Stochastic Complexity and Model Selection II. Histograms.

By

Bin Yu and T.P. Speed

Department of Statistics
University of California, Berkeley, CA 94720

Technical Report No. 241
March 1990

Department of Statistics
University of California
Berkeley, California

Stochastic complexity and model selection II. Histograms

by
T.P. Speed and Bin Yu
Department of Statistics
University of California, Berkeley

1. Introduction

Dawid (1984) introduced his prequential approach to statistics, arguing that the purpose of statistics is to make sequential probability forecasts for future observations. This led him to selection procedures which compare different models on the basis of their accumulated prediction error, which for a given model is measured by

$$\sum_{t=1}^{n} -\log g_{t-1}\left(x_t \mid x^{t-1}\right) \tag{1.1}$$

where $x^n = (x_1, \ldots, x_n)$, and $g_{t-1}\left(x_t \mid x^{t-1}\right)$ is a fully specified predictive density for x_t, given by some fixed procedure using x^{t-1} and the model. From a quite different perspective Rissanen (1986), see Rissanen (1989) for a comprehensive discussion and references to earlier work, studied in the same sum (1.1), viewing it as the length of a predictive code for x^n corresponding to the distribution g, where

$$g(x^n) = \prod_{i=1}^{n} g_{t-1}\left(x_t \mid x^{t-1}\right). \tag{1.2}$$

Key notions in Rissanen's approach to statistics are the description length, the predictive description length, which is just another name for (1.1), and the stochastic complexity of a set of data x^n relative to a class of probability models. In terms of each of these lengths, there is an associated model selection procedure, which we call MDL (minimum description length), PMDL (minimum predictive description length) and MSC (minimum stochastic complexity).

In the case where all the models are smoothly finitely parametrized, Rissanen (1986) has shown that all these model selection criteria give, on the average,

asymptotically equivalent results. Indeed in this case they are all asymptotically equivalent to the criterion now termed BIC. The almost sure equivalence of these criteria is also likely to be true, although it has only been proved in fully certain special cases, see Speed and Yu (1989) and references therein.

It is not at all clear that the prequential or PMDL approach, and the MDL and MSC criteria will lead to asymptotically equivalent results in the case of infinite-dimensional (also called nonparametric) models, either on average, or almost surely. One of the aims of this paper is to explore this topic with a simple infinite dimensional model class and its simplest finite-dimensional approximations: smooth densities on the unit interval, and histograms with equal bin-widths. This study was begun by Hall and Hannan (1988), and continued in Rissanen *et al.* (1989), and a number of our results are extensions or refinements of results found in these references. In particular, we give expansions of (1.1) valid a.s. and in expectation, when the elements of x^n are i.i.d. with a density on $[0,1]$ satisfying some standard regularity conditions, and g is based upon histograms with both a fixed and a time-varying number of bins. In a former case, it turns out that the various model selection procedures mentioned above can be regarded as equivalent, but this equivalence breaks down if the number of bins is permitted to vary with t.

As well as studying the behavior of the model selection criteria mentioned above, Rissanen (1986) derived an interesting lower bound for the code length achievable using finite-dimensional parametric families, when the data is generated by a member of one of these families. Specialized to the case of i.i.d. random variables, the bound involves a sequence of probability models $\{f_{k,\theta} : \theta \in \Theta_k\}$, $k = 1, 2, \ldots$, where Θ_k is a compact subset of $\mathbb{R}^k$ with non-empty interior, and we assume that the densities $f_{k,\theta}$ satisfy certain standard regularity conditions. It states that for all distributions g of x^n and for all $k \geq 1$, there is a subset A_g of Θ_k with Lebesgue measure zero, such that for $\theta \notin A_g$,

$$\liminf_{n\to\infty} \frac{E_{k,\theta} \log \left[\frac{f^n_{k,\theta}(x^n)}{g(x^n)} \right]}{\frac{1}{2} k \log n} \geq 1 . \tag{1.3}$$

Here $f^n_{k,\theta}$ denotes n-dimensional product of the density $f_{k,\theta}$, and the expectation is taken with respect to this density. Because of the well-known equivalence between prefix codes and probability distributions, see Rissanen (1989), (1.3) may be paraphrased as follows: without knowing the true source distribution $f^n_{k,\theta}$, we have to use (asymptotically) an extra $n^{-1}\,\frac{1}{2}\,k\,\log n$ bits per symbol, to encode a sequence x^n generated by $f^n_{k,\theta}$, no matter what prefix code we use. This rate $\frac{1}{2}\,k\,\log n$ depends critically on the assumption that the true source distribution belongs to a smooth finitely parametrized family, and the second main topic of this paper is to seek lower bounds of the form (1.3) in the infinite-dimensional case. One such has been established in Rissanen *et al.* (1990), and in §3 below we establish a minimax lower bound for the redundancy, in the spirit of Davisson (1983). A minimax analog of (1.3) above in the finite-dimensional case would refer to the class G_n of all densities of x^n, and have the form

$$\liminf_{n\to\infty} \frac{\min_{g\in G_n}\,\max_{\theta\in\Theta_k}\,E_{k,\theta}\,\log\left[\frac{f^n_{k,\theta}(x^n)}{g(x^n)}\right]}{\frac{1}{2}\,k\log n} \geq 1\,. \tag{1.4}$$

This is readily proved (in the i.i.d. case) from results in Clarke (1989). Our minimax lower bound in the infinite-dimensional case has the same general form as (1.4), but with $\frac{1}{2}\,k\log n$ replaced by $n^{-\frac{2}{3}}$.

We turn now to a brief description of our results, expressed in the terminology of coding theory, see Hamming (1986) and Rissanen (1989) for this background. For the most part, it is straightforward to translate back to the prequential statistics terminology of Dawid (1984, 1989).

In §2 below we give an expression of the form (1.1), which defines the length of a predictive code for x^n based upon a histogram with m equal-width bins. We need to modify the obvious expression slightly, and it turns out that the natural modification coincides with expression (2.3) of Hall and Hannan (1988). Thus the asymptotic expansions we derive parallel theirs, although we obtain ones valid a.s. and in expectation whereas theirs were only established in probability. We also interpret the expression obtained as a two-part code, thus providing a natural link between

the infinite and the finite dimensional cases. Finally, we give some asymptotic results concerning the data-determined number $\hat{m}_n$ of bins selected by our criterion, and establish optimality and consistency results which are quite analogous to those which hold with finite-dimensional models and for histograms with other selection criteria, see Stone (1985).

The subject of minimax lower bounds for expressions of the form (1.1) is addressed in §3, where a bound is derived by modifying familiar arguments from density estimation, see e.g. Devroye (1987). For the class F of boundedly differentiable densities which is used in most of our discussion, the redundancy can decrease at a rate no faster than $n^{-\frac{2}{3}}$, although with further smoothness assumptions this can be increased to $n^{-\frac{2s}{2s+1}}$.

We then turn to the construction of codes, equivalently probability densities, which achieve these lower bounds, or at least achieve the same rate of decrease of the redundancy. Here it becomes clear that the codes (densities) discussed by Hall and Hannan (1988) and Dawid (1989) do not achieve the lower bound rate, basically because they are insufficiently adaptive. The code described in Rissanen *et al.* (1989) was shown there to achieve the lower bound rate, on the average, and in §4 we describe a modification of the earlier code which a.s. achieves the same rate.

The proofs of results in §2 and §4 are deferred to §5 and §6 respectively, and an Appendix collects some results on so-called Poissonization, necessary for establishing the a.s. approximations required for our theorems. Unfortunately the proofs are all rather lengthy, but we have been unable to simplify them appreciably.

2. The selection rule

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with density f on $[0,1]$. For any fixed integer m, write $I_{k,m} = [(k-1)/m, k/m]$ and denote by H_m the family

of histograms with m equal-width bins, i.e.,

$$H_m = \left\{ h : h = \sum_{k=1}^{m} c_k 1_{I_{k,m}}, \sum c_k = m, c_k \geq 0 \right\}.$$

Moreover, let $N_{k,m}(t) = \sum_{s=1}^{t} 1_{I_{k,m}}(x_s)$ be the counts of those x_s falling into the kth bin $I_{k,m}$ in the first t observations.

For each m we will take the histogram with m bins based on x^t as the predictive density g_t, and construct a density, and hence a prefix code, on the n-tuples. However, we must modify the naive histogram estimator to avoid the problem created by the fact that the predictive density will take value zero on some of the bins at the beginning of the encoding. We add to the beginning of our n-tuple x^n m numbers $y_1, \ldots, y_m$, where y_k is taken to be an observation from the uniform distribution on $I_{k,m}$, with different y_k being independent. Denote the conditional predictive density of x^n given $y_1, \ldots, y_m$ by

$$g_{m,n,y}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \hat{f}^*_{m,t-1}(x_t \mid x^{t-1}, y_1, \ldots, y_m)$$

where $\hat{f}^*_{m,t-1}$ is the histogram based on H_m and data $x^{t-1}, y_1, \ldots, y_m$:

$$\hat{f}^*_{m,t-1}(x_t \mid x^{t-1}, y_1, \ldots, y_m) = m \frac{N_{k,m}(t-1)+1}{t-1+m} \qquad \text{on } I_{k,m}.$$

Note that the form of $\hat{f}^*_{m,t-1}$ is independent of the particular values of the y_k, and so we can integrate out y_k and get the same expression as the unconditional density on x^n. Denote the result by $g_{m,n}$.

The expression for $g_{m,n}$ can be greatly simplified, as we can reorder the terms in the product without changing its value. Letting $k(x_t)$ be the unique k such that $x_t \in I_{k,m}$,

$$\begin{aligned} g_{m,n}(x^n) &= \prod_{t=1}^{n} m \frac{N_{k(x_t),m}(t-1)+1}{t-1+m} \\ &= \frac{m^n (m-1)!}{(n+m-1)!} \prod_{t=1}^{n} (N_{k(x_t),m}(t-1)+1) \\ &= \frac{m^n (m-1)!}{(n+m-1)!} \prod_{k=1}^{n} \prod_{k(x_t)=k} (N_{k(x_t),m}(t-1)+1) \\ &= \frac{m^n (m-1)!}{(n+m-1)!} \prod_{k=1}^{n} N_{k,m}(n)!. \end{aligned} \tag{2.1}$$

The above expression is exactly the same as equation (2.3) in Hall and Hannan (1988), who took the stochastic complexity version of the code length, that is, they took the marginal distribution of x^n obtained by integrating out the multinomial parameters with respect to a uniform prior. This is known to be equivalent to adding m more observations, which is what we have done.

By Stirling's formula $n! = n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}\, e^{c_n}$, where $|c_n| < \frac{1}{12}(n+1)$, we can expand the factorials in (2.1). Cancelling out terms and using the convention that $\log N_k = 0$, if $N_k = 0$, we get:

$$\begin{aligned} -\log g_{m,n}(x^n) &= -\sum_{t=1}^{n} \log \hat{f}_{m,n}(x_t) - \tfrac{1}{2} \sum_{k=1}^{m} \log N_k + m \log \frac{n}{m} + R_{m,n} \\ &= -\sum_{t=1}^{n} \log \hat{f}_{m,n}(x_t) + \tfrac{1}{2} \sum_{k=1}^{m} \log \frac{n^2}{m^2 N_k} + R_{m,n}\,, \end{aligned} \tag{2.2}$$

where $\hat{f}_{m,n}(x) = m\, n^{-1} \sum_{k=1}^{m} N_k\, 1_{I_{k,m}}(x)$ is the histogram density estimator based on x^n, and $N_{k,m}(n)$ is abbreviated N_k for simplicity. Moreover, by the bound on c_n, the remainder term $R_{m,n}$ is bounded by an $O(m)$ term.

We can interpret expression (2.2) as a two-part code length: the first term is the code length for the data x^n using the fitted model from H_m, and the second is the code length required to encode the estimators of parameters of the fitted model. (The number of bits required to encode a parameter estimate which has standard deviation σ is approximately $-\log \sigma$, and the value $mn^{-1}N_k$ which $\hat{f}_{m,n}$ takes on $I_{k,m}$ has a variance of approximately $m^2 n^{-2} N_k$, provided m and n are sufficiently large.)

Now we have obtained a class $\{g_{m,n}\}$ of densities on $[0,1]^n$. If we want to use one of them to encode data with an underlying density f, the best code in the average sense will be the one which gives the smallest redundancy. Therefore we would like to find the $g_{m,n}$ minimizing $E_f \log f^n/g_{m,n} = E_f \log f^n - E_f \log g_{m,n}$, that is, the one (those) minimizing $-E_f \log g_{m,n}$. As will be seen later (Theorem 2.4), the minimizer is asymptotically unique in the sense that the ratio of any two minimizers tends to 1 as n tends to infinity. However, the quantity $-E_f \log g_{m,n}$,

is unknown since we do not know the underlying density f, but we can always minimize the data-driven quantity $-\log g_{m,n}(x^n)$ and hope that it will behave like its expectation. Hence we propose the following selection criterion for the code or bin-width:

Selection rule. Choose $\hat{m}_n$ to minimize $-\log g_{m,n}(x^n)$ over a reasonable range of m, where the reasonable range will be specified later. In practice, the range can be chosen according to ad hoc rules.

In order to analyze the behavior of our selection rule further, we need to put some smoothness conditions on the underlying true density f.

Let F denote the family of densities on $[0,1]$ which are uniformly bounded above and below, and also have uniform bound on the first derivative:

$$F = \left\{ f : 0 < c_0 \le f(x) \le c_1 < \infty, \, |f'(x)| \le c_2, \, x \in [0,1]; \, \int_0^1 f \, dx = 1 \right\}.$$

For results in §5, we will restrict the range of m for selection to

$$A_n = \left\{ m : n^{\epsilon_1} \le m \le n^{\epsilon_2} \right\}$$

where ϵ_1 and ϵ_2 are two small positive constants satisfying $0 < \epsilon_1 < \epsilon_2 < 1$.

On the range A_n, we have the following expansions for the per symbol redundancy using $g_{m,n}$.

Theorem 2.1. *Suppose that f is in F, x^n denotes* i.i.d *observations from f, and $f \neq 1$. Then as $n \to \infty$, uniformly in m on A_n:*

$$\frac{1}{n} \log \frac{f^n(x^n)}{g_{m,n}(x^n)} = \left(\tfrac{1}{2} c_f \frac{1}{m^2} + \tfrac{1}{2} \frac{m}{n} \log \frac{n}{m} \right)(1 + o(1)) \quad \text{a.s.}$$

where $c_f = \frac{1}{12} \int_0^1 \frac{f'^2}{f} \, dx$.

It is obvious that c_f measures the smoothness of the density function f, but it is zero if f is the uniform density on $[0,1]$. In this case we will offer a separate discussion, see Theorem 2.5 below. The above expansion is also true in expectation.

Theorem 2.2. *Suppose that f is in F, x^n denotes* i.i.d *observations from f, and $f \neq 1$. Then as $n \to \infty$, uniformly in m on A_n:*

$$\frac{1}{n} E \log \frac{f^n(x^n)}{g_{m,n}(x^n)} = \left(\tfrac{1}{2} c_f \frac{1}{m^2} + \tfrac{1}{2} \frac{m}{n} \log \frac{n}{m} \right)(1 + o(1)) .$$

It is clear that selecting m by minimizing $-\log g_{m,n}$ is asymptotically equivalent, almost surely and in expectation, to trading off $\frac{1}{2} c_f \frac{1}{m^2}$ and $\frac{1}{2} m \log \frac{n}{m}$. When m gets large, the first of these terms gets smaller, and the second gets larger. In fact, the first term is an approximation to the model error between f and H_m, and the second an accumulated estimation error from estimating parameters in H_m. We now make these remarks more precise.

Let f_m denote the density in H_m which assigns the same probability to each bin $I_{k,m}$ as f does, i.e., for $x \in [0,1]$ let

$$f_m(x) = m \sum_{k=1}^{m} \phi_k \, 1_{I_{k,m}}(x) ,$$

where $\phi_k = \int_{I_{k,m}} f \, dx$.

We then have the following theorem. It shows that f_m is the closest element of H_m to f in the sense of the Kullback–Leibler (K–L) divergence, and also gives an approximation to this closest "distance". The proof for the first part is straightforward application of Shannon's inequality. The second part is an exercise in analysis; we refer to Lemma 4.6 for further details.

Theorem 2.3. (i) *For any fixed m,*

$$\min_{h \in H_m} E_f \log \frac{f}{h} = E_f \log \frac{f}{f_m} .$$

(ii) *For m uniformly in the range A_n:*

$$E_f \log \frac{f}{f_m} = \tfrac{1}{2} c_f \frac{1}{m^2} (1 + o(1)) .$$

We now offer a heuristic proof for the expansion in expectation, which shows clearly how the above two terms arise.

