Записки научных
семинаров ПОМИ
Том 363, 2009 г.

F. El Haje Hussein, Yu. Golubev

ON ENTROPY ESTIMATION BY M-SPACING METHOD

Abstract. The m-spacing method is a very popular statistical tool in entropy estimation and in goodness of fit testing. In this text, we focus on the case, where the underlying probability density may have an unbounded support or may vanish and show that under mild conditions the m-spacing entropy estimators have standard Gaussian limits.

1. Introduction

Suppose we are given a random vector $\mathbf{X}^n = (X_1, \dots, X_n)^T$ whose components are i.i.d. random variables with an unknown probability density $p(x)$, $x \in \mathbb{R}^1$. Our goal is to estimate the entropy

$$H(p) = -\int_{-\infty}^{\infty} \log[p(x)]p(x)\,dx \tag{1}$$

with the help of $\mathbf{X}^n$. This statistical problem can be viewed as a particular case of the general theory of nonlinear functional estimation developed in [11, 12, 7–9]. However, the entropy is a very specific functional enabling to construct its nontrivial estimators and the study of these estimators is the main theme of the present paper.

Undoubtedly, principal difficulties in entropy estimation result from two obvious facts:

- $H(p)$ is the nonlinear functional of p;
- $\log[p(x)] \to -\infty$ as $p(x) \to 0$.

Indeed, if our target functional would be linear, for instance,

$$L(p) = \int_{-\infty}^{\infty} l(x)p(x)\,dx,$$

then we could use the standard δ-method providing the following estimate

$$L(\mathbf{X}^n) = L(\widehat{p}), \quad \text{with} \quad \widehat{p}(x, \mathbf{X}^n) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i),$$

where $\delta(x)$ is the standard Dirac δ-function. Since

$$L(\mathbf{X}^n) = \frac{1}{n}\sum_{i=1}^{n} l(X_i),$$

statistical analysis of this estimate is simple because it is related to the standard probabilistic facts like the law of large numbers and the central limit theorem (for mathematical details concerning the δ - method we refer interested readers to [15]).

Although, the naive idea to plug-in $\widehat{p}(x, \mathbf{X}^n)$ in $H(p)$ obviously fails, but it prompts a structure of reasonable entropy estimators

$$\widehat{H}(\mathbf{X}^n) = -\frac{1}{n}\sum_{i=1}^{n} \log[\widetilde{p}(X_i, \mathbf{X}^n)],$$

where $\widetilde{p}(X_i, \mathbf{X}^n)$ is a probability density estimator. This idea reduces entropy estimation to recovering probability density. Intuitively, it is clear that the popular plug-in principle saying that *good density estimators result in good nonlinear functional estimators*, does not work in this situation. This phenomenon admits a simple explanation because $\widehat{H}(\mathbf{X}^n)$ is based on the averaging with respect to the empirical measure. So, it finds out that the variance of the density estimator is not determinative because of the averaging, but its bias is really important. Therefore to construct good entropy estimators, we need density estimators with finite variances but very small biases.

Such density estimators can be obtained by numerical differentiation of the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leqslant x\}.$$

For instance, we can use the following density estimator

$$\widehat{p}_1(X_{(i)}, \mathbf{X}^n) = \frac{F_n[X_{(i+1)}] - F_n[X_{(i)}]}{X_{(i+1)} - X_{(i)}} = \frac{1}{n[X_{(i+1)} - X_{(i)}]},$$

where $X_{(1)} \leqslant \cdots \leqslant X_{(n)}$ stands the nondecreasing permutation of $X_1, \ldots, X_n$. This estimator admits a natural generalization

$$\widehat{p}_m(X_{(i)}, \mathbf{X}^n) = \frac{F_n[X_{(i+m)}] - F_n[X_{(i)}]}{X_{(i+m)} - X_{(i)}} = \frac{m}{n[X_{(i+m)} - X_{(i)}]},$$

which is called m-spacing density estimator. With this density estimator we arrive at the so-called m-spacing entropy estimator

$$\widehat{H}_m(\mathbf{X}^n) = \frac{1}{n} \sum_{i=1}^{n-m} \log \frac{n[X_{(i+m)} - X_{(i)}]}{m}. \tag{2}$$

The idea of this estimator goes back to [4]. We also refer interested readers to the paper [5] which contains a reach bibliography on spacings. Notice also that a slightly modified version of this estimator provides powerful methods for testing True Random Numbers Generators [6].

