ..., x_{t-k})$. On the other hand, $e(x_{t-d-k+1}, ..., x_{t-k}) := u(x_{t-d-k+1}, ..., x_{t-k}) - 1 = f_{\theta_{i+}} - 1$ is an odd function of x_{t-k} on I_{k+1}^*, because $x_{t-k} \in I_{k+1}^*$

so $(x_{t-d-k+1},...,x_{t-k}) \notin A_i = I_d^* \times ... \times I_1^*$. Therefore, $v_k e$ is an odd function of x_{t-k} on I_{k+1}^*, and

$$\begin{aligned} & \int_{I_{k+1}^*} v_k(x_{t-d},...,x_{t-k}) u(x_{t-d-k+1},...,x_{t-k}) dx_{t-k} \\ = & \int_{I_{k+1}^*} v_k(x_{t-d},...,x_{t-k})[1 + e(x_{t-d-k+1},...,x_{t-k})] dx_{t-k} \\ = & \int_{I_{k+1}^*} v_k(x_{t-d},...,x_{t-k}) dx_{t-k} = v_{k+1}. \end{aligned}$$

Finally,

$$\begin{aligned} & \int_{[-1/2,1/2]^{t-d}} \int_{I_d^*} ... \int_{I_2^*} \int_{I_1^*} u(x_{t-d+1},...,x_t) dx_t u(x_{t-d},...x_{t-1}) dx_{t-1} ... dx^{t-d} \\ = & \int_{[-1/2,1/2]^{t-d}} \int_{I_d^*} ... \int_{I_2^*} v_1(x_{t-d},...,x_{t-1}) u(x_{t-d},...x_{t-1}) dx_{t-1} ... dx^{t-d} \\ = & ... \\ = & \int_{[-1/2,1/2]^{t-d}} \int_{I_d^*} v_{d-1}(x_{t-d+1}) u(x_{t-2d+1},...x_{t-d+1}) dx_{t-d+1} ... \\ & v_{d-1}(x_{t-d+1}).,x_{t-1}) u(x_{t-d},...x_{t-1}) dx_{t-1} ... dx^{t-d} \\ = & v_d \int_{[-1/2,1/2]^{t-d}} u(x^{t-d}) dx^{t-d} = Q_d \cdot Q_{t-d}. \quad \square \end{aligned}$$

Similarly to the proof of (iii) above, we can prove the following lemma.

Lemma A

$$\begin{aligned} & \int_{J^{t-d}} \int_{A_i^*} (\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} - (\sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 \min(f_{\theta_{i+}}(x^{t-1}), f_{\theta_{i-}}(x^{t-1})) dx_{t-d+1}^t dx_{t-d} \\ = & \int_{A_i^*} (\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} - (\sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 dx_{t-d+1}^t \int_{J^{t-d}} \times \\ & \min(f_{\theta_{i+}}(x^{t-d}), f_{\theta_{i-}}(x^{t-d})) dx_{t-d} \qquad \text{(A.1)} \\ \geq & 2^d \cdot \int_{[-h/2,h/2]^d} q^2(x^d) dx^d \cdot \int_{J^{t-d}} \min(f_{\theta_{i+}}(x^{t-d}), f_{\theta_{i-}}(x^{t-d})) dx_{t-d} \qquad \text{(A.2)} \end{aligned}$$

Proof: Note that for $x_{t-d+1}^{t-1} \in I_d^* \times ... \times I_2^*$, $x_{t-d+1-j}^{t-j} \notin A_i^* = I_d^* \times ... \times I_1^*$, $j = 1,...,d-1$. Hence

$$f_{\theta_{i+}}(x^{t-1}) = f_{\theta_{i+}}(x_{t-d},...,x_{t-1}) \times f_{\theta_{i+}}(x_{t-d-1},...,x_{t-2}) \times ... \times$$

$$
\begin{aligned}
& f_{\theta_{i+}}(x_{t-2d},...,x_{t-d+1}) \times f_{\theta_{i+}}(x^{t-d}) \\
:= \;& u(x_{t-d},...,x_{t-1}) \times u(x_{t-d-1},...,x_{t-2}) \times ... \times u(x_{t-2d},...,x_{t-d+1}) \times f_{\theta_{i+}}(x^{t-d}); \\
f_{\theta_{i-}}(x^{t-1}) \;=\; & f_{\theta_{i-}}(x_{t-d},...,x_{t-1}) \times f_{\theta_{i-}}(x_{t-d-1},...,x_{t-2}) \times ... \times \\
& f_{\theta_{i-}}(x_{t-2d},...,x_{t-d+1}) \times f_{\theta_{i-}}(x^{t-d}) \\
:= \;& u(x_{t-d},...,x_{t-1}) \times u(x_{t-d-1},...,x_{t-2}) \times ... \times u(x_{t-2d},...,x_{t-d+1}) \times f_{\theta_{i-}}(x^{t-d}).
\end{aligned}
$$