Firstly, we insert $f_m^n(x^n)$, break $E_f \log f^n/g_{m,n}$ into two terms by the additivity of the log function, and observe that $f_m^n(x^n)/g_{m,n}(x^n)$ depends on x^n only through the counts N_k. These counts have the same distribution under f and f_m, and so we have

$$\begin{aligned} E_f \log \frac{f^n}{g_{m,n}} &= E_f \log \frac{f^n}{f_m^n} + E_f \log \frac{f_m^n}{g_{m,n}} \\ &= E_f \log \frac{f^n}{f_m^n} + E_{f_m} \log \frac{f_m^n}{g_{m,n}} \\ &\approx n \tfrac{1}{2} c_f \frac{1}{m^2} + E_{f_m} \log \frac{f_m^n}{g_{m,n}} . \end{aligned}$$

The second term is known to have the order $\frac{1}{2} m \log \frac{n}{m}$ by results in the parametric case, see Rissanen (1986). This is because under f_m we are dealing with the multinomial parametric family and $\frac{1}{2} m \log \frac{n}{m}$ is the rate we can get in such a case. Note that the asymptotic expression of $-\log g_{m,n}$ has a unique minimum. This expression can be viewed as the accumulated estimation error for estimating the multinomial parameters, or equivalently as the smallest redundancy achievable by using a code rather than the true density f_m^n.

We turn now to a discussion of the optimality and consistency of the selection rule.

It is clear that the rule

$$\hat{m}_n = \arg(\min_{m \in A_n} (-\log g_{m,n}(x^n)))$$

will also minimize the data-dependent redundancy

$$\log f^n(x^n)/g_{m,n}(x^n) = -\log g_{m,n}(x^n) - (-\log f^n(x^n)) .$$

Note, however, that $\hat{m}_n$ might not be unique for any particular data string x^n. To be precise, we should define $\hat{m}_n$ as a minimizer of $-\log g_{m,n}$. Fortunately, $\hat{m}_n$ is asymptotically unique in the sense described in Theorem 2.4 below.

The asymptotic agreement of the data-dependent redundancy and the expected redundancy suggests that our $\hat{m}_n$ will approach the m which minimizes the expected

redundancy, that is, $\hat{m}_n$ will approach

$$m_n^* = \arg\left(\min_{m \in A_n} E_f \log \frac{f^n(x^n)}{g_{m,n}(x^n)} \right)$$

as $n \to \infty$. This can be stated formally as follows.

Theorem 2.4. *Under the assumptions of Theorem 2.1, we have the following expression as* $n \to \infty$.

(i) $m_n^* = (3c_f\, n/\log n)^{\frac{1}{3}}\,(1 + o(1))$.

(ii) $\hat{m}_n = m_n^*(1 + o(1))$ a.s.

(iii) $E_f \log f^n(x^n)/g_{m_n^*,n}(x^n) = (3c_f)^{\frac{1}{3}}\,(n^{-1}\log n)^{\frac{2}{3}}\,(1 + o(1))/2$.

(iv) $\log f^n(x^n)/g_{\hat{m}_n,n}(x^n) = \left\{E_f \log f^n(x^n)/g_{m_n^*,n}(x^n)\right\}(1 + o(1))$ a.s.

As remarked by Hall and Hannan (1989), the rate of $\hat{m}_n \sim (n/\log n)^{\frac{1}{3}}$ coincides with the rate for the bin width which minimizes the largest deviation of the histogram and the true density.

A question left from the earlier discussion is the case $f \equiv 1$, for here we do not have the asymptotic expansions given in Theorems 2.1 and 2.2. In this case, the model error term is zero, i.e., $f_m \equiv f \equiv 1$ for all m. Fortunately, our selection rule $\hat{m}_n$ will lead us to the uniform density as n gets large, but to allow this, we have to give $\hat{m}_n$ a range which permits the choice $m = 1$. Let us define the range B_n by

$$B_n = \left\{m : 1 \le m \le n^{\epsilon_2}\right\},$$

where $0 < \epsilon_2 < 1$.

Theorem 2.5. *If* x^n *is a sequence of* i.i.d *observations from the uniform density on* $[0,1]$, *then* $\hat{m}_n \equiv 1$ a.s. *as* $n \to \infty$.

Simulations show that $\hat{m}_n$ will correctly choose the uniform density even for moderate sample sizes, for example, $n = 25$ to 50. If f is a stepwise density, i.e.,

belongs to H_m for some $m > 1$, $\hat{m}_n$ then will also pick up the correct m eventually. For those cases where f is close the uniform density, i.e., when the jumps between the blocks are not very big, our selection rule tends to pick the uniform density until the sample size gets very lage. An example is that when f is a density with a jump of size 0.2 at $\frac{1}{2}$, 1500 data points are still not enough to recognize the jump. This behavior is not unreasonable if we want a density estimate as being close in global measure to the true one, since it will pick up the uniform density which is close to the true stepwise density, but it is less satisfactory as a device for locating the positions of the jumps. The same phenomenon is observed in the context of spectral density estimation, when a similar minimum code length criterion can be used, see Hannan, Cameron and Speed (1990).

3. Minimax lower bounds

In the previous section we found the best code $g_{m_n^*,n}$ among a specified class from a data-driven criterion (2.1). In fact, we showed two things. First, we calculated a lower bound on the expected redundancy using one of the codes in our class $\{g_{m,n} : m \in A_n\}$, i.e., for any fixed f in F:

$$\min_{m \in A_n} \frac{1}{n} E_f \log \frac{f^n(x^n)}{g_{m,n}(x^n)} \geq (3c_f)^{\frac{1}{3}} (n^{-1} \log n)^{\frac{2}{3}} (1 + o(1))/2 . \qquad (3.1)$$

The constant c_f can be large or small depending on the smoothness of f. Since f is usually unknown, the lower bound in (3.1) does not tell us as much as when the constant were independent of f. This leaves the question of whether there exists a uniform lower bound over f in F.

Second, we proved that $g_{\hat{m}_n,n}$ achieves this lower bound. Note that we restricted ourselves to a very particular class $\{g_{m,n}, m \in A_n\}$ of codes. We therefore ask: can we do better than this if we search through a larger class?

Strictly speaking, $g_{\hat{m}_n,n}$ is not a density on n-tuples because $\hat{m}_n$ depends on our data x^n. To make $g_{\hat{m}_n,n}$ a code, we need to encode the estimator of the model class, i.e., the integer $\hat{m}_n \approx [(3c_f\, n/\log n)^{\frac{1}{3}}]$. Using the universal prior on integers

(Rissanen (1983)) will cost $n^{-1} \log \hat{m}_n \approx O(n^{-1} \log n)$ extra bits per symbol, which is negligible compared with the order of the lower bound $(n^{-1} \log n)^{\frac{2}{3}}$. Thus, it causes no problem to treat $g_{\hat{m}_n,n}$ as a code on n-tuples, but it is still not a predictive code. We need to scan all the data before we can decide which code we are going to use for this particular data string. Obviously, this can sometimes be inconvenient.

In this section we will try to provide answers to the two questions asked above. In other words, what is the minimax lower bound on the expected redundancy? And, can we achieve this lower bound using a predictive code, i.e. a code that can be implemented using only one pass through the data?

For any code $g(x^n)$, i.e., density function on $[0,1]^n$, we can write the redundancy encoding messages of length n from a continuous source f as an accumulated prediction error as follows:

$$E_{f^n} \log \frac{f^n}{g} = \sum_{t=1}^{n} E_{f^{t-1}} \int \log \frac{f(x_t)}{g(x_t \mid x^{t-1})} f(x_t)\, dx_t \tag{3.2}$$

where f^{t-1} is the (product) density function of x^{t-1}.

It is very useful for us to recognize that each term in the summation is the density estimation error of g_{t-1} expressed in terms of K–L divergence. There is a large body of work in the literature on the minimax rate for density estimation when f is assumed to be a member of a smooth family, see Assouad (1983), Birge (1985), and Devroye (1987) and references cited there. There are different techniques to prove minimax results, but they are similar in one way: they all try to find the worst (hardest) densities to estimate in the family for a fixed sample size. The hope is that the rate at which these densities can be estimated is close to the optimal minimax rate. The worst density class is chosen in such a way that its elements are far apart from each other in all the possible directions so that they cannot all be estimated at the same time at a good rate by any estimator. Minimax results in density estimation tend to be in terms of the L^1, L^2 or Hellinger norms rather than K–L divergence, but we have the followimg well-known inequality which relates K–L divergence to the Hellinger norm.

Lemma 3.1. *For any two densities f and g on $[0,1]$,*

$$\int f \log \frac{f}{g} \geq \left(\sqrt{f} - \sqrt{g}\right)^2 . \tag{3.3}$$

There is another inequality which we will use later.

Lemma 3.2. *For any two densities f and g on $[0,1]$,*

$$\int \min(f,g) \geq 1 - H_2(f,g)$$

where $H_2(f,g) = \int \left(\sqrt{f} - \sqrt{g}\right)^2$.

Unfortunately, we cannot get our results as direct consequences of existing mimimax lower bounds in density estimation since we cannot freely exchange the "min max" and the sum. However, we can repeat the argument used in density estimation for the accumulated estimation error, and get a version of Assouad's minimax lower bound, this time in accumulated Hellinger norm.

Lemma 3.3. *Let F_r be the set of densities on $[0,1]$ with 2^r elements denoted f_θ, where $\theta = (\theta_1, \ldots, \theta_r)$ is an r-tuple of -1's and $+1$'s. Let $\{A_i, i = 1, \ldots, r\}$ be a partition of $[0,1]$ into r disjoint subsets, such that there exists $\alpha > 0$ and $\beta > 0$ with the following properties: for all θ and i we have*

$$\int_0^1 \sqrt{f_\theta \, f_{\theta_i'}} \geq \beta$$

and

$$\int_{A_i} \left(\sqrt{f_\theta} - \sqrt{f_{\theta_i'}}\right)^2 \geq \alpha ,$$

where $(\theta_i')_j = \theta_j$ except at $j = i$, and $(\theta_i')_i = -\theta_i$. Thus for all g in G_n, the set of all densities on $[0,1]^n$, we have

$$\min_{g \in G_n} \max_{f \in F_r} \frac{1}{n} E_{f^n} \log \frac{f^n}{g} \geq \tfrac{1}{4} r\alpha \left(1 - \frac{2\sqrt{2}}{3} \sqrt{n} \sqrt{(1-\beta)}\right) .$$

Proof. We will use the trivial fact that the maximum of r numbers is greater or equal to the average of these r numbers, and then interchange summations.

For any fixed g in G_n,

$$\begin{aligned}
\max_{f\in F_r} \frac{1}{n} E_{f^n} \log \frac{f^n}{g} &\geq 2^{-r} \sum_{f_\theta \in F_r} \frac{1}{n} E_{f_\theta^n} \log \frac{f_\theta^n}{g} \\
&= 2^{-r} \sum_{f_\theta \in F_r} \sum_{t=1}^{n} \frac{1}{n} E_{f_\theta^{t-1}} \int f_\theta \log \frac{f_\theta}{g_{t-1}}\, dx \\
&\geq 2^{-r} \sum_{t=1}^{n} \frac{1}{n} \sum_{f_\theta \in F_r} E_{f_\theta^{t-1}} \int f_\theta \log \frac{f_\theta}{g_{t-1}}\, dx \\
&\geq 2^{-r} \sum_{t=1}^{n} \frac{1}{n} \sum_{f_\theta \in F_r} E_{f_\theta^{t-1}} \int_0^1 \left(\sqrt{f_\theta} - \sqrt{g_{t-1}}\right)^2 .
\end{aligned}$$

Now we look at one term. It can be written as

$$\begin{aligned}
&2^{-r} \sum_{\theta} E_{f_\theta^{t-1}} \int_0^1 \sum_{i=1}^{r} \int_{A_i} \left(\sqrt{f_\theta} - \sqrt{g_{t-1}}\right)^2 \\
&\quad = 2^{-r} \sum_{i=1}^{r} \tfrac{1}{2} \sum_{\theta} \left[E_{f_{\theta_{i+}}^{t-1}} \int_{A_i} \left(\sqrt{f_{\theta_{i+}}} - \sqrt{g_{t-1}}\right)^2 \right. \\
&\qquad \left. + E_{f_{\theta_{i-}}^{t-1}} \int_{A_i} \left(\sqrt{f_{\theta_{i-}}} - \sqrt{g_{t-1}}\right)^2 \right] \\
&\quad \geq 2^{-r-1} \sum_{i=1}^{r} \sum_{\theta} \left[\int_{A_i} \left(\sqrt{f_{\theta_{i+}}} - \sqrt{g_{t-1}}\right)^2 \right. \\
&\qquad \left. + \int_{A_i} \left(\sqrt{f_{\theta_{i-}}} - \sqrt{g_{t-1}}\right)^2 \right] dx_t \min\left(f_{\theta_{i+}}^{t-1}, f_{\theta_{i-}}^{t-1}\right) dx^{t-1} \qquad (3.4) \\
&\quad \geq 2^{-r-2} \sum_{i=1}^{r} \sum_{\theta} \left[\int_{A_i} \left(\sqrt{f_{\theta_{i+}}} - \sqrt{f_{\theta_{i-}}}\right)^2 \right] dx_t \min\left(f_{\theta_{i+}}^{t-1}, f_{\theta_{i-}}^{t-1}\right) dx^{t-1} \\
&\quad \geq \tfrac{1}{4} r\alpha \inf_{\theta,i} \int \min\left(f_{\theta_{i+}}^{t-1}, f_{\theta_{i-}}^{t-1}\right)) \\
&\quad \geq \tfrac{1}{4} r\alpha \inf_{\theta,i} \left(1 - H_2\left(f_{\theta_{i+}}^{t-1}, f_{\theta_{i-}}^{t-1}\right)\right) \\
&\quad \geq \tfrac{1}{4} r\alpha \left(1 - \sqrt{2 - 2\beta^{t-1}}\right) \\
&\quad \geq \tfrac{1}{4} r\alpha \left(1 - \sqrt{2t(1-\beta)}\right) .
\end{aligned}$$

Thus, summing up over t,

$$\begin{aligned}
\max_{f\in F_r} \frac{1}{n} E_{f^n} \log \frac{f^n}{g} &\geq \tfrac{1}{4} r\alpha \frac{1}{n} \sum_{t=1}^{n} \left(1 - \sqrt{2t(1-\beta)}\right) \\
&\geq \tfrac{1}{4} r\alpha \left(1 - \sqrt{n}\, \frac{2\sqrt{2}}{3} \sqrt{(1-\beta)}\right) .
\end{aligned}$$

□

At first sight, it seems that we can try the same argument with the K–L divergence so that we would not have to go through the Hellinger norm, and we might even have a better lower bound this way. However, the above argument will break down at least at one point; equality (3.4) will not hold for K–L divergence since K–L divergence is not a metric, and therefore there is no triangle inequality.

For our class F, we need to find the right subsets F_r with α and β.

Theorem 3.1. *For all $r \geq 1$, there exists a subset F_r of F and a partition $\{A_i, i = 1,\ldots,r\}$ of $[0,1]$ such that the two inequalities in Lema 3.3 hold with $\alpha = \frac{3c_2^2}{32\pi^2} r^{-3}$ and $\beta = 1 - \alpha$. Then*

$$\min_{g \in G_n} \max_{f \in F} \frac{1}{n} E_{f^n} \log \frac{f^n}{g} \geq B_1 \, c_2^{\frac{2}{3}} \, n^{-\frac{2}{3}}$$

where $B_1 \approx 0.011$.