In spite of the simplicity of $\widehat{H}_m(\mathbf{X}^n)$, its statistical analysis is not banal. In this paper, we use the famous Pyke theorem [13] to compute statistical characteristics of this estimator.

Theorem 1. *Let $U_1, \ldots, U_n$ be independent random variables uniformly distributed on $[0,1]$ and $e_1, \ldots, e_{n+1}$ be independent exponentially distributed random variables $\mathbf{P}\{e_i > x\} = \exp(-x)$. Then*

$$U_{(k)} \stackrel{\mathcal{D}}{=} \sum_{i=1}^{k} e_i \Big/ \sum_{i=1}^{n+1} e_i. \tag{3}$$

Let us look at heuristically how does this theorem work. Denote by $F(x)$ the distribution function of X_i. To compute the limit distribution of $\widehat{H}_m(\mathbf{X}^n)$, note that

$$F(X_{(i)}) = U_{(i)}$$

and therefore we can write

$$\begin{aligned} U_{(i+m)} - U_{(i)} &= F(X_{(i+m)}) - F(X_{(i)}) \\ &= p(X_{(i)})[X_{(i+m)} - X_{(i)}] \times \frac{F(X_{(i+m)}) - F(X_{(i)})}{p(X_{(i)})[X_{(i+m)} - X_{(i)}]}. \end{aligned}$$

Thus we have

$$X_{(i+m)} - X_{(i)} = \frac{U_{(i+m)} - U_{(i)}}{p(X_{(i)})} \times \frac{p(X_{(i)})[X_{(i+m)} - X_{(i)}]}{F(X_{(i+m)}) - F(X_{(i)})}$$

and substituting this in (2) and using Pyke's theorem, we obtain

$$\widehat{H}_m(\mathbf{X}^n) = -\frac{1}{n}\sum_{i=1}^{n-m}\log[p(X_{(i)})] + \frac{1}{n}\sum_{i=1}^{n-m}\log\left[\frac{1}{m}\sum_{k=i}^{i+m-1}e_k\right] - \left(1-\frac{m}{n}\right)\log\left[\frac{1}{n}\sum_{k=1}^{n+1}e_k\right] + \frac{\epsilon_n}{\sqrt{n}}, \tag{4}$$

where

$$\epsilon_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n-m}\log\frac{F[X_{(i+m)}] - F[X_{(i)}]}{p(X_{(i)})[X_{(i+m)} - X_{(i)}]}.$$

Statistical properties of the first term at the right-hand side of (4) can be easily analyzed by standard probabilistic methods. Indeed, by the central limit theorem,

$$\sqrt{n}\left[-\frac{1}{n}\sum_{i=1}^{n-m}\log[p(X_{(i)})] - H(p)\right] \approx -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\log[p(X_i)] + H(p)\} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2(p)), \tag{5}$$

where

$$\sigma^2(p) = \int_{-\infty}^{\infty}\log^2(p(x))p(x)\,dx - H^2(p).$$

Next note that the second and the third terms in (4) do not depend on $p(\cdot)$. It is easy to see, using the Taylor formula, that

$$\mathbf{E}\log\left[\frac{1}{n}\sum_{k=1}^{n+1}e_k\right] = O\left(\frac{1}{n}\right)$$

and

$$\mathbf{E}\frac{1}{n}\sum_{i=1}^{n-m}\log\left[\frac{1}{m}\sum_{k=i}^{i+m-1}e_k\right] = \Psi(m) - \log(m) + O\left(\frac{1}{n}\right),$$

where $\Psi(m)$ is digamma function

$$\Psi(m) = \mathbf{E}\log\left[\sum_{k=1}^{m}e_k\right] = \frac{1}{\Gamma(m)}\int_0^{\infty}\log(x)x^{m-1}\exp(-x)\,dx = \frac{\Gamma'(m)}{\Gamma(m)}.$$

Note also that, by the Taylor formula,

$$\sqrt{n}\log\left[\frac{1}{n}\sum_{k=1}^{n+1} e_k\right] \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(e_i - 1) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

To analyze the second term at the right-hand side of (4), we decompose it as follows

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n-m}\left\{\log\left[\sum_{k=i}^{i+m-1} e_k\right] - \Psi(m)\right\}$$
$$= \frac{1}{\sqrt{m}}\sum_{l=1}^{m}\frac{1}{\sqrt{n/m}}\sum_{s=0}^{n/m}\left\{\log\left[\sum_{k=l+ms}^{l+(m+1)s-1} e_k\right] - \Psi(m)\right\}.$$