$$
\begin{aligned}
& \int_{J^{t-d}} \int_{A_i^*} (\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} - (\sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 \times \min(f_{\theta_{i+}}(x^{t-1}), f_{\theta_{i-}}(x^{t-1}))dx_{t-d+1}^t dx_{t-d} \\
= \;& \int_{J^{t-d}} \int_{A_i^*} (\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} - (\sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 \times u(x_{t-d},...,x_{t-1}) \times \\
& u(x_{t-d-1},...,x_{t-2}) \times ... \times u(x_{t-2d},...,x_{t-d+1}) \times \min(f_{\theta_{i+}}(x^{t-d}), f_{\theta_{i-}}(x^{t-d}))dx_t
\end{aligned}
$$

which, by a similar argument as in proving Lemma 1 (iii), equals the right hand side of $(A.1)$. $(A.1)$ is straightforward after noting that on A_i

$$
(\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} - \sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 = \frac{(f_{\theta_{i+}}(x_{t-d+1}^t) - f_{\theta_{i-}}(x_{t-d+1}^t))^2}{\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} + \sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2},
$$

$$
\begin{aligned}
& (\sqrt{f_{\theta_{i+}}(x_{t-d+1}^t)} + (\sqrt{f_{\theta_{i-}}(x_{t-d+1}^t)})^2 = f_{\theta_{i+}} + f_{\theta_{i-}} + 2\sqrt{f_{\theta_{i+}} f_{\theta_{i-}}} \\
= \;& 1 + q_i + 1 - q_i + 2\sqrt{1 - q_i^2} \leq 4, \text{ and} (f_{\theta_{i+}} - f_{\theta_{i-}})^2 = 4q_i^2. \quad \square
\end{aligned}
$$

Proof of Lemma 2: For any $1 \leq k < d$, define

$$
\begin{aligned}
T_k(t) := & \int_{[-1/2,1/2]^{t-d}} \{ \int_{I_{d-k}^*} ...[\int_{I_1{}^*} u(x_{t-k-d+1},...,x_{t-k})dx_{t-k}] \times ... \times \\
& u(x_{t-2d+2},...,x_{t-d+1})dx_{t-d+1}\} u(x_{t-2d+1},...,x_{t-d})dx_{t-d} \times ... \times u(x_1,...,x_d)dx_d dx_{d-1}...dx_1.
\end{aligned}
$$

Moreover, let $B_k := \{(i_1, i_2, ..., i_d) \in R : i_{k+1} = d-k, ..., i_d = 1\}$, then $B_0 = \{(d, d-1, ..., 1)\}$, $|B_0| = 1$, $B_1 = \{(j, d-1, ..., 1) : j = 1, 2, ..., r_0\}$, $|B_1| = r_0$, and $B_0 \cap B_1 = \emptyset$. In general, $|B_k| = r_0^k$ and $B_i \cap B_j = \emptyset$ for $i \neq j$.

For any $(i_1, i_2, ..., i_d) \notin \cup_{k=0}^{d-1} B_k$, $f_{\theta_{i+}} \equiv f_{\theta_{i-}}$ on $I_{i_1}^* \times ... \times I_{i_d}^*$, by symmetry of $f's$ in $\mathcal{F}_{r,d}$,

$$\int_{J^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 = (2h) \int_{J^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_{d-1}}^*} u(x^{t-1}) dx_{t-1} ... dx_1$$

$$= \quad ... = (2h)^d \cdot \int_{[-1/2,1/2]^{t-d}} u(x^{t-d}) dx^{t-d} = (2h)^d \cdot Q_{t-d}.$$

For any $(i_1, i_2, ..., i_d) \in B_k$ $(k \geq 1)$, similar arguments give

$$\int_{[-1/2,1/2]^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 = (2h)^k \cdot T_k(t).$$

If $(i_1, i_2, ..., i_d) \in B_0$, $(i_1, i_2, ..., i_d) = (d, d-1, ..., 1)$. Then by Lemma 1 (i) and (ii),

$$\int_{[-1/2,1/2]^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 = Q_d \cdot Q_{t-d} = ((2h)^d - c_h) Q_{t-d}.$$