Proof. Let

$$A_i = \left(\frac{i}{r}, \frac{i+1}{r} \right], \qquad i = 0,1,\ldots,r-1 .$$

We now define functions f_i and g_i on A_i, which are building blocks of f_θ in F_r:

$$f_i(x) = \begin{cases} 1 + \frac{c_2}{4r\pi} \left\{ \cos\left(4r\pi\left(x - \frac{i}{r}\right)\right) - 1 \right\} & \text{for } x \in \left(\frac{i}{r}, \frac{i}{r} + \frac{1}{2r}\right] \\ 1 - \frac{c_2}{4r\pi} \left\{ \cos\left(4r\pi\left(x - \frac{i}{r}\right)\right) - 1 \right\} & \text{for } x \in \left(\frac{i}{r} + \frac{1}{2r}, \frac{i}{r} + \frac{1}{r}\right] , \end{cases}$$

and

$$g_i(x) = \begin{cases} 1 - \frac{c_2}{4r\pi} \left\{ \cos\left(4r\pi\left(x - \frac{i}{r}\right)\right) - 1 \right\} & \text{for } x \in \left(\frac{i}{r}, \frac{i}{r} + \frac{1}{2r}\right] \\ 1 + \frac{c_2}{4r\pi} \left\{ \cos\left(4r\pi\left(x - \frac{i}{r}\right)\right) - 1 \right\} & \text{for } x \in \left(\frac{i}{r} + \frac{1}{2r}, \frac{i}{r} + \frac{1}{r}\right] , \end{cases}$$

Then f_i and g_i have the following properties:

$$f_i\left(\frac{i}{r}\right) = f_i\left(\frac{i}{r} + \frac{1}{2r}\right) = f_i\left(\frac{i+1}{r}\right) = 1 ;$$

$$g_i\left(\frac{i}{r}\right) = g_i\left(\frac{i}{r} + \frac{1}{2r}\right) = g_i\left(\frac{i+1}{r}\right) = 1 ;$$

$$f_i'\left(\frac{i}{r}\right) = f_i'\left(\frac{i}{r} + \frac{1}{2r}\right) = f_i'\left(\frac{i+1}{r}\right) = 0 ;$$

$$g_i'\left(\frac{i}{r}\right) = g_i'\left(\frac{i}{r} + \frac{1}{2r}\right) = g_i'\left(\frac{i+1}{r}\right) = 0 ;$$

and

$$\int_{A_i} f_i\,dx = \int_{A_i} g_i\,dx = \frac{1}{r}\,;$$

$$c_0 \le f_i, g_i \le c_1 \text{ if } r > \frac{c_2}{2\pi}\max\left(\frac{1}{1-c_0}, \frac{1}{c_1-1}\right). \tag{3.5}$$

For any $\theta = (\theta_1, \ldots, \theta_r)$, $\theta_i \in \{-1, +1\}$, let $F_r = \{f_\theta\}$, where

$$f_\theta = \begin{cases} f_i & \text{on } A_i, \text{ if } \theta_i = +1\,; \\ g_i & \text{on } A_i, \text{ if } \theta_i = -1\,. \end{cases}$$

By the properties of f_i and g_i, it is clear that $f_\theta \in F$ if (3.5) holds.

Now we are ready to prove that F_r thus defined satisfy the conditions of Lemma 3.3. Let us start with the following claim:

$$\alpha \le \int_{A_i} \left(\sqrt{f_i} - \sqrt{g_i}\right)^2 dx \le \alpha'$$

where $\alpha = \frac{3c_2^2}{32\pi^2}\,r^{-3}$ and $\alpha' = 2\alpha$.

Proof of the claim. By symmetry and periodicity of the cosine function

$$\begin{aligned} T_i &= \int_{A_i} \left(\sqrt{f_i} - \sqrt{g_i}\right)^2 dx \\ &= 4\int_0^{\frac{1}{4r}} \left(\sqrt{f_0} - \sqrt{g_0}\right)^2 dx \\ &= 4\int_0^{\frac{1}{4r}} \frac{\left(f_0 - g_0\right)^2}{\left(\sqrt{f_0} + \sqrt{g_0}\right)^2}\,dx\,. \end{aligned}$$

Note that $f_0 + g_0 = 2$, thus

$$2 = f_0 + g_0 \le \left(\sqrt{f_0} + \sqrt{g_0}\right)^2 \le 2(f_0 + g_0) \le 4\,.$$

Moreover,

$$\int_0^{\frac{1}{4r}} \left(\sqrt{f_0} - \sqrt{g_0}\right)^2 dx = \frac{c_2^2}{(2r\pi)^2}\int_0^{\frac{1}{4r}} (\cos(4r\pi x) - 1)^2\,dx = \frac{3c_2^2}{32\pi^2} r^{-3}\,,$$

hence the claim is proved, as part of the theorem. The proof for the other part is trivial after realizing that

$$\begin{aligned} \int_0^1 \sqrt{f_{\theta_i}, f_{\theta_{i+}}}\,dx &= 1 - \frac{1}{r}\int_{A_i} \sqrt{f_i\,g_i}\,dx \\ &= 1 - \tfrac{1}{2}\int_{A_i} \left(\sqrt{f_i} - \sqrt{g_i}\right)^2 dx \\ &\ge 1 - \tfrac{1}{2}\,\alpha' \\ &= 1 - \alpha = \beta\,. \end{aligned}$$

We can substitute the α and β into the Lemma to get the minimax lower bound. For details, see the proof of the next theorem for the case $s = 1$. □

Generally, the same trick can be used to obtain minimax lower bounds for other classes of smooth densities, see e.g. Devroye (1987). For example, the family of densities on $[0,1]$ with the s-th derivative uniformly bounded

$$F'_s = \left\{ f \in C^s[0,1] : c_0 \le f \le c_1,\ |f^{(j)}| \le c_{j+1},\ j = 1, \ldots, s \right\}$$

or the family of densities in a Sobolev ball

$$W_s = \left\{ f \text{ on } [0,1] : c_0 \le f \le c_1,\ \|f^{(j)}\|_2 \le c_{j+1},\ j = 1, \ldots, s \right\}.$$

We can define f_i and g_i for $s > 1$ by replacing $\frac{c_2}{4r\pi}$ by $\frac{c_{s+1}}{(4r\pi)^s}$ and define $F_r(s) = \{f_\theta\}$ the same way as in Theorem 3.1. Note that $F_r(s)$ is a subset of W_s but not of F'_s (unless $s = 1$) because of the discontinuity of $F_\theta^{(j)}$, $(j = 2, 3, \ldots, s)$ at the end points of A_i. Although we believe that F_θ can be smoothed at the end points of A_i to satisfy the pointwise differentiability so that the modified $F_r(s)$ will be contained in F'_s, we will not try to do it here since it would only be a tedious exercise in analysis.

Theorem 3.2. *For all $r \ge 1$, there exist a subset $F_r(s)$ of W_s and a partition $\{A_i, i = 1, \ldots, r\}$ of $[0,1]$ such that the two inequalities in Lemma 3.3 hold with $\alpha = \frac{3c_s^2+1}{2(4\pi)^{2s}} r^{-(2s+1)}$ and $\beta = 1 - \alpha$. Then*

$$\min_{g \in G_n} \max_{f \in W_s} \frac{1}{n} E_{f^n} \log \frac{f^n}{g} \ge B_s \, c_{s+1}^{\frac{2s}{2s+1}} \, n^{-\frac{2s}{2s+1}},$$

where

$$B_s = \frac{1}{4} \left(\frac{A_s}{9k_s^2} \right)^{-\frac{2s}{2s+1}} (1 - k_s),$$

$$k_s = \frac{4s}{6s+1}$$

and

$$A_s = \frac{3}{2(4\pi)^{2s}}.$$

Proof. As in the proof of Theorem 3.1, we can get

$$a \le \int_{A_i} \left(\sqrt{f_i} - \sqrt{g_i}\right)^2 dx \le \alpha'$$

with $\alpha' = 2\alpha = 2\frac{3c_s^2+1}{2(4\pi)^{2s}} r^{-(2s+1)}$ if $r > \frac{c_s+1}{2\pi} \max\left(\frac{1}{1-c_0}, \frac{1}{c_1-1}\right)$. Hence the two inequalities in Lema 3.3 are satisfied. Substituting α and β into the minimax lower bound gives

$$\tfrac{1}{4} r\alpha\left(1 - \frac{2\sqrt{2}}{3}\sqrt{n}\sqrt{(1-\beta)}\right) = \tfrac{1}{4} A_s r^{-2s}\left(1 - \frac{2\sqrt{2}}{3}\sqrt{n}\sqrt{A_s} r^{-s+\frac{1}{2}}\right).$$

This expression is minimized when $\frac{2\sqrt{2}}{3}\sqrt{n}\sqrt{A_s} r^{-s+\frac{1}{2}} = k_s$, i.e. $r = \left(\frac{A_s n}{9k_s^2}\right)^{\frac{1}{2s+1}}$. Substituting this r into the lower bound, we get $B_s c_{s+1}^{\frac{2s}{2s+1}} n^{-\frac{2s}{2s+1}}$. □

By numerical calculation, B_s seems to be a decreasing function of s and has asymptotic value 0.0033 when $s \to \infty$. For $s = 1$, $B_s \approx 0.01$. It approaches 0.0033 when s is around 20.

4. Achievability of the minimax bounds

For the family F, it has been shown that the rate $n^{-\frac{2}{3}}$ is actually optimal since we can construct a predictive code g^* to achieve this rate in expectation, see Rissanen, Speed and Yu (1989), namely

$$g^*(x^n) = \prod_{t=1}^{n} g_{t-1}(x_t)$$

where

$$g_{t-1}(x) = \sum_{k=1}^{m_t} m_t \frac{N_{k,m_t} + 1}{t - 1 + m_t} 1_{I_{k,m_t}},$$

and $m_t = [(t-1)^{\frac{1}{3}}]$.

Clearly, g^* is the histogram based on x^{t-1}, modified to keep it away from zero. Moreover, g^* can be further modified to achieve this rate almost surely.

Theorem 4.1 (Rissanen, Speed and Yu (1989)). *For all f in F, when n is sufficiently large:*

$$\frac{1}{n} E_{f^n} \log \frac{f^n(x^n)}{g^*(x^n)} \leq A_F\, n^{-\frac{2}{3}}(1 + o(1))$$

where $o(1) \to 0$ uniformly in f in F, as n tends to infinity, and A_F is a constant depending on the class F.

This result is proved in the reference cited.

Theorem 4.2. *Let f^* be a density on x^n defined predictively as follows:*

$$f^*(x^n) = g_{m_1}(x^{\bar{n}}) \prod_{i=n^{4/15}}^{n/\Delta n} g_{m_i}(x_t)$$

where

$$g_{m_t}(x) = \sum_{k=1}^{m_t} m_t \frac{N_{k,m_t} + 1}{t - 1 + m_t} 1_{I_{k,m_t}},$$

$\Delta n = n^{\frac{2}{3}+\delta}$ *for some* $\delta \in (0, \frac{1}{15})$, *and*

$$m_{t-1} = \begin{cases} \left[(\Delta n \cdot n^{\frac{4}{15}})^{\frac{1}{3}}\right] & \text{if } 1 \leq t \leq \left[\Delta n \cdot n^{\frac{4}{15}}\right] \\ \left[(\Delta n \cdot i)^{\frac{1}{3}}\right] & \text{if } (i-1)\Delta n \leq t \leq i \cdot \Delta n \quad \text{and} \quad i \geq n^{\frac{4}{15}}. \end{cases}$$

Then, almost surely as $n \to \infty$,

$$\frac{1}{n} \log \frac{f^n(x^n)}{f^*(x^n)} \leq A_F\, n^{-\frac{2}{3}}(1 + o(1)).$$

This theorem will be proved in §6 below.

The best $g_{\hat{m}_n,n}$ in §2 is no longer optimal over the larger class G_n (compare $(n^{-1} \log n)^{\frac{2}{3}}$ with the minimax rate $n^{-\frac{2}{3}}$), since it is not flexible enough to adopt the shape of the underlying density when more and more data are obtained. This is because we restricted ourselves to a class with histograms with fixed bin width for all the predictive densities g_{t-1}. The truly optimal g^* is the one which updates the bin width each time a new observation comes along, and this way, we do not have to scan the whole string to get the code. Note that the bin-width updating is

at the same rate as would be chosen by an AIC type criterion based on x^{t-1}, since we are in a similar situation to that in Breiman and Freedman (1983), where the model error is never zero.

One problem, however, is the sample stability of g^*, since updating at each new observation is too frequent. On the other hand, $g_{\hat{m}_n,n}$ behaves well along almost every sample path, although it is not optimal in the minimax sense. A natural compromise is f^*, which does not update the bin width as frequently as g^*, but updates frequently enough to keep the optimal rate. As stated in Theorem 4.2, the density f^* behaves as expected.

For the other classes of density families, W_s and F'_s ($s > 1$), g^* will not be minimax optimal. Since these classes are subsets of F, Theorems 4.1 and 4.2 still hold, but $n^{-\frac{2}{3}}$ is slower than the optimal rate, which is $n^{-\frac{2s}{2s+1}}$ for the model classes W_s and F'_s ($s > 1$). Intuitively, the histograms with different bin widths are not smooth enough to be close to the much smoother densities in W_s and F'_s at the optimal rate. We plan to address this topic in a future paper.

A related result in which the same rate arises is given in Barron and Cover (1989). They consider the problem of using the shortest code length principle to derive a density estimator, and describe one which, when the true density lies in a Sobolev ball W_s, converges in the Hellinger distance and in probability at rate $n^{-\frac{2s}{2s+1}}$.

5. Proofs for Section 2

In this section, we will provide proofs of the results in §2. We will start with the proofs of the asymptotic expansions of the redundancy of $g_{m,n}$, both almost surely and in expectation. We then show the optimal properties of $g_{\hat{m}_n,n}$ in the class of codes $\{g_{m,n} : m \in A_n = [n^{\epsilon_1}, n^{\epsilon_2}]\}$. Finally, when f is uniform on $[0,1]$, we will prove that $\hat{m}_n = 1$ a.s., when $\hat{m}_n$ is selected over range $B_n = [1, n^{\epsilon_2}]$.

Proofs of the asymptotic expansions

We assume that x^n is a sequence of i.i.d. observations from the density f in the class F defined in §2. Recall that, by Stirling's formula, we can expand the code length for data x^n using $g_{m,n}$ as:

$$-\log g_{m,n}(x^n) = -\sum_{t=1}^{n} \log \hat{f}_{m,n}(x_t) - \tfrac{1}{2} \sum_{k=1}^{m} \log N_k + m \log \frac{n}{m} + O(m) \quad (5.1)$$

where $N_k = \sum_{t=1}^{n} 1_{I_{k,m}}(x_t)$, $\hat{f}_{m,n}(x) = m n^{-1} \sum_{k=1}^{m} N_k 1_{I_{k,m}}(x)$, and we adopt the convention that $\log N_k = 0$ if $N_k = 0$.

Thus the redundancy using $g_{m,n}$ is

$$\log \frac{f^n(x^n)}{g_{m,n}(x^n)} = \sum_{t=1}^{n} \log \frac{f(x_t)}{f_m(x_t)} + \sum_{t=1}^{n} \log \frac{f_m(x_t)}{\hat{f}_{m,n}(x_t)} + \left[-\tfrac{1}{2} \sum_{k=1}^{m} \log N_k + m \log \frac{n}{m} \right] + O(m). \quad (5.2)$$

We will show that the first sum in (5.2) is the model error ($c_f n/2m^2$), the second is $o(m) \log n$ and so negligible, and the third is the code length for the parameters $\left(\frac{m}{2} \log \frac{n}{m}\right)$. The following Lemma will also be needed.

Lemma 5.1. *Suppose that $f \in F$ and write*

$$J_{m,n} = \max_{1 \le k \le m} |m N_k/n - m \phi_k| .$$

Then for any integer $q > 0$ and m in $A_n = \{m : n^{\epsilon_1} < m < n^{\epsilon_2}\}$, there is a constant a_q such that

(i) $$E J_{m,n}^{2q} \le a_q c_1^q n^{-(1-\epsilon_2)q+\epsilon_2} ,$$

(ii) and, uniformly in $m \in A_n$, as $n \to \infty$

$$J_{m,n} = o(1) \quad \text{a.s.}$$

Proof. (i) By Lemma 3 in the Appendix, for any integer $q > 0$, there is an $a_q > 0$ such that

$$E(N_k - n \phi_k)^{2q} \le a_q (n \phi_k)^q .$$

Therefore,

$$\begin{aligned} E\, J_{m,n}^{2q} &\le (m/n)^{2q} \sum_{k=1}^{m} E(N_k - n\,\phi_k)^{2q} \\ &\le a_q (m/n)^{2q} \sum_{k=1}^{m} n^q\, \phi_k^2 \\ &\le a_q\, c_1^q\, n^{-(1-\epsilon_2)q+\epsilon_2} \end{aligned}$$

since $\phi_k \le c_1\, m^{-1}$ and $m < n^{\epsilon_2}$.

(ii) By Markov's inequality, and taking $q > \frac{2\epsilon_2+q}{1-\epsilon_2}$ in (i),

$$\begin{aligned} \sum_{n}^{\infty} \sum_{m\in A_n} P(J_{m,n} > \epsilon) &\le \sum_{n}^{\infty} \sum_{m\in A_n} \epsilon^{-2q}\, E\, J_{m,n}^{2q} \\ &\le \epsilon^{-2q}\, a_q\, c_1^q \sum_{n}^{\infty} \sum_{m\in A_n} n^{-(1-\epsilon_2)q+\epsilon_2} \\ &\le \epsilon^{-2q}\, a_q\, c_1^q \sum_{n}^{\infty} n^{-(1-\epsilon_2)q+\epsilon_2} \\ &< \infty\,. \end{aligned}$$

By the Borel-Cantelli lemma, (ii) is proved. □

Lemma 5.2. *Suppose that $f \in F$. Then as $n \to \infty$, uniformly in $m \in A_n$,*

(i) $$\sum_{t=1}^{n} \log \frac{f_m(x_t)}{\hat{f}_{m,n}(x_t)} = O(m) \quad \text{a.s.},$$

(ii) $$E \sum_{t=1}^{n} \log \frac{f_m(x_t)}{\hat{f}_{m,n}(x_t)} = O(m)\,.$$

Proof.

$$\begin{aligned} \sum_{t=1}^{n} \log \frac{f_m(x_t)}{\hat{f}_{m,n}(x_t)} &= \sum_{k=1}^{m} N_k \log N_k/n\,\phi_k \\ &= \sum_{k=1}^{m} N_k \log(1 + (N_k - n\,\phi_k)/n\,\phi_k)\,. \end{aligned}$$

Expanding $\log(1 + (N_k - n\,\phi_k)/n\,\phi_k)$, we have

$$\begin{aligned} \sum_{k=1}^{m} N_k \log(1 + (N_k - n\,\phi_k)/n\,\phi_k) &= \sum_{k=1}^{m} N_k(N_k - n\,\phi_k)/n\,\phi_k) \\ &\quad - \tfrac{1}{2} \sum_{k=1}^{m} N_k(1+\theta_k)^{-2}\,(N_k - n\,\phi_k)^2/(n\,\phi_k)^2 \\ &= L + Q\,, \end{aligned}$$

where θ_k is in between 0 and $(N_k - n\phi_k)/n\phi_k$.

(i) Let us look at the first term L.

$$\begin{aligned} L &= \sum_{k=1}^{m} N_k(N_k - n\phi_k)/n\phi_k \\ &= \sum_{k=1}^{m} (N_k - n\phi_k)^2/n\phi_k + \sum_{k=1}^{m} (N_k - n\phi_k) \\ &= \sum_{k=1}^{m} (N_k - n\phi_k)^2/n\phi_k\,, \quad \text{since } \sum_{k=1}^{m} N_k = 1 \\ &\le c_0\, m\, n^{-1} \sum_{k=1}^{m} (N_k - n\phi_k)^2\,, \end{aligned}$$

and so $L = O(m)$, since the sum is of order $n(1 + o(1))$ by Lemma 4(i) of the Appendix.

Moreover, by Lemma 5.1(ii), for n large, a.s. and uniformly in $m \in A_n$,

$$\max_k |\theta_k| \le c_0\, J_{m,n} = o(1) < \tfrac{1}{2}\,.$$

Hence, a.s. and uniformly in m, the second term

$$\begin{aligned} |Q| &= \tfrac{1}{2} \sum_{k=1}^{m} N_k(1+\theta_k)^{-2}\,(N_k - n\phi_k)^2/(n\phi_k)^2 \\ &\le \left(\tfrac{1}{8}\right) \sum_{k=1}^{m} N_k(N_k - n\phi_k)^2/(n\phi_k)^2 \\ &\le \left(\tfrac{1}{8}\right) \sum_{k=1}^{m} (N_k - n\phi_k)^3/(n\phi_k)^2 + \left(\tfrac{1}{8}\right) \sum_{k=1}^{m} (N_k - n\phi_k)^2/n\phi_k \\ &\le \left(\tfrac{1}{8}\right) (c_0\, I_{m,n} + 1)(m/n) \sum_{k=1}^{m} (N_k - n\phi_k)^2 \\ &= m(o(1) + 1)\, O(1) \\ &= O(m)\,. \end{aligned}$$

(ii) If we take the expectation of L, we get

$$E\,L = \sum_{k=1}^{m} E(N_k - n\phi_k)^2/n\phi_k = \sum_{k=1}^{m} (1 - \phi_k) = m - 1\,.$$

It now suffices to show that $E\,|Q| = O(m)$ is true. But

$$2E\,|Q| = 2E\,|Q|\,1_{\{J_{k,m} > \epsilon\}} + 2E\,|Q|\,1_{\{J_{k,m} \le \epsilon\}}\,.$$

Note that on $\{N_k > 0\}$, $\theta_k + 1 \geq (1 - n\phi_k)/n\phi_k + 1 \geq 1/n\phi_k$, and hence

$$\begin{aligned}
2E\,|Q|\,1_{\{J_{k,m}>\epsilon\}} &= \sum_{k=1}^{m} EN_k(N_k - n\phi_k)^2/(n\phi_k)^2\,(1+\theta_k)^{-2}\,1_{\{J_{k,m}>\epsilon\}} \\
&= E\sum_{N_k>0} N_k(N_k - n\phi_k)^2/(n\phi_k)^2\,(1+\theta_k)^{-2}\,1_{\{I_{k,m}>\epsilon\}} \\
&\leq E\sum_{N_k>0} N_k(N_k - n\phi_k)^2\,1_{\{J_{k,m}>\epsilon\}} \\
&= E\sum_{k=1}^{m} N_k(N_k - n\phi_k)^2\,1_{\{J_{k,m}>\epsilon\}} \\
&= \sum_{k=1}^{m} E(N_k - n\phi_k)^3\,1_{\{J_{k,m}>\epsilon\}} \\
&\quad + \sum_{k=1}^{m} En\phi_k(N_k - n\phi_k)^2\,1_{\{J_{k,m}>\epsilon\}}\,. \qquad (5.3)
\end{aligned}$$

By Schwartz's inequality, together with Lemma 3 in the Appendix and Lemma 5.1(i), if we take $q > (3+\epsilon_2)/(1-\epsilon_2)$, we obtain

$$\begin{aligned}
E(N_k - n\phi_k)^2\,1_{\{J_{k,m}>\epsilon\}} &\leq \sqrt{E(N_k - n\phi_k)^6}\,P^{\frac{1}{2}}\,(J_{m,n} > \epsilon) \\
&\leq \sqrt{a_6}\,(n\phi_k)^{\frac{3}{2}}\,\epsilon^{-q}\,\sqrt{E\,J_{m,n}^{2q}} \\
&\leq \sqrt{a_6}\,a_q\,c_1^{q-\frac{3}{2}}\,\epsilon^{-q}\,n^{-[(1-\epsilon_2)q-\epsilon_2]/2+\frac{3}{2}} \\
&= O(1)\,.
\end{aligned}$$

Similarly we can get

$$\sum_{k=1}^{m} En\phi_k(N_k - n\phi_k)^2\,1_{\{J_{k,m}>\epsilon\}} = O(m)\,.$$

Thus

$$E\,|Q|\,1_{\{J_{k,m}>\epsilon\}} = O(m)\,. \qquad (5.4)$$

Since $|\theta_k| \leq J_{m,n}$,

$$\begin{aligned}
2E\,|Q|\,1_{\{J_{k,m}\leq\epsilon\}} &= E\sum_{k=1}^{m}(N_k - n\phi_k)^2/(n\phi_k)^2\,(1+\theta_k)^{-2}\,1_{\{J_{k,m}\leq\epsilon\}} \\
&\leq (1-\epsilon)^{-2}\,E\sum_{k=1}^{m} N_k(N_k - n\phi_k)^2/(n\phi_k)^2\,1_{\{J_{k,m}\leq\epsilon\}}
\end{aligned}$$

$$\leq (1-\epsilon)^{-2} \sum_{k=1}^{m} E\, N_k (N_k - n\phi_k)^2/(n\phi_k)^2$$

$$\leq (1-\epsilon)^{-2}\, c_0^{-2} (m/n)^2 \left\{ \sum_{k=1}^{m} E(N_k - n\phi_k)^3 + \sum_{k=1}^{m} E(N_k - n\phi_k)^2 \right\}$$

$$\leq (1-\epsilon)^{-2}\, c_0^{-2} (m/n)^2\, 2 \sum_{k=1}^{m} n\phi_k$$

$$\leq 2(1-\epsilon)^{-2}\, c_0^{-2} (m/n)^2\, n$$

$$\leq m\, 2(1-\epsilon)^{-2}\, c_0^{-2} (m/n)^2\, n^{(1-\epsilon_2)}$$

$$= O(m)\,. \tag{5.5}$$

Putting (5.4) and (5.5) together, we have $E\,|Q| = O(m)$, and the proof is complete.□

Similarly, we can show

Lemma 5.3. *For $f \in F$, as $n \to \infty$, uniformly in $m \in A_n$:*

(i) $$\sum_{N_k>0} \log \frac{N_k}{n\phi_k} = O(m) \quad \text{a.s.},$$

(ii) $$E_f \sum_{N_k>0} \log \frac{N_k}{n\phi_k} = O(m)\,.$$

Proof. As in the proof of Lemma 5.2, we use a Taylor exansion to get

$$\sum_{N_k>0} \log \frac{N_k}{n\phi_k} = \sum_{N_k>0} (N_k - n\phi_k)/n\phi_k - \left(\tfrac{1}{2}\right) \sum_{N_k>0} (1+\theta_k)^{-2} (N_k - n\phi_k)^2/(n\phi_k)^2\,.$$

By Lemma 4(ii) in the Appendix, we note that the first sum is $o(m)$ a.s. and uniformly in $m \in A_n$, and its expectation is zero. The second sum is bounded by the corresponding (absolute value of) quadratic term $|Q|$, since

$$\sum_{N_k>0} (1+\theta_k)^{-2} (N_k - n\phi_k)^2/(n\phi_k)^2 \leq \sum_{N_k>0} N_k (1+\theta_k)^{-2} (N_k - n\phi_k)^2/(n\phi_k)^2$$

$$= \sum_{k=1}^{m} N_k (1+\theta_k)^{-2} (N_k - n\phi_k)^2/(n\phi_k)^2$$

$$= |Q|\,.$$

By the bound on $|Q|$ from the last lemma, Lemma 5.3 is proved. □

The almost sure statements in Lemma 5.2(i) and Lemma 5.3(i) were obtained, using Borel-Cantelli lemma, by showing

$$\sum_{n=1}^{\infty} P\Big(\max_{m \in A_n} \Big| \sum_{k=1}^{m} N_k \log \frac{N_k}{n\,\phi_k} \Big| > \epsilon m \Big) < \infty ,$$

and

$$\sum_{n=1}^{\infty} P\Big(\max_{m \in A_n} \Big| \sum_{k=1}^{m} \log \frac{N_k}{n\,\phi_k} \Big| > \epsilon m \Big) < \infty .$$

We refer to the proof of Lemma 4 in the Appendix for details.

Note also that the foregoing two lemmas actually hold on $B_n \supset A_n$, while the following two lemmas hold only for A_n.

Lemma 5.4. *For f in F, uniformly in $m \in A_n$, as $n \to \infty$*

$$\sum_{t=1}^{n} \log \frac{f(x_t)}{f_m(x_t)} = n\,E_f \log \frac{f}{f_m} + o\Big(\frac{n}{m^2} + m \log \frac{n}{m} \Big) \quad \text{a.s.}$$

Proof. Since $m > n^{\epsilon_1}$ on A_n, it is sufficient to show that

$$\sum_{t=1}^{n} \log \frac{f(x_t)}{f_m(x_t)} = n\,E_f \log \frac{f}{f_m} + o\Big(\frac{n}{m^2} + m \log n \Big) \quad \text{a.s.}$$

For any fixed m in A_n, let $Y_{t,m} = \log \frac{f(x_t)}{f_m(x_t)}$.

Then the $Y_{t,m}$ are i.i.d. for $t = 1, 2, \ldots, n$, and

$$|Y_{t,m}| \le \frac{1}{c_0} \max_x |f(x) - f_m(x)| \le \frac{c_2}{c_0}\, m^{-1} .$$

Therefore

$$|Y_{t,m} - E\,Y_{t,m}| \le 2\,\frac{c_2}{c_0}\, m^{-1}$$

and

$$V = 4n\,\frac{c_2^2}{c_0^2}\, m^{-2} \ge \sum_{t=1}^{n} \operatorname{var} Y_{t,m} .$$

By Bernstein's inequality, for any $\epsilon > 0$:

$$P\left(\left|\sum_{t=1}^{n}(Y_{t,m} - E\,Y_{t,m})\right| > \eta\right) \leq 2\,\exp\left(-\tfrac{1}{2}\,\eta^2/(V + \tfrac{1}{3}\,M\,\eta)\right), \qquad (5.6)$$

where $M = 2\,\frac{c_2}{c_0}\,m^{-1}$ and $\eta = n\,\epsilon(m^{-2} + m\,n^{-1}\,\log\,n)$.

Note that

$$\begin{aligned} V + \tfrac{1}{3}\,M\,\eta &= 2n\frac{c_2}{c_0}\,m^{-2} + \tfrac{2}{3}\,\frac{c_2}{c_0}\,m^{-1}\,n\,\epsilon(m^{-2} + m\,n^{-1}\,\log\,n) \\ &= 2n\,\frac{c_2}{c_0}\,m^{-2}\left(1 + \frac{\epsilon}{3}\,(m^{-2} + (m/n)^2\,\log\,n)\right) \\ &= 2n\,\frac{c_2}{c_0}\,m^{-2}\left(1 + \frac{\epsilon}{3}\,(n^{-2\epsilon_1} + n^{-2(1-\epsilon_2)}\,\log\,n)\right) \\ &\leq 4n\,\frac{c_2}{c_0}\,m^{-2}, \end{aligned}$$

for n sufficiently large that $\frac{\epsilon}{3}\,(n^{-2\epsilon_2} + m^{-2(1-\epsilon_2)}\,\log\,n) < 1$. Thus,

$$\begin{aligned} \eta^2/(V + \tfrac{1}{3}\,M\,\eta) &\geq \eta^2/\left(4n\,\frac{c_2}{c_0}\,m^{-2}\right) \\ &= \frac{\epsilon^2\,c_0}{4c_2}\,n(m^{-1} + m^2\,n^{-1}\,\log\,n)^2 \\ &\geq \frac{\epsilon^2\,c_0}{4c_2}\,n\,\min_{m\in A_n}(m^{-1} + m^2\,n^{-1}\,\log\,n)^2 \\ &= O\big(n^{\frac{2}{3}}\,(\log\,n)^{\frac{1}{3}}\big). \end{aligned}$$

Substituting this bound into (5.6), we get

$$P\left(\left|\sum_{t=1}^{n}(Y_{t,m} - E\,Y_{t,m})\right| > \eta\right) \leq 2\,\exp\left(-O\big(n^{\frac{2}{3}}(\log\,n)^{\frac{1}{3}}\big)\right).$$

Thus as $n \to \infty$,

$$\begin{aligned} \sum_{n=1}^{\infty}\sum_{m\in A_n} P\left(\left|\sum_{t=1}^{n}(Y_{t,m} - E\,Y_{t,m})\right| > \eta\right) &\leq 2\sum_{n=1}^{\infty}\sum_{m\in A_n}\exp\left(-O\big(n^{\frac{2}{3}}(\log\,n)^{\frac{1}{3}}\big)\right) \\ &\leq 2\sum_{n=1}^{\infty} n^{\epsilon_2}\,\exp(-O\big(n^{\frac{2}{3}}(\log\,n)^{\frac{1}{3}}\big)) \\ &< \infty. \end{aligned}$$

Using the Borel-Cantelli lemma, the proof of Lemma 5.4 is complete. □

We need a result from Freedman and Diaconis (1981) to obtain Lemma 5.6, which gives an approximation in the K–L divergence between the true density f and the model class H_m.

Lemma 5.5 (Freedman and Diaconis (1981)). *Suppose that f is in F, and let $h = m^{-1}$. Then*

$$\sum_{k=1}^{m-1} \left| \int_{kh}^{(k+1)h} \left(\frac{1}{h^2} (f - f_m)^2 - \tfrac{1}{12} f'^2 \right) \right| dx = o(1) \quad \text{as} \quad h \to 0 .$$

The proof of this Lemma is essentially contained in the proof of Proposition (2.7) of Freedman and Diaconis (1981).