On the other hand,

$$\begin{aligned}
Q_t &= \sum_{(i_1,...,i_d) \notin \cup_{k=0}^{d-1} B_k} \int_{[-1/2,1/2]^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 \\
&\quad + \sum_{k=0}^{d-1} \sum_{(i_1,...,i_d) \in B_k} \int_{[-1/2,1/2]^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 \\
&= (r_0^d - \sum_{k=0}^{d-1} |B_k|)(2h)^d Q_{t-d} + [(2h)^d - c_h) Q_{t-d} \\
&\quad + \sum_{k=1}^{d-1} \sum_{(i_1,...,i_d) \in B_k} \int_{[-1/2,1/2]^{t-d}} \int_{I_{i_1}^*} ... \int_{I_{i_d}^*} u(x^t) dx_t dx_{t-1} ... dx_1 \\
&= (r_0^d - r_0^{d-1} - ... - r_0)(2h)^d Q_{t-d} - c_h Q_{t-d} + \sum_{k=1}^{d-1} r_0^k (2h)^k T_k(t) \\
&= (r_0^d - r_0^{d-1} - ... - r_0)(2h)^d Q_{t-d} - c_h Q_{t-d} + \sum_{k=1}^{d-1} T_k(t).
\end{aligned}$$

That is,

$$Q_t = (r_0^d - r_0^{d-1} - ... - r_0)(2h)^d Q_{t-d} - c_h Q_{t-d} + \sum_{k=1}^{d-1} T_k(t). \tag{A.3}$$

Similarly, $T_1(t) = Q_{t-1} - (r_0^{d-1} - r_0^{d-2} - ... - 1)(2h)^{d-1} Q_{t-d} - T_2(t) - T_3(t) - ... - T_{d-1}(t)$.

Hence $\sum_{k=1}^{d-1} T_k(t) = Q_{t-1} - (r_0^d - r_0^{d-1} - ... - r_0)(2h)^d Q_{t-d}$, which, combined with (A.3),

gives $Q_t = Q_{t-1} - c_h Q_{t-d}$. □

Proof of Lemma 3:

$$
\begin{aligned}
Q_t &= Q_{t-1} - c_h Q_{t-d} \\
&= Q_{t-2} - c_h Q_{t-d-1} - c_h Q_{t-d} \text{ since } Q_{t-1} = Q_{t-2} - c_h Q_{t-d-1} \\
&= Q_{t-3} - c_h Q_{t-d-2} - c_h Q_{t-d-1} - c_h Q_{t-d} \text{ since } Q_{t-2} = Q_{t-3} - c_h Q_{t-d-2} \\
&= ... \\
&= (1-c_h)Q_{t-d} - \sum_{j=1}^{d-1} Q_{t-d-j} \\
&= (1-c_h)Q_{t-d-1} - (1-c_h)c_h Q_{t-2d} - c_h(Q_{t-d-1} + Q_{t-d-2} + ... + Q_{t-2d+1}) \\
&= (1-2c_h)Q_{t-d-1} - c_h(Q_{t-d-2} + Q_{t-d-3} + ... + Q_{t-2d}) + c_h^d Q_{t-2d} \\
&\geq (1-2c_h)Q_{t-d-1} - c_h(Q_{t-d-2} + Q_{t-d-3} + ... + Q_{t-2d}) \\
&= (1-3c_h)Q_{t-d-2} - c_h(Q_{t-d-3} + Q_{t-d-4} + ... + Q_{t-2d-1}) + 2c_h^d Q_{t-2d-1} \\
&\geq (1-3c_h)Q_{t-d-2} - c_h(Q_{t-d-3} + Q_{t-d-4} + ... + Q_{t-2d-1}) \\
&\geq ... \\
&\geq [1-(t-2d+1)c_h]Q_d - c_h(Q_{d-1} + Q_{d-2} + ... + Q_1) \\
&= [1-(t-2d+1)c_h)(1-c_h) - (d-1)c_h \\
&\geq 1-(t-2d+1+1+d-1)c_h = 1-(t-d+1)c_h. \quad \square
\end{aligned}
$$

References

[1] Assouad, P. (1983). Deux remarques sur l'estimation. *Comptes Rendus de l'Academie des Sciences de Paris.* **296** 1021-1024.

[2] Birgé, L. (1985). Non-asymptotic minimax risk for Hellinger balls. *Probability and Mathematical Statistics* **5** 21-29.

[3] Clarke, B. S. and Barron, A. R. (1990). Information theoretic asymptotics of Bayes methods. *IEEE Trans. Infor. Th.* **36** 453-471.

[4] Devroye, L. (1983). On arbitrary slow rates of global convergence in density estimation. *Z. Wahrsch. Verw. Gebiete* **62** 475-483.

[5] Devroye, L. (1987). *A course in density estimation.* Progress in probability and statistics **14** Birkhauser.

[6] Donoho, D. L., Liu, R. C., and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416-1437.

[7] Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

[8] Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14** 1080-1100.

[9] Shields, P. C. (1991). Universal redundancy rates don't exist. Preprint.

[10] Yu, B. and Speed, T. P. (1992). Data compression and histograms. *Probab. Th. Rel. Fields* **2** 195-229.

[11] Yu, B. (1992). A simple proof of the nonexistence of universal redundancy rates. (in preparation)