Lemma 5.6. *For f in F and $f \neq 1$, we have as $m \to \infty$*

$$E_f \log \frac{f}{f_m} = \tfrac{1}{2} c_f \frac{1}{m^2} (1 + o(1)) .$$

Proof. For any fixed m, we write $h = m^{-1}$ and $L(m) = E_f \frac{(f - f_m)^2}{f^2}$. If $\ell(x) = 1/f(x)$, then $\ell'(x) = -f'(x)/f^2(x)$, and $|\ell'(x)| \leq \frac{c_2}{c_0^2} = A$. Moreover, on $I_k = [k/m, (k+1)/m] = [kh, (k+1)h]$, $\ell(x)$ can be written as

$$\ell(x) = \ell(kh) + \int_{kh}^{x} \ell'(t)\, dt .$$

Writing $a_k = \ell(kh)$, we have

$$\begin{aligned}
h^{-2} L(m) - c_f &= \sum_{k=1}^{m-1} \int_{I_k} \left(h^{-2} f(x)^{-1} (f(x) - f_m(x))^2 - \tfrac{1}{12} f(x)^{-1} f'^2(x) \right) dx \\
&= \sum_{k=1}^{m-1} \int_{I_k} \left(a_k + \int_{kh}^{x} \ell'(t) dt \right) \left\{ h^{-2} (f(x) - f_m))^2 - \tfrac{1}{12} f'^2(x) \right\} dx \\
&= \sum_{k=1}^{m-1} a_k \int_{I_k} \left\{ h^{-2} (f(x) - f_m(x))^2 - \tfrac{1}{12} f'^2(x) \right\} dx \\
&\quad + \sum_{k=1}^{m-1} \int_{I_k} \int_{kh}^{x} \ell'(t)\, dt \left\{ h^{-2} (f(x) - f_m))^2 - \tfrac{1}{12} f'^2(x) \right\} dx \\
&= I_1 + I_2 .
\end{aligned}$$

Note that $|f - f_m| \leq c_2 h$, $|f'| \leq c_2$, and $\left| \int_{kh}^{x} \ell'(t)\, dt \right| \leq \int_{kh}^{x} |\ell'(t)|\, dt \leq Ah$ for x in I_k, and so we get

$$I_2 \leq hA \left(c_2 + \frac{c_2^2}{12} \right) \int_{I_k} dx = O(h) = o(1) .$$

The last bound on I_2 together with $I_1 = o(1)$ (by Lemma 5.5) gives

$$h^{-2}\, L(m) - c_f = o(1) \quad \text{as} \quad h \to 0\,.$$

That is

$$E_f\, \frac{(f - f_m)^2}{f^2} = c_f\, m^{-2}(1 + o(1))\,.$$

Using a Taylor expansion,

$$\log \frac{f}{f_m} = -\frac{f - f_m}{f} + \tfrac{1}{2}\, \frac{(f - f_m)^2}{f^2} - \tfrac{1}{3}\, \frac{1}{(1 + \theta(x))^3}\, \frac{(f - f_m)^3}{f^3}$$

where $|\theta(x)| \le \frac{|f - f_m|}{f} \le \frac{c_2}{c_0}\, h$. Hence

$$\begin{aligned} E_f \log \frac{f}{f_m} &= E_f - \frac{f - f_m}{f} + E_f\, \tfrac{1}{2}\, \frac{(f - f_m)^2}{f^2} - E_f\, \tfrac{1}{3}\, \frac{1}{(1 + \theta(x))^3}\, \frac{(f - f_m)^3}{f^3} \\ &= \tfrac{1}{2}\, L(m) - E_f\, \tfrac{1}{3}\, \frac{1}{(1 + \theta(x))^3}\, \frac{(f - f_m)^3}{f^3} \\ &= \tfrac{1}{2}\, L(m) + O(h)\, L(m)\,, \end{aligned}$$

since

$$\left| E_f\, \tfrac{1}{3}\, \frac{1}{(1 + \theta(x))^3}\, \frac{(f - f_m)^3}{f^3} \right| \le \frac{1}{(1 - m^{-1})^3}\, E_f\, \frac{|f - f_m|^3}{f^3} \le \frac{1}{(1 - m^{-1})^3}\, \frac{c_2}{c_0}\, hL(m)\,.$$

Finally, we have

$$\begin{aligned} E_f \log \frac{f}{f_m} &= \tfrac{1}{2}\, L(m) + O(h)\, L(m) \\ &= \tfrac{1}{2}\, L(m)(1 + o(1)) \\ &= \tfrac{1}{2}\, c_f\, m^{-2}(1 + o(1))\,. \end{aligned}$$

□

Because $c_0 \le m\, \phi_k \le c_1$, we have

$$\sum_{k=1}^{m} \log n\, \phi_k = \sum_{k=1}^{m} \log \frac{n}{m} \sum_{k=1}^{m} \log m\, \phi_k = m \log \frac{n}{m} + O(m)\,.$$

Now we are ready to obtain the asymptotic expansions given in §2.

Proof of Theorem 2.1

Combining the above observations with Lemma 5.2(i), Lemma 5.3(i), Lemma 5.4 and Lemma 5.6, we can complete the proof of Theorem 2.1.

The pointwise redundancy is

$$\begin{aligned}
\log \frac{f^n(x^n)}{g_{m,n}(x^n)} &= \sum_{t=1}^{n} \log \frac{f(x_t)}{f_m(x_t)} + \sum_{t=1}^{n} \log \frac{f_m(x_t)}{\hat{f}_{m,n}(x_t)} \\
&\quad + \left[-\tfrac{1}{2} \sum_{k=1}^{m} \log N_k + m \log \frac{n}{m} \right] + O(m) \\
&= n\, E_f \log \frac{f}{f_m} + O(m) + \left[-\tfrac{1}{2} \sum_{k=1}^{m} \log N_k / n\phi_k \right] \\
&\quad - \tfrac{1}{2} m \log \frac{n}{m} + m \log \frac{n}{m} \\
&= \tfrac{1}{2} c_f\, n m^{-2} + \tfrac{1}{2} m \log \frac{n}{m} + O(m) \quad \text{a.s.}
\end{aligned}$$

□

Similarly, we have

Proof of Theorem 2.2

Combining the above observations with Lemma 5.2(ii), Lemma 5.3(ii), Lemma 5.4 and Lemma 5.6, we complete the proof of Theorem 2.2. □

There is a nice little fact which asserts that f_m is the "closest" element of H_m to f in a certain sense. The Pythagorean identity is known to be satisfied by the I-projection, but f_m is not the I-projection of f to H_m: we are minimizing the K–L "distance" in the other variable.

Lemma 5.7. *For any h in H_m and f in F,*

$$E_f \log \frac{f}{h} = E_f \frac{f}{f_m} + E_{f_m} \log \frac{f_m}{h},$$

where $f_m = \sum_{k=1}^{m} m\phi_k 1_{I_{k,m}}$. Equivalently,

$$K(f,h) = K(f,f_m) + K(f_m,h).$$

Proof. For any $h = \sum_{k=1}^{m} p_k\, m\, I_{k,m}$ in H_m with $\sum_{k=1}^{m} p_k = 1$,

$$\begin{aligned} E_f \log \frac{f}{h} - E_f \frac{f}{f_m} &= E_f \log \frac{f_m}{h} \\ &= \sum_{k=1}^{m} \int_{I_{k,m}} f \log \frac{m\,\phi_k}{m\,p_k} \\ &= \sum_{k=1}^{m} \phi_k \log \frac{\phi_k}{p_k} \\ &= E_{f_m} \log \frac{f_m}{h}\,. \end{aligned}$$

□

Proof of Theorem 2.3

Part (i) is proved in Lemma 5.6, so it is sufficient to prove part (ii). It follows from Lemma 5.7 that

$$E_f \log \frac{f}{h} - E_f \frac{f}{f_m} = E_{f_m} \log \frac{f_m}{h} \geq 0\,,$$

and this completes the proof for (ii) since h was arbitrary. □

Proofs of the optimality and consistency of our selection rule

This section contains proofs of the results of §2 concerning (1) the optimality of $g_{\hat{m}_n,n}$ in the class of codes $\{g_{m,n} : m \in A_n\}$, and (2) the consistency of $g_{\hat{m}_n,n}$ if the true density $f \equiv 1$.

Write

$$\begin{aligned} S(m) &= \tfrac{1}{2}\, c_f\, m^{-2} + \tfrac{1}{2}\, \frac{m}{n} \log \frac{n}{m}\,, \\ R(m) &= \frac{1}{n} \log \frac{f^n(x^n)}{g_{m,n}(x^n)}\,, \end{aligned}$$

and

$$L(m) = E_f\, R(m)\,.$$

Denote by $\bar{m}_m$, $\hat{m}_n$ and m_n^* the minimizers of $S(m)$, $R(m)$ and $L(m)$ respectively over a range A_n or B_n of m to be specified later. Then we have the following lemma.

Lemma 5.8. *If we choose $\bar{m}_m$, $\hat{m}_n$ and m_n^* over the range A_n, then as $n \to \infty$*

(i) $L(m_n^*)/S(\bar{m}_n) \to 1\,.$

(ii) $R(\hat{m}_n)/S(\bar{m}_n) \to 1$ a.s.

(iii) $\bar{m}_n = (3c_f\, n/\log n)^{\frac{1}{3}}(1 + o(1))\,.$

(iv) $m_n^*/\bar{m}_n \to 1\,.$

(v) $\hat{m}_n/\bar{m}_n \to 1$ a.s.

Proof. (i) The asymptotic expansion in Theorem 2.2 gives

$$L(m_n^*) = S(m_n^*)(1 + o(1))\,. \tag{5.7}$$

Since m_n^* is the minimizer of $L(m)$,

$$L(m_n^*) \le L(\bar{m}_n) = S(\bar{m}_n)(1 + o(1))$$

which yields

$$L(m_n^*)/S(\bar{m}_n) \le (1 + o(1))\,. \tag{5.8}$$

Similarly, by Theorem 2.2

$$L(\bar{m}_n) = S(\bar{m}_n)(1 + o(1)) \tag{5.9}$$

and

$$S(\bar{m}_n) \le S(m_n^*) = L(m_n^*) - o(1)\, S(m_n^*)$$

which together give

$$L(m_n^*)/S(\bar{m}_n) \ge 1 + o(1)\, S(m_n^*)\, S(\bar{m}_n)\,. \tag{5.10}$$

Note that $S(m_n^*)/S(\bar{m}_n)$ is bounded, since otherwise $S(m_n^*)/S(\bar{m}_n) \to \infty$ for some subsequence of n. For the sake of simplicity, we assume this is the case for the whole sequence. The $L(m_n^*)/L(\bar{m}_n) \to \infty$ by (5.7) to (5.9). This contradicts the

fact that $L(m_n^*) \leq L(\bar{m}_n)$, because m_n^* is the minimizer of $L(m)$. Combining (5.8) and (5.10) completes the proof of (i).

(ii) Similar to the proof of (i) by Theorem 2.1.

(iii) Putting the derivative of $L(m)$ with respect to m equal to zero yields

$$c_f\, m^{-3} = n^{-1} \log n - n^{-1} \log m - n^{-1}\,.$$

Thus, $\bar{m}_n$, as the root of this equation, takes the form

$$\bar{m}_n = (3c_f)^{\frac{1}{3}}\,(n/\log n)\,a_n(f)\,,$$

where $a_n(f) \to 1$ as $n \to \infty$. Alternatively, we can write

$$\bar{m}_n = (3c_f)^{\frac{1}{3}}\,(n/\log n)^{\frac{1}{3}}\,(1 + o(1))\,. \tag{5.11}$$

(iv) Substituting (5.11) into S, we have

$$S(\bar{m}_n) = \tfrac{1}{2}\,(3c_f)^{\frac{1}{3}}\,(n^{-1} \log n)^{\frac{2}{3}}\,(1 + o(1))\,.$$

By Theorem 2.2,

$$L(m_n^*)/S(m_n^*) \to 1\,,$$

which, together with (i), implies that

$$S(m_n^*)/S(\bar{m}_n) \to 1\,,$$

that is

$$S(m_n^*)/S(\bar{m}_n) = \frac{\left\{c_f\, m_n^{*-2} + m_n^*\, n^{-1} \log(n/m_n^*)\right\}}{\left\{(3c_f)^{\frac{1}{3}}\,(n^{-1} \log n)^{\frac{2}{3}}\right\}} \to 1\,. \tag{5.12}$$

Since both terms in the numerator are positive, we have

$$m_n^{*-2}/\left\{(3c_f)^{\frac{1}{3}}\,(n^{-1} \log n)^{\frac{2}{3}}\right\} \to c\,,$$

where c is a non-negative finite constant.

We now show c is non-zero by contradiction. If $c = 0$, the $(n/\log n)^{-\frac{1}{3}} = o(m_n^*)$ and $m_n^* n^{-1} \log(n/m_n^*)/\{(3c_f)^{\frac{1}{3}} (n^{-1} \log n)^{\frac{2}{3}}\} \to 1$, but these cannot both hold at the same time.

Therefore, $m_n^* = b_n(f)\bar{m}_n$. By (4.12), the constant $b_n(f)$ equals $(3c_f)^{\frac{1}{3}}$ $(1 + o(1))$. This completes the proof of (iv).

(v) This is similar to the proof of (iv), but using Theorem 2.1 instead of Theorem 2.2. □

Proof of Theorem 2.3

(i) Combining (iv) and (iii) of Lemma 5.8 gives,

$$m_n^* = \bar{m}_n(1 + o(1)) = (3c_f\, n/\log n)^{\frac{1}{3}} (1 + o(1)).$$

(ii) Combining (v), (iv) of Lemma 5.8 gives

$$\hat{m}_n = \bar{m}_n(1 + o(1)) = m_n^*(1 + o(1)).$$

(iii) Combining (i) and (iii) of Lemma 5.8 gives

$$\begin{aligned} \frac{1}{n} E_f \log f^n(x^n)/g_{m_n^*,n}(x^n) &= L(m_n^*) \\ &= S(\bar{m}_n)(1 + o(1)) \\ &= \tfrac{1}{2} (3c_f)^{\frac{1}{3}} (n^{-1} \log n)^{\frac{2}{3}} (1 + o(1)). \end{aligned}$$

(iv) Combining (i) and (ii) of Lemma 5.8 gives,

$$\begin{aligned} \frac{1}{n} \log f^n(x^n)/g_{\hat{m}_n,n}(x^n) &= R(\hat{m}_n) \\ &= S(\bar{m}_n)(1 + o(1)) \\ &= L(m_n^*)(1 + o(1)) \\ &= \{E_f \log f^n(x^n)/g_{m_n^*,n}(x^n)\}(1 + o(1)). \end{aligned}$$ □

Now we turn to the proof of consistency if the true density f is uniform. If $f \equiv 1$, then $R(m) = -\log g_{m,n}(x^n)$ and $S(m) = \frac{m}{2} \log \frac{n}{m}$. For any fixed $\delta > 0$, let

$$T_n(\delta) = \{|R(m) - S(m)| < \delta S(m) \quad \text{for all} \quad m \in B_n = [1, n^{\epsilon_2}]\}.$$

Lemma 5.9.

$$\sum_{n=1}^{\infty} P\big(T_n^c(\delta)\big) < \infty .$$

Proof. It follows from $f \equiv 1$ that $\phi_k \equiv m^{-1}$ and $f(x_t) \equiv 1$. Thus $R(m) - S(m)$ can be written as

$$\begin{aligned} -\sum_{t=1}^{n} \log\{\hat{f}_{m,n}(x_t)/f_m(x_t)\} - \tfrac{1}{2}\sum_{k=1}^{m} \log N_k/n\,\phi_k + O(m) \\ = -\sum_{k=1}^{m} N_k \log \frac{N_k}{n\,\phi_k} - \tfrac{1}{2}\sum_{k=1}^{m} \log N_k/n\,\phi_k + O(m). \end{aligned} \tag{5.13}$$

By the remark after Lemma 5.4

$$\sum_{n=1}^{\infty} P\Big(\max_{m\in B_n} \Big|\sum_{k=1}^{m} N_k \log \frac{N_k}{n\,\phi_k}\Big| > \epsilon m\Big) < \infty$$

and

$$\sum_{n=1}^{\infty} P\Big(\max_{m\in B_n} \Big|\sum_{k=1}^{m} \log \frac{N_k}{n\,\phi_k}\Big| > \epsilon m\Big) < \infty$$

which imply that $\sum_{n=1}^{\infty} P\big(T_n^c(\delta)\big) < \infty$ by (5.13) and $S(m) = O(m \log n)$. □

Proof of Theorem 2.4

It suffices to show that $\hat{m}_n = 1$ on the set $T_n(\delta)$ for some δ, since it then follows that

$$\sum_{n=1}^{\infty} P(\hat{m}_n \neq 1) \leq \sum_{n=1}^{\infty} P\big(T_n^c(\epsilon)\big) < \infty .$$

which by the Borel-Cantelli lemma yields $\hat{m}_n \to 1$. Since $\hat{m}_n$ is integer-valued, it follows that, for almost all sample paths x^∞, there is a $n(x^\infty) > 0$ such the $\hat{m}_n = 1$ for $n > n(x^\infty)$.

A direct calculation of the derivative of $S(m)$ with respect to m shows that $S(m)$ is strictly increasing as m increases on B_n provided that $n > e^{1/\epsilon_2}$. Moreover, for any $\delta > 0$, $\frac{1+\delta}{1-\delta}\, S(1) < S(2)$ is equivalent to $\delta < \frac{\log n - \log 2}{3\log n - 2\log 2}$, which holds for $\delta = \frac{1}{4}$ if $n > 4$. For this particular δ, when $n > \max(4,\, e^{1/\epsilon_2})$, we have on $T_n(\delta)$,

$$|R(m) - S(m)| < \delta S(m)$$

for all m in B_n.

Taking the values 1 and $\hat{m}_n$ as m in the above inequality gives $(1-\delta)\,S(\hat{m}_n) \leq R(\hat{m}_n)$ and $R(1) < S(1)(1+\delta)$. Together with the fact that $R(\hat{m}_n) \leq R(1)$ because $\hat{m}_n$ is the minimizer of $R(m)$, the two inequalilties yield

$$S(\hat{m}_n) \leq S(1)\left(\frac{1+\delta}{1-\delta}\right) < S(2)$$

where the last inequality holds by the choice of δ. Since $S(m)$ is strictly increasing on B_n, this implies $\hat{m}_n = 1$ and this completes the proof. □

6. The a.s. achievability of the minimax rate

To get the sample stability and the minimax rate at the same time, we use a predictive density f^* defined on x^n. Intuitively, f^* is constructed by updating the binwidth of predictive histograms only at the beginning of each block of size Δn, where we decompose x^n into blocks. The first block has size $\bar{n} = \left[\Delta n \cdot n^{\frac{4}{15}}\right]$, longer than the rest, which have equal size Δn. More precisely,

$$f^*(x^n) = \prod_{t=1}^{n} g_{m_{t-1}}\left(x_t \mid x^{t-1}\right) = g_{m_1}(x^{\bar{n}}) \prod_{i=n^{\frac{4}{15}}}^{n/\Delta n} g_{m_i}(x_t)$$

where $g_{m_{t-1}}(x_t)$ is the histogram based on x^{t-1}, with m_{t-1} equal bins, $\Delta n = n^{\frac{2}{3}+\delta}$, $0 < \delta < \frac{1}{15}$, and

$$m_{t-1} = \begin{cases} \left[\left(\Delta n \cdot n^{\frac{4}{15}}\right)^{\frac{1}{3}}\right] & \text{if} \quad 1 \leq t \leq \left[\Delta n \cdot n^{\frac{4}{15}}\right] \\ \left[\left(\Delta n \cdot i\right)^{\frac{1}{3}}\right] & \text{if} \quad (i-1)\Delta n \leq t \leq i \cdot \Delta n \quad \text{and} \quad i \geq n^{\frac{4}{15}}. \end{cases}$$

We deal with the singularity problem the same way as before, i.e., we add to the beginning of each block, $y_1, \ldots, y_{m_i}$, where y_j uniform on $\left[\frac{j-1}{m_i}, \frac{j}{m_i}\right]$, and the $\{y_j\}$ are mutually independent.

Alternatively, we could take Hall and Hannan's stochastic complexity version of the density, which will give rise to the same predictive density. Thus, f^* can be written as

$$f^*(x^t) = g_{m_1}(x^1) \prod_{i=n^{\frac{4}{15}}}^{n/\Delta n} g_{m_i}(x^i \mid x^1, \ldots, x^{i-1}),$$

where $x^1 = (x_1, \ldots, x_{[\Delta n \cdot n^{\frac{4}{15}}]}) = x^n$ and $x^i = (x_{(i-1)\Delta n+1}, \ldots, x_{i\Delta n})$ for $i = n^{\frac{4}{15}}, \ldots, n/\Delta n$.

To be quite precise, we should take $\left[\frac{n}{\Delta n}\right] - \left[n^{\frac{4}{15}}\right]$ blocks of size Δn from the end of the data string, and put the remainder in the first block. This will increase the size of the first block by a size of smaller order than $\left[\Delta n \cdot n^{\frac{4}{15}}\right]$.

Note that by Theorem 2.1, if $m_1 = \left[\Delta n \cdot n^{\frac{4}{15}}\right]^{\frac{1}{3}}$, then as $n \to \infty$

$$-\log g_{m_1}(x^1)/f_{m_1}(x^1)$$
$$= \Delta n \cdot n^{\frac{4}{15}} \left[\tfrac{1}{2} c_f \frac{1}{m_1^2} + \tfrac{1}{2} \frac{m_1}{\Delta n \cdot n^{\frac{4}{15}}} \log \frac{\Delta n \cdot n^{\frac{4}{15}}}{m_1}\right](1 + o(1)) \quad \text{a.s.}$$

Moreover,

$$g_{m_i}(x^i \mid x^1, \ldots, x^{i-1}) = \frac{g_{m_i}(x^1, \ldots, x^{i-1}, x^i)}{g_{m_i}(x^1, \ldots, x^{i-1})},$$

which gives

$$L_i = -\log \frac{g_{m_i}(x^i \mid x^1, \ldots, x^{i-1})}{f_{m_i}(x^i)} = -\log g_{m_i}(x^1, \ldots, x^i)$$
$$+ \log g_{m_i}(x^1, \ldots, x^{i-1}) + \log f_{m_i}(x^i) . \quad (6.1)$$

We will expand the summands of the expansion

$$\log[f^n(x^n)/f^*(x^n)] = \sum_{i=1}^{n} \{L_i + \log[f(x^i)/f_{m_i}(x^i)]\}$$

for each i, using Taylor's theorem, cf. the proof of Theorem 2.1, and then use results from the Appendix to complete the proof. Denote by

$$N_{k,i} = \sum_{x \in (x^1, \ldots, x^i)} 1_{I_{k,m_i}}(x),$$

the count of x's in the first $i-1$ blocks falling into I_{k,m_i}, and by

$$M_{k,i} = \sum_{x \in x^i} 1_{I_{k,m_i}}(x)$$

the count of x's in the ith block falling into I_{k,m_i}. For simplicity we will suppress the i in $N_{k,i}$ and $M_{k,i}$, since we will be expanding (6.1) for a particular i.

Let $r_i = (i-1)\,\Delta n$. By equation (2.2)

$$\begin{aligned} L_i &= -\log \frac{g_{m_i}(x^i \mid x^1, \ldots, x^{i-1})}{f_{m_i}(x^i)} \\ &= -\sum_{k=1}^{m_i} (N_k + M_k) \log \frac{N_k + M_k}{i\Delta n} \cdot m_i + \tfrac{1}{2} \sum_{k=1}^{m_i} \log(N_k + M_k) \\ &\quad + m_i \log \frac{i\Delta n}{m_i} + R^1_{m,n} + \sum_{k=1}^{m_i} N_k \log \frac{N_k}{r_i} m_i - \tfrac{1}{2} \sum_{k=1}^{m_i} \log N_k \\ &\quad - m_i \log \frac{r_i}{m_i} + R^2_{m,n} + \sum_{k=1}^{m_i} M_k \log \phi_k \cdot m_i \end{aligned}$$

where $\log 0 = 0$, and

$$\begin{aligned} \left|R^1_{m,n}\right| &< \sum_{N_k > 0} \frac{1}{N_k}, \\ \left|R^2_{m,n}\right| &< \sum_{N_k + M_k > 0} \frac{1}{N_k + M_k}. \end{aligned}$$

Note that $-\log g_{m_i}(x^i | x^1, \ldots, x^{i-1})$ can be written $L_i - \log f_{m_i}(x^i)$.

Proposition 6.1. *We have the following expression uniformly in* $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$, *as* $n \to \infty$

$$L_i = \tfrac{1}{2} \frac{m_i}{i} [1 + o(1)] \quad \text{a.s.}$$

The proof consists of a series of lemmas.

Rewrite L_i as

$$\begin{aligned} &-\sum_{k=1}^{m_i} N_k \log \frac{N_k + M_k}{N_k} \cdot \frac{i-1}{i} - \sum_{k=1}^{m_i} M_k \log \frac{N_k + M_k}{i\Delta n} \frac{m_k}{\phi_k m_i} \\ &\qquad - \tfrac{1}{2} \sum_{k=1}^{m_i} \log \frac{N_k + M_k}{i} \cdot \frac{i-1}{N_k} + \tfrac{1}{2} m_i \log \frac{i}{i-1} + R_i \end{aligned}$$

where $R_i = R^1_{m,n} + R^2_{m,n}$ satisfies

$$R_i \le \sum_{N_k < 0} \frac{1}{N_k} + \sum_{N_k + M_k > 0} \frac{1}{N_k + M_k} \le 2 \sum_{N_k > 0} \frac{1}{N_k},$$

and express L_i as

$$I_1 + I_2 + I_3 + I_4 + R_i$$

where

$$I_1 = -\sum_{k=1}^{m_i} N_k \log \frac{N_k + M_k}{N_k} \cdot \frac{i}{i-1},$$

$$I_2 = -\sum_{k=1}^{m_i} M_k \log \frac{N_k + M_k}{i\Delta n\phi_k},$$

$$I_3 = -\tfrac{1}{2} \sum_{k=1}^{m_i} \log \frac{N_k + M_k}{i} \cdot \frac{i-1}{N_k},$$

and

$$I_4 = \tfrac{1}{2}\, m_i \log \frac{i}{i-1}.$$

It is easy to see that $I_4 = \frac{1}{2}\frac{m_i}{i}(1 + o(1))$ uniformly for $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$ as $n \to \infty$.

Lemma 6.1. *Uniformly in* $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right] = \left[n^{\frac{4}{15}}, n^{\frac{1}{3}-\delta}\right]$, *as* $n \to \infty$

$$I_1 = \tfrac{1}{2}\frac{m_i}{i}(1 + o(1)) \quad \text{a.s.}$$

Proof. We can expand

$$\begin{aligned} I_1 &= -\sum_{k=1}^{m_i} N_k \log \frac{N_k + M_k}{N_k} \cdot \frac{(i-1)\Delta n\phi_k}{i\Delta n\phi_k} \\ &= -\sum_{k=1}^{m_i} N_k \log\left(1 + \frac{N_k + M_k - i\Delta n\phi_k}{i\Delta n\phi_k}\right) + \sum_{k=1}^{m_i} N_k \log\left(1 + \frac{N_k - r_i\phi_k}{r_i\phi_k}\right) \\ &= T_1 + T_2\,, \text{say}\,. \end{aligned}$$

We now introduce the notation

$$\Delta M_k = M_k - EM_k = M_k - \Delta n\phi_k$$

$$\Delta N_k = N_k - EN_k = N_k - r_i\phi_k\,,$$

and expand T_1, T_2 into four terms as follows:

$$T_j = L_j + Q_j + C_j + Re_j\,, \quad j = 1, 2\,.$$

Then, after some simplification,

$$\begin{aligned} L_1 + L_2 &= -\sum_{k=1}^{m_i} N_k \frac{N_k + M_k - i\Delta n\phi_k}{i\Delta n\phi_k} + \sum_{k=1}^{m_i} N_k \frac{N_k - r_i\phi_k}{r_i\phi_k} \\ &= \sum_{k=1}^{m_i} \frac{(N_k - r\phi_k)^2}{i(i-1)\Delta n\phi_k} - \sum_{k=1}^{m_i} \frac{(N_k - r\phi_k)(M_k - \Delta n\phi_k)}{i\Delta n\phi_k} \\ &= \frac{m_i}{i}(1 + o(1)) \quad \text{a.s.} \end{aligned}$$

by (6) and (12) in the Appendix.

Similarly we can get

$$Q_1 - Q_2 = -\frac{m_i}{2i}(1 + o(1)) \quad \text{a.s.}$$

and both C_1 and C_2 are smaller order than m_i/i, hence $Re_i = o(m_i/i)$ and $I_i = (m_i/2i)(1 + o(1))$ a.s. as asserted.

In fact $Q_1 - Q_2$ can be simplified to

$$\begin{aligned}
&= -\tfrac{1}{2}\sum_{k=1}^{m_i} \Delta N_k^3 \left[\frac{2i-1}{i^2(i-1)^2\Delta n^2 \phi_k^2}\right] + \tfrac{1}{2}\sum_{k=1}^{m_i} \Delta N_k^2 \frac{-2i+1}{i^2(i-1)\Delta n}\frac{1}{\phi_k} \\
&\quad + \tfrac{1}{2}\sum_{k=1}^{m_i} \frac{2\Delta N_k^2 \Delta M_k + \Delta N_k \Delta M_k^2}{i^2 \Delta n^2 \phi_k^2} \\
&\quad + \tfrac{1}{2}\sum_{k=1}^{m_i} \frac{(i-1)2\Delta N_k \Delta M_k + \Delta M_k^2}{i^2 \Delta n^2 \phi_k} \\
&= o\left(\frac{m_i}{i}\right) - \frac{m_i}{i}(1 + o(1)) + \tfrac{1}{2}\frac{m_i}{i}(1 + o(1)) \quad \text{a.s.}
\end{aligned}$$

by (8), (6), (13), (14), (12) and (7) of the Appendix,

$$= -\tfrac{1}{2}\frac{m_i}{i}(1 + o(1)) \quad \text{a.s.} \quad \text{as} \quad n \to \infty .$$

Furthermore,

$$\begin{aligned}
C_1 &= -\tfrac{1}{3}\sum_{k=1}^{m_i} \Delta N_k \frac{[\Delta N_k + \Delta M_k]^3}{i^3 \Delta n^3 \phi_k^3} - \tfrac{1}{3}\sum_{k=1}^{m_i} \frac{r_i(\Delta N_k + \Delta M_k)^3}{i^3 \Delta n^3 \phi_k^3} \\
&= o\left(\frac{m_i}{i}\right) \quad \text{a.s.} \quad n \to \infty
\end{aligned}$$

by (10), (8), (21), (20), (8), (13), (14) and (9) of the Appendix. Similarly $C_2 = o(m_i/i)$ a.s. as $n \to \infty$. □

Lemma 6.2. *Uniformly in* $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$, *as* $n \to \infty$,

$$I_2 = -\frac{m_i}{2i}(1 + o(1)) \quad \text{a.s.}$$

Proof. $I_2 = -\sum_{k=1}^{m_i} M_k \log \frac{N_k + M_k}{i\Delta n \phi_k} = L + Q + C + R$, where

$$L = -\sum_{k=1}^{m_i} M_k \cdot \frac{\Delta N_k + \Delta M_k}{i\Delta n \phi_k}$$

$$= -\sum_{k=1}^{m_i} \frac{\Delta N_k \cdot \Delta M_k}{i\Delta n\phi_k} - \sum_{k=1}^{m_i} \frac{\Delta M_k^2}{i\Delta n\phi_k}$$
$$= -\frac{m_i}{i}(1 + o(1)) \quad \text{a.s.}$$

by (12) and (7) of the Appendix. Similarly

$$Q = \tfrac{1}{2} \sum_{k=1}^{m_i} \frac{\Delta M_k \Delta N_k^2 + 2\Delta M_k^2 \Delta N_k + \Delta M_k^3}{i^2 \Delta n^2 \phi_k^3}$$
$$+ \tfrac{1}{2} \frac{1}{i^2 \Delta n} \sum_{k=1}^{m_i} \frac{[\Delta N_k + \Delta M_k]^2}{i^2 \Delta n \phi_k}$$
$$= \tfrac{1}{2} \frac{m_i}{i}(1 + o(1)) \quad \text{a.s.} \quad \text{as} n \to \infty$$

by (13), (14), (9), (6), (7) and (12) of the Appendix.

Finally,

$$C = -\tfrac{1}{3} \sum_{k=1}^{m_i} M_k \cdot \left[\frac{\Delta N_k + \Delta M_k}{i\Delta n\phi_k}\right]^3$$
$$= -\tfrac{1}{3} \sum_{k=1}^{m_i} \frac{\Delta M_k \left[\Delta N_k^3 + 3\Delta N_k^2 \Delta M_k + 3\Delta N_k \Delta M_k^2 + \Delta M_k^3\right]}{i^3 \Delta n^2 \phi_k^3}$$
$$- \tfrac{1}{3} \sum_{k=1}^{m_i} \frac{\left[\Delta N_k^3 + 3\Delta N_k^2 \Delta M_k + 3\Delta N_k \Delta M_k^2 + \Delta M_k^3\right]}{i^3 \Delta n^2 \phi_k^2}$$
$$= o\left(\frac{m_i}{i}\right) \quad \text{a.s.} \quad \text{as} n \to \infty$$

by (19), (21), (20), (11), (8), (13), (14) and (9) of the Appendix. Thus $R = o\left(\frac{m_i}{i}\right)$ a.s. as $n \to \infty$ and we conclude that

$$I_2 = -\frac{m_i}{2i}(1 + o(1)) \quad \text{a.s.}$$

Lemma 6.3. *Uniformly in* $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$, *as* $n \to \infty$,

$$I_3 = o\left(\frac{m_i}{i}\right) \quad \text{a.s.}$$

Proof. The proof is similar to the proofs of Lemma 6.1 and Lemma 6.2 and will be omitted.

The only thing remaining in the proof of Proposition 6.1 is to bound the remainder term, which has the order

$$O\Big(\sum_{N_k>0}\frac{1}{N_k}\Big).$$

Lemma 6.4. *Uniformly in* $i \in \big[\cdot n^{\frac{4}{15}}, \frac{n}{\Delta n}\big]$, *as* $n \to \infty$,

$$\sum_{N_k>0}\frac{1}{N_{k,i}} = o\Big(\frac{m_i}{i}\Big) \quad \text{a.s.}$$

Proof. Let $A_{n,i} = \Big\{ \max_{1\le k\le m_i} |N_{k,i} - r_i\phi_{ki}| < (i\Delta n)^{\frac{1}{3}+\epsilon}\Big\}$ for some small $\epsilon > 0$. Then on $A_{n,i}$,

$$\begin{aligned}
O < \sum_{N_{k,i}>0}\frac{1}{N_{k,i}} &= \sum_{N_{k,i}>0}\frac{1}{N_{k,i} - r_i\phi_k + r_i\phi_k} \\
&= \sum_{N_{k,i}>0}\frac{1}{[1+\Delta N_{k,i}/r_i\phi_k]r_i\phi_k} \\
&\le \sum_{i=1}^{m_i}\frac{1}{r_i\phi_k}\left[1+\frac{\Delta N_{k,i}}{r_i\phi_k}\right] \\
&\le \sum_{i=1}^{m_i}\frac{1}{r_i\phi_k}\left(1+\frac{(i\Delta n)^{\frac{1}{3}+\epsilon}}{i\Delta n}m_i\right) \\
&\le \sum_{i=1}^{m_i}\frac{1}{r_i\phi_k}\left(1+\frac{1}{(i\Delta n)^{\frac{1}{3}-\epsilon}}\right) \\
&= \left(\sum_{i=1}^{m_i}\frac{1}{r_i\phi_k}\right)(1+o(1)) \\
&= O\Big(\frac{m_i^2}{i\Delta n}\Big)(1+o(1)) \\
&= \frac{m_i}{i}\cdot O\Big(\frac{(i\Delta n)^{\frac{1}{3}}}{\Delta n}\Big) \\
&= \frac{m_i}{i}\cdot O\Big[\frac{[n/\Delta n]^{\frac{1}{3}}}{\Delta n^{\frac{2}{3}}}\Big] = \frac{m_i}{i}O\Big(\frac{n^{\frac{1}{3}}}{\Delta n}\Big) = \frac{m_i}{i}O\Big(\frac{1}{n^{\frac{1}{3}+\delta}}\Big) = o(\frac{m_i}{i}).
\end{aligned}$$

To complete the proof, it is enough to show that

$$\sum_n \sum_{i=n^{\frac{4}{15}}\cdot\Delta n}^{n/\Delta n} P(A_{n,i}^c) < \infty. \tag{6.2}$$

Note that

$$P(A_{n,i}^c) \le \sum_{k=1}^{m_i} P\big(|N_{k,i} - r_i\phi_k| > (i\Delta n)^{\frac{1}{3}+\epsilon}\big)$$
$$\le \sum_{k=1}^{m_i} \frac{E|N_{k,i} - r_i\phi_k|^{2\ell}}{(i\Delta n)^{\frac{2\ell}{3}+\epsilon\ell}}$$
$$\le \sum_{k=1}^{m_i} \frac{O(r_i\phi_k)^\ell}{(i\Delta n)^{\frac{2\ell}{3}+2\epsilon\ell}} = O\Big(\frac{m_i}{(i\Delta n)^{2\epsilon\ell}}\Big) = O\Big(\frac{1}{(i\Delta n)^{2\epsilon\ell-\frac{1}{3}}}\Big),$$

since

$$\frac{(r_i\phi_k)^\ell}{i\Delta n^{\frac{2}{3}\ell}} = O\frac{\big[i\Delta n \cdot \frac{1}{m_i}\big]^\ell}{i\Delta n^{\frac{2}{3}\ell}} = O(1).$$

As $i\Delta n > \Delta n \cdot n^{\frac{14}{15}} = n^{\frac{4}{15}+\delta}$, taking ℓ large enough will guarantee that (6.2) holds.□

Proof of Proposition 6.1. Combine Lemmas 6.1–6.4.

Proposition 6.2. *If $f \not\equiv 1$, then uniformly in $i \in \big[n^{\frac{4}{15}}, \frac{n}{\Delta n}\big]$ as $n \to \infty$*

$$\log \frac{f(x^i)}{f_{m_i}(x^i)} = \sum_{t=(i-1)\Delta n+1}^{i\Delta n} \log \frac{f(x_t)}{f_{m_i}(x_t)}$$
$$= \Delta n\left[\frac{c_f}{2m_i^2} + o\Big(\frac{1}{m_i^2} + \frac{m_i}{i}\Big)\right] \quad \text{a.s.}$$

Proof. Let $Y_{t,m_i} = \log \frac{f(x_t)}{f_{m_i}(x_t)}$, as in the proof of Lemma 5.4. By the method used there we can obtain

$$P\Big(\Big|\sum_{t=r_i+1}^{i\Delta n} Y_{t,m_i} - EY_{t,m_i}\Big| > \epsilon\Big(\frac{\Delta n}{m_i^2} + \frac{\Delta n m_i}{i}\Big)\Big)$$
$$\le \exp\big(-O(\Delta n^{\frac{2}{3}})\big) = \exp\big(-O(n^{\frac{4}{9}})\big).$$

Therefore

$$\sum_n P\Big(\max_i \big|\sum Y_{t,m_i} - \Delta n\, E\, Y_{t,m_i}\big| > \epsilon\Big(\frac{\Delta n}{m_i^2} + \frac{\Delta n m_i}{i}\Big)\Big)$$
$$\le \sum_{n=1}^{\infty} \cdot \sum_{i=n^{\frac{4}{15}}}^{n/\Delta n} \exp\big(-O(n^{\frac{4}{9}})\big)$$
$$\le \sum_{n=1}^{\infty} \frac{n}{\Delta n} \cdot \exp\big(-O(n^{\frac{4}{9}})\big) < \infty.$$

By Borel-Cantelli lemma, Proposition 6.2 is proved, provided that $E_f \log \frac{f}{f_{m_i}} = \frac{1}{2}\frac{c_f}{m_i^2}(1+o(1))$, which is true by Lemma 5.6. □

Proposition 6.3. *If $f \not\equiv 1$, then as $n \to \infty$,*

$$\log \frac{f^n(x^n)}{f^*(x^n)} = \tfrac{3}{2}(1+c_f)\,n^{\frac{1}{3}}(1+o(1)) \quad \text{a.s.}$$

Proof. Let the size of the first block be denoted $\bar{n} = n^{\frac{4}{15}} \times \Delta n = n^{\frac{4}{15}+\delta} \ll n$, and write $m_1 = \left[\frac{\bar{n}}{\log \bar{n}}\right]^{\frac{1}{3}}$. By Theorem 2.1 above, as $n \to \infty$ we have

$$\log \frac{f(x^1)}{f^*(x^1)} = \bar{n}\left[\tfrac{1}{2}\,c_f \frac{1}{m_1^2} + \tfrac{1}{2}\frac{m_1}{\bar{n}} \log \frac{\bar{n}}{m_1}\right](1+o(1)) \quad \text{a.s.}$$

By Propositions 6.1 and 6.2 above, for $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$, if $m_i = [(i\Delta n)^{\frac{1}{3}}]$

$$\begin{aligned}\log \frac{f(x^i)}{f^*(x^i)} &= \frac{m_i}{2i} + \frac{\Delta n\, c_f}{2m_i^2}[1+o(1)] \\ &= \tfrac{1}{2}(1+c_f)\left[\frac{\Delta n^{\frac{1}{3}}}{i^{\frac{2}{3}}}\right](1+o(1)) \quad \text{a.s.}\end{aligned}$$

Therefore

$$\begin{aligned}\log \frac{f(x^i)}{f^*(x^i)} &= \bar{n}\left[\tfrac{1}{2}\,c_f \frac{1}{m_1^2} + \frac{m_1}{\bar{n}} \log \frac{\bar{n}}{m_1}\right](1+o(1)) \\ &\quad + \sum_{i=n^{\frac{4}{15}}}^{n/\Delta n} \tfrac{1}{2}(1+c_f)\left[\frac{\Delta n^{\frac{1}{3}}}{i^{\frac{2}{3}}}\right](1+o(1)) \\ &= O\big(\bar{n}^{\frac{1}{3}}(\log \bar{n})^{\frac{2}{3}}\big) + \tfrac{1}{2}(1+c_f)(\Delta n)^{\frac{1}{3}} \cdot 3 \cdot \left[\left[\frac{n}{\Delta n}\right]^{\frac{1}{3}} - n^{\frac{4}{15}+\frac{1}{3}}\right](1+o(1)) \\ &= o(n^{\frac{1}{3}}) + \tfrac{3}{2}(1+c_f)\,n^{\frac{1}{3}}[1+o(1)] = \tfrac{3}{2}(1+c_f)\,n^{\frac{1}{3}}(1+o(1)) \quad \text{a.s.}\end{aligned}$$

The justification for our summing up over i while keeping the $o(1)$ term is that $o(1)$ is uniform in i. The constant $\frac{3}{2}(1+c_f)$ is not the best possible, but without knowing c_f, we might as well take $m_i = [i\Delta n]^{\frac{1}{3}}$ instead of $\hat{A}_f[i\Delta n]^{\frac{1}{3}}$ with $\hat{A}_f$ the adaptive best constant. The best constant is $\frac{3}{2} \cdot \left[\frac{c_f}{4}\right]^{\frac{1}{3}}$.

This completes the proof of Proposition 6.3 and hence of Theorem 4.2. □

7. Acknowledgements

We would like to thank Jorma Rissanen for his sustained interest in this topic, Lucien Birgé for a particularly helpful discussion, and the Statistics Research Section at the Australian National University for their hospitality when the manuscript was being prepared. Norma Chin is especially to be thanked for her extraordinarily fast and accurate typing. Partial support from the NSF (grant DMS 8802378) to both of us, is also gratefully acknowledged.

8. References

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

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

Barron, A.R. and Cover, T.M. (1989). Minimum complexity density estimation. *IEEE Trans. Inf. Th.*, (To appear).

Breiman, L.A. and Freedman, D.F. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78**, 131-136.

Clarke, B.S. (1989). *Asymptotic Cumulative Risk and Bayes Risk Under Entropy, with Applications.* PhD thesis, University of Illinois at Urbana–Champaign.

Davisson, L.D. (1983). Minimax noiseless universal coding for Markov sources. *IEEE Trans. Inf. Th.* **29**, 211-215.

Dawid, A.P. (1984). Present position and potential developments: some personal views, statistical theory, the prequential approach. *J. Roy. Statist. Soc. Ser. B* **147**, 278–292.

Dawid, A.P. (1989). Prequential data analysis. In *Issues and Controversies in Statistical Inference*. Essays in Honor of D. Basu's 65th birthday, eds. M. Ghosh and P.K. Pathak.

Devroye, L. (1987). *A Course in Density Estimation.* Progress in Probability and Statistics **14**, Birkhauser.

Freedman, D.A. and Diaconis, P. (1981). On the histogram as a density estimator: L^2 theory. *Zeit. für Wahr. und Ver. Geb.* **57**, 453–475.

Hall, P. and Hannan, E.J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika* **74**, 705–714.

Hamming, R.W. (1986). *Coding and Information Theory*. Englewood Cliffs, N.J.: Prentice-Hall.

Hannan, E.J., Cameron, M.A. and Speed, T.P. (1990). Estimating spectra and prediction variance. (manuscript)

Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **15**, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity (with discussion). *J. Roy. Statist. Soc. Ser. B* **49**, 223–239, 252–265.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Books.

Rissanen, J., Speed, T.P. and Yu, Bin (1989). Density estimation by stochastic complexity. (submitted to *IEEE Trans. IT*).

Speed, T.P. and Yu, Bin (1989). Stochastic complexity in model selection. (submitted to *Ann. Statist.*)

Stone, C.J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. *Recent Advances in Statistics*, 393– 406. Academic Press, Inc.

Stone, C.J. (1985). An asymptotic optimal histogram selection rule. *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. II, L.M. Le Cam and R.A. Ohshen, eds., 513-520. Wadsworth, Inc.

APPENDIX

The purpose of this Appendix is to derive almost sure asymptotic expansions for the certain expressions encountered in §5, 6, such as

$$T_1 = \sum_{k=1}^{m} (N_k - n\,\phi_k)^2\,,$$

$$T_2 = \sum_{k=1}^{m} \frac{(N_k - n\,\phi_k)}{n\,\phi_k}\,,$$

where the $\{N_k\}$ are multinomial random variables with parameters $\{\phi_k\}$ and $\sum N_k = n$. Note that $c_0\,m^{-1} \le \phi_k \le c_1\,m^{-1}$.

Stone (1985) gives moment bounds on T_1 through Poissonization as follows.

Lemma 1 (Stone (1985)). *For each positive integer q, there is an universal constant a_q such that*

$$E\left\{\sum_{k=1}^{m}((N_k - n\,\phi_k)^2 - n)\right\}^{2q} \le c_q\,n^q(1 + n^q\,m^{-2q})\,.$$

We will employ the same Poissonization tecqhnique to obtain bounds on moments of T_2 and other expressions. We first find bounds for independent Poisson variables instead of multinomials, and then observe that Stone (1985) has a general argument (Lemma 4) which guarantees that bounds having the same rate (as $n \to \infty$) will hold for multinomials.

Lemma 2. *Let M_ℓ ($\ell = 1, 2, \ldots, m$) be independent Poisson random variables with means λ_ℓ such that $0 < \lambda = \sum_{\ell=1}^{m} \lambda_\ell < \infty$. Set $M = \sum_{\ell=1}^{m} M_\ell$, $p_\ell = \lambda_\ell/\lambda$ and $c_0\,\bar{p} \le p_\ell \le c_1\,\bar{p}$ ($c_0, c_1 > 0$). For each positive integer q there is a finite positive universal constant A'_q such that*

$$E\left\{\sum_{\ell=1}^{m} \frac{(M_\ell - M\,p_\ell)}{p_\ell}\right\}^{2q} \le a'_q\,\lambda^q\,\bar{p}^{-2q}\,.$$

Corollary. *Let* $\{N_k\}$ *be multinomial random variables with parameters* $\{\phi_k\}$ *such that* $\sum_{k=1}^m \phi_k = 1$, $c_0\, m^{-1} \le \phi_k \le c_1\, m^{-1}$, *and* $\sum_{k=1}^m N_k = n$. *Then for each integer* q, *there is an universal constant* a_q *such that*

$$E\left\{ \sum_{k=1}^{m} \frac{(N_k - n\,\phi_k)}{\phi_k} \right\}^{2q} \le a_q\, n^q\, m^{2q}\,.$$

Stone's (1985) argument for his Lemma 4 is general enough to derive the corollary from Lemma 2 above. We refer to the paper for details.

Proof of Lemma 2. We begin by rewriting

$$\sum_{\ell=1}^{m} \frac{(M_\ell - M\,p_\ell)}{p_\ell} = \lambda \sum_{\ell=1}^{m} \frac{(M_\ell - \lambda_\ell)}{\lambda_\ell} + m(M-\lambda)\,.$$

By the inequality $(a+b)^{2q} \le 2^{2q}\,(a^{2q} + b^{2q})$, we get

$$E\left\{ \sum_{\ell=1}^{m} \frac{(M_\ell - M\,p_\ell)}{p_\ell} \right\}^{2q} \le (2\lambda)^{2q}\, E\left\{ \sum_{\ell=1}^{m} \frac{(M_\ell - M\,\lambda_\ell}{\lambda_\ell)} \right\}^{2q} + m^{2q}\, E(M-\lambda)^{2q}\,. \quad (1)$$

Setting $T = \sum_{\ell=1}^m \frac{M_\ell - M\,\lambda_\ell}{\lambda_\ell}$, the $2q$-th moments of T can be written as a sum of its cumulants,

$$E\,T^{2q} = \sum c_{(\ell_1,\ldots,\ell_j)} \prod_{i=1}^{j} \kappa_{2\ell_i}(T) \quad (2)$$

where the sum is taken over all the partitions of q such that $\sum_{i=1}^j \ell_i = q$ ($j < q$, $\ell_i > 0$), since $\kappa_1(T) = 0$.

Because the M_ℓ are independent Poisson random variables, it follows that

$$\begin{aligned} \kappa_{2\ell_i}(T) &= \sum_{\ell=1}^{m} \kappa_{2\ell_i}\left(\frac{M_\ell - M\,\lambda_\ell}{\lambda_\ell} \right) \\ &= \frac{\lambda_\ell}{\lambda_\ell^{2\ell_i}} \le \lambda(\lambda\bar{p}/c_0)^{-2\ell_i}\,. \end{aligned}$$

Here and below we will use the same notation for possibly different constants. Substituting this last bound in (2), we find that

$$\begin{aligned} E\,T^{2q} &\le c_q \sum \prod_{\ell=1}^{j} \lambda(\lambda\,\bar{p}/c_0)^{-2\ell_i} \\ &\le c_q \sum \lambda^j\, \lambda^{-2q}\, \bar{p}^{-2q} \le c_q\, \lambda^{-q}\, \bar{p}^{-2q}\,. \end{aligned} \quad (3)$$

The second term in (1) is $E(M-\lambda)^{2q}$, which is known to be a polynomial of order q. Thus, there is a constant c_q' such that

$$E(M-\lambda)^{2q} \le c_q' \lambda^q .$$

The last bound together with (1) and (3) give

$$E\Big\{ \sum_{\ell=1}^{m} \frac{(M_\ell - M p_\ell)}{p_\ell} \Big\}^{2q} \le 2^{2q}\big(c_q \bar{p}^{-2q} + c_q' m^{2q} \lambda^q\big) \le a_q' \bar{p}^{-2q} \lambda^q .$$

A similar Poissonization argument will give the following

Lemma 3. *Suppose that N is a binomial random variable with mean np. Then for any integer $q > 0$, there is an $a_q > 0$ such that*

$$E(N-np)^{2q} \le a_q (np)^q .$$

Now we are ready to derive the a.s. expansions of T_1 and T_2.

Lemma 4. *Under the assumptions of the Corollary to Lemma 2, uniformly in $m \in A_n$ as $n \to \infty$*

(i) $$T_1 = \sum_{k=1}^{m} (N_k - n\phi_k)^2 = n(1+o(1)) \quad \text{a.s.},$$

(ii) $$T_2 = \sum_{k=1}^{m} \frac{(N_k - n\phi_k)}{n\phi_k} = o(m) \quad \text{a.s.}$$

Proof. (i) It suffices to show that $\max_{m\in A_n} \Big| \sum_{k=1}^{m} ((N_k - n\phi_k)^2 - n) \Big| = o(n)$ a.s. as $n \to \infty$. Note that, for any $\epsilon > 0$,

$$P\Big(\max_{m\in A_n} \Big| \sum_{k=1}^{m} ((N_k - n\phi_k)^2 - n) \Big| \ge \epsilon n \Big) \le \sum_{m\in A_n} P\Big(\Big| \sum_{k=1}^{m} ((N_k - n\phi_k)^2 - n) \Big| \ge \epsilon n \Big) . \tag{4}$$

By Markov's inequality and Lemma 1,

$$\begin{aligned} P\Big(\Big| \sum_{k=1}^{m} ((N_k - n\phi_k)^2 - n) \Big| \ge \epsilon n \Big) &\le c_q \epsilon^{-2q} m^{-2q} n^q (1 + n^q m^{-2q}) \\ &\le c_q \epsilon^{-2q} \big(n^{-q} + n^{-2\epsilon_1 q} \big) . \end{aligned} \tag{5}$$

Combining (4) and (5), we obtain

$$P\Big(\max_{m\in A_n}\Big|\sum_{k=1}^{m}((N_k-n\,\phi_k)^2-n)\Big|\geq \epsilon n\Big)\leq \sum_{m\in A_n} c_q\,\epsilon^{-2q}\big(n^{-q}+n^{\epsilon_2-2\epsilon_2 q}\big)$$
$$\leq c_q\,\epsilon^{-2q}\big(n^{\epsilon_2-q}+n^{2\epsilon_2-2\epsilon_2 q}\big)\,.$$

If we take

$$q>\max\left(\epsilon_2+1,\frac{1+\epsilon_2}{2\epsilon_2}\right),$$

the above series converges. Hence by the Borel-Cantelli Lemma,

$$\max_{m\in A_n}\Big|\sum_{k=1}^{m}((N_k-n\,\phi_k)^2-n)\Big|=o(n)\quad\text{a.s.}$$

Part (ii) is proved similarly using the corollary to Lemma 2 and taking $q>\epsilon_2+1$.□

Similarly we can obtain the following expansions, needed in the proofs of Propositions 6.1 and 6.2. to which we refer for the notation. Uniformly in $i\in\left[n^{\frac{4}{15}},\frac{n}{\Delta n}\right]$, as $n\to\infty$, with $r_i=(i-1)\Delta n$,

$$\sum_{k=1}^{m_i}\frac{(N_k-r_i\phi_k)^2}{\phi_k}=r_i(m_i-1)(1+o(1))\quad\text{a.s.}\tag{6}$$

$$\sum_{k=1}^{m_i}\frac{(M_k-\Delta n\phi_k)^2}{\phi_k}=\Delta n(m_i-1)(1+o(1))\quad\text{a.s.}\tag{7}$$

$$\sum_{k=1}^{m_i}\frac{(N_k-r_i\phi_k)^3}{\phi_k^2}=O\big(r_i\cdot m_i^2\big)\quad\text{a.s.}\tag{8}$$

$$\sum_{k=1}^{m_i}\frac{(M_k-\Delta n\phi_k)^3}{\phi_k^2}=O\big(\Delta n\cdot m_i^2\big)\quad\text{a.s.}\tag{9}$$

$$\sum_{k=1}^{m_i}\frac{(N_k-r_i\phi_k)^4}{\phi_k^3}=r_i^2\left(\sum_{k=1}^{m_i}\frac{1}{\phi_k}\right)(1+o(1))\quad\text{a.s.}\tag{10}$$

$$\sum_{k=1}^{m_i}\frac{(M_k-\Delta n\phi_k)^4}{\phi_k^3}=\Delta n^2\left(\sum_{k=1}^{m_i}\frac{1}{\phi_k}\right)(1+o(1))\quad\text{a.s.}\tag{11}$$

$$\sum_{k=1}^{m_i}\frac{1}{\phi_k}\,\Delta M_k\,\Delta N_k=o\big(i^{\frac{1}{2}}\cdot\Delta n\cdot n^{y}\big)\quad\text{a.s.}\tag{12}$$

$$\sum_{k=1}^{m_i}\frac{1}{\phi_k^2}\,\Delta M_k\,\Delta N_k^2=o\big(i^{\frac{3}{2}}\,\Delta n^{\frac{3}{2}}\cdot m_i\cdot n^{y}\big)\quad\text{a.s.}\tag{13}$$

$$\sum_{k=1}^{m_i}\frac{1}{\phi_k^2}\,\Delta M_k^2\,\Delta N_k=o\big(i^{\frac{1}{2}}\,\Delta n^{\frac{3}{2}}\cdot m_i\cdot n^{y}\big)\quad\text{a.s.}\tag{14}$$

$$\sum_{k=1}^{m_i} \frac{\Delta N_k}{\phi_k} = o\big(m_i \cdot (i\Delta n)^{\frac{1}{2}} \cdot n^y\big) \quad \text{a.s.} \tag{15}$$

$$\sum_{k=1}^{m_i} \frac{\Delta M_k}{\phi_k} = o\big(m_i \cdot (\Delta n)^{\frac{1}{2}} \cdot n^y\big) \quad \text{a.s.} \tag{16}$$

$$\sum_{k=1}^{m_i} \Delta N_k^2 = (i-1)\Delta n(1+o(1)) \quad \text{a.s.} \tag{17}$$

$$\sum_{k=1}^{m_i} \Delta M_k^2 = \Delta n(1+o(1)) \quad \text{a.s.} \tag{18}$$

$$\sum_{k=1}^{m_i} \frac{\Delta M_k \cdot \Delta N_k^3}{\phi_k^3} = o\big(\Delta n^2 \cdot m_i^{\frac{3}{2}} \cdot i \cdot n^y\big) \quad \text{a.s.} \tag{19}$$

$$\sum_{k=1}^{m_i} \frac{\Delta N_k \cdot \Delta M_k^3}{\phi_k^3} = o\big(\Delta n^2 \cdot m_i^{\frac{3}{2}} \cdot i^{\frac{1}{2}} n^y\big) \quad \text{a.s.} \tag{20}$$

$$\sum_{k=1}^{m_i} \frac{\Delta N_k^2 \Delta M_k^2}{\phi_k^3} = O\big(i \cdot m_i^{-1}\, \Delta n^2\big) \quad \text{a.s.} \tag{21}$$

In (12), (13), (14), (15) and (16) $y > 0$ is arbitrary, whereas in (19) and (20) we only need the results for some $y > 0$.

Let T_i be any of the expressions on the left-hand side of (6),...,(21). As with Lemma 4 (i) and (ii) above, the approach we adopt is to obtain a moment bound via Poissonization on $P(|T_i - ET_i| > a_{n,i})$, where $(a_{n,i})$ is a suitable sequence of numbers. The a.s. assertion then follows using the Borel-Cantelli lemma. To avoid a tedious repetition of arguments, we select one of (6),...,(21) and only prove that, leaving the rest to the reader. Take (20):

$$T_i = \sum_{k=1}^{m_i} \frac{\Delta N_k\, \Delta M_k^3}{\phi_k^3}\,.$$

Since ΔN_k and ΔM_k are independent, $ET_i = 0$.

Lemma 5. *If $\{N_k'\}$ and $\{M_k'\}$ are independent arrays of independent Poisson random variables with $EN_k' = \lambda\phi_k$, $EM_k' = \mu\phi_k$, $k = 1,\ldots,m_i$, and $T_i' = \sum_{k=1}^{m_i} \phi_k^{-3}\, \Delta N_k'(\Delta M_k')^3$, then there exists a constant a_q such that*

$$E|T_i'|^{2q} \le a_q\, m_i^q\, \bar{p}^{-2q}\, \lambda^q\, \mu^{3q}\,.$$

Proof. As in the proof of Lemma 2 above,

$$E|T_i'|^{2q} = \sum_{\ell_1+\ldots\ell_j=q} c_{\ell_1,\ldots,\ell_j} \prod_{i=1}^{j} \kappa_{2\ell_i}(T_i') ,$$

where

$$\begin{aligned}\kappa_{2\ell_i}(T_i) &= \sum_{k=1}^{m_i} \kappa_{2\ell_i}\left(\frac{\Delta N_k' \, \Delta M_k^3}{\phi_k}\right) \\ &= \sum_{k=1}^{m_i} \phi_k^{-6\ell_i} \, \kappa_{2\ell_i}(\Delta N_k' \, \Delta M_k'^3) .\end{aligned}$$

Now

$$\begin{aligned}\kappa_{2\ell_i}(\Delta N_k' \, \Delta M_k'^3) &= \sum_{t_1+\ldots+t_j=\ell_i} c_{t_1,\ldots,t_j} \prod_{s=1}^{j} E(\Delta N_k' \, \Delta M_k'^3)^{2t_s} \\ &\le a_{\ell_i} \sum \prod_{s=1}^{j} (\lambda p_k)^{t_s} \, (\mu p_k)^{3t_s} \\ &\le a'_{\ell_i} \, \lambda^{\ell_i} \, \mu^{3\ell_i} \, p_k^{4\ell_i} .\end{aligned}$$

Thus

$$\begin{aligned}\kappa_{2\ell_i}(T_i') &\le a'_{\ell_i} \sum_{k=1}^{m_i} p_k^{-6\ell_i} \, \lambda^{\ell_i} \, \mu^{3\ell_i} \, p_k^{4\ell_i} \\ &\le a'_{\ell_i} \, \lambda^{\ell_i} \, \mu^{3\ell_i} \, m_i \, \bar{p}^{-2\ell_i} .\end{aligned}$$

Substituting this into the expression for $E|T_i'|^{2q}$ we get

$$\begin{aligned}E|T_i'|^{2q} &\le a_q \sum_{\ell_1+\ldots+\ell_j=q} \prod_{i=1}^{j} m_i \, \bar{p}^{-2\ell_i} \, \lambda^{\ell_i} \, \mu^{3\ell_i} \\ &\le a_q \sum m_i^j \, \bar{p}^{-2q} \, \lambda^q \, \mu^{3q} \\ &\le a_q \, m_i^q \, \bar{p}^{-2q} \, \lambda^q \, \mu^{3q} .\end{aligned}$$

□

Using Lemma 4 of Stone (1985) we are able to conclude that for our original multinomial sum T_i we have

Corollary.

$$E|T_i|^{2q} \le a'_q \, m_i^{3q} \, i^q (\Delta n)^{4q} .$$

Proof of (20). For any $\epsilon > 0$ we have

$$P\Big(\max_i \left|a_{n,i}^{-1} T_i\right| > \epsilon\Big) \le \sum_i (a_{n,i}\,\epsilon)^{-2q}\, E|T_i|^{2q}$$
$$\le \sum_i a'_q\, a_{n,i}^{-2q}\, \epsilon^{-2q}\, m_i^{3q}\, i^q\, \Delta n^{4q}\,.$$

If we take $a_{n,i} = m_i^{\frac{3}{2}} \cdot i^{\frac{1}{3}} \cdot \Delta n^2 \cdot n^y$ for some $y > 0$, the last expression is easily seen to be bounded by $a'_q \sum_k n^{-2yq} \le a'_q \frac{n}{\Delta n} n^{-2yq}$. Now we can take q large enough to ensure that this last series converges, and so deduce from the Borel-Cantelli lemma that as $n \to \infty$,

$$\max_i \left|a_{n,i}^{-1} T_i\right| = o(1) \quad \text{a.s.}$$

where $a_{n,i}$ is as described, i.e., uniformly in $i \in \left[n^{\frac{4}{15}}, \frac{n}{\Delta n}\right]$

$$T_i = o\big(m_i^{\frac{3}{2}} \cdot i^{\frac{1}{3}} \cdot \Delta n^2 \cdot n^y\big) \quad \text{a.s.}$$

This completes our illustrative proof. □

For those expressions amongst (6),...,(21) for which the left-hand side has non-zero expectation, the following fact is needed to get bounds on their moments.

Suppose that $\{N_k\}$, $\{M_k\}$ are similarly indexed independent multinomial arrays with parameters $(n, \{\phi_k\})$ and $(m, \{\phi_k\})$ respectively, that $\{N'_k\}$, $\{M'_k\}$ are independent arrays of independent Poissons with means $\{\lambda p_k\}$ and $\{\mu p_k\}$ respectively, where $c_0\bar{p} \le p_k \le c_1\bar{p}$ and $c'_0\bar{\phi} \le \phi_k \le c'\bar{\phi}$ for all k, and that

$$\lambda, \mu \to \infty\,, \quad \bar{p} \to 0\,, \quad \lambda\bar{p}\,, \quad \mu\bar{p} \to \infty\,,$$
$$n, m \to \infty\,, \bar{\phi} \to 0\,, \quad n\bar{\phi}\,, \quad m\bar{\phi} \to \infty\,.$$

Then as these limits are approached, for any integer j there is a constant c_j such that

$$E(N_k - n\phi_k)^j = c_j\, n^{q_j}\, \phi_k^{q_j}\,(1 + o(1))$$
$$E(N'_k - \lambda p_k)^j = c_j\, \lambda^{q_j}\, p_k^{q_j}\,(1 + o(1))$$

where $q_j = \frac{1}{2}\, j$ if j is even, and $q_j = \frac{1}{2}\,(j-1)$ if j is odd. This implies that under the same conditions,

$$E(N_k - n\phi_k)^j (M_k - m\phi_k)^\ell = c_j c_\ell n^{q_j} \phi_k^{q_j + q_\ell} m^{q_\ell} (1 + o(1))$$

and

$$E(N'_k - \lambda\phi_k)^j (M'_k - m\phi_k)^\ell = c_j c_\ell \lambda^{q_j} \phi_k^{q_j + q_\ell} \mu^{q_\ell} (1 + o(1))$$

since the $\{N_k\}$ and $\{M_k\}$ are independent, and similarly for $\{N'_k\}$ and $\{M'_k\}$.