By the central limit theorem,

$$\frac{1}{\sqrt{n/m}}\sum_{s=0}^{n/m}\left\{\log\left[\sum_{k=l+ms}^{l+(m+1)s-1} e_k\right] - \Psi(m)\right\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2(m)),$$

where

$$\Sigma^2(m) = \frac{1}{\Gamma(m)}\int_0^\infty \log^2(x) x^{m-1}\exp(-x)\,dx = \frac{\Gamma''(m)}{\Gamma(m)} - \Psi^2(m) = \Psi'(m).$$

and therefore the third term at (4) has also a Gaussian limit. The calculation of the variance of this Gaussian low is not very easy and we refer the interested reader to [3], where it was proved that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n-m}\left\{\log\left[\sum_{k=i}^{i+m-1} e_k\right] - \Psi(m)\right\} - \left(1 - \frac{m}{n}\right)\sqrt{n}\log\left[\frac{1}{n}\sum_{k=1}^{n+1} e_k\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2), \tag{6}$$

where

$$\Sigma^2 = (2m^2 - 2m + 1)\Psi'(m) - 2m + 1.$$

Therefore combining (6) and (5), we can hope that $\sqrt{n}[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)]$ converges in distribution to $\mathcal{N}(0, \sigma^2(p) + \Sigma^2)$ as $n \to \infty$. To prove this fact it remains to check that

- the remainder term ϵ_n is small, i.e., $\lim\limits_{n\to\infty} \mathbf{E}\epsilon_n^2 = 0$;
- the first term in (4) is weakly correlated with the others.

All these facts can be easily proved if we suppose the density $p(\cdot)$ has a compact support and strictly bounded from zero over its support (see [1, 2]), and the derivative $p'(x)$ is bounded over the support. For instance, to prove the first assertion, we obtain by the Taylor formula

$$\frac{F[X_{(i+m)}] - F[X_{(i)}]}{p(X_{(i)})[X_{(i+m)} - X_{(i)}]} = 1 + \frac{p'(\xi_i)}{2p(X_{(i)})}[X_{(i+m)} - X_{(i)}],$$

where ξ_i belongs to $[X_{(i)}, X_{(i+m)}]$. Thus, we have using that $\log(1+x) \leqslant x$, $x \geqslant 0$,

$$\epsilon_n \leqslant \frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \log[1 + C(X_{(i+m)} - X_{(i)})]$$
$$\leqslant \frac{C}{\sqrt{n}} \sum_{i=1}^{n-m} [(X_{(i+m)} - X_{(i)})] \leqslant \frac{Cm}{\sqrt{n}}. \quad (7)$$

Here and later, on C denotes a generic constant. To bound ϵ_n from below, notice by the Taylor formula

$$\epsilon_n \geqslant \frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \log[1 - C(X_{(i+m)} - X_{(i)})]$$
$$\geqslant -\frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \frac{C[X_{(i+m)} - X_{(i)}]}{1 - C[X_{(i+m)} - X_{(i)}]}$$
$$\geqslant -\frac{Cm}{\sqrt{n}\{1 - C\max_i [X_{(i+m)} - X_{(i)}]\}}.$$

Next, since $p(x)$ is assumed to be strictly bounded from zero, we have

$$U_{(i+m)} - U_{(i)} = F[X_{(i+m)}] - F[X_{(i)}]$$
$$= p(\xi_i)[X_{(i+m)} - X_{(i)}] \geqslant C[X_{(i+m)} - X_{(i)}],$$

and therefore obviously

$$X_{(i+m)} - X_{(i)} \leqslant C[U_{(i+m)} - U_{(i)}] \leqslant Cm \max_i [U_{(i+1)} - U_{(i)}].$$

Thus, using Pyke's theorem, we get for some $C > 0$

$$\mathbf{P}\left\{\epsilon_n \leqslant -\frac{2Cm}{\sqrt{n}}\right\} \leqslant \mathbf{P}\left\{\max_i [U_{(i+1)} - U_{(i)}] \geqslant \frac{1}{2C}\right\} \leqslant n \exp\left[-n/(2C)\right].$$

Combining this with (7), we see that with a high probability $\epsilon_n \leqslant Cm/\sqrt{n}$.

Obviously, these arguments fail when the density $p(x)$ has an unbounded support or vanishes. However, in spite of the fact that unbounded support densities are widely used in statistical practice, there are only a few papers dealing with this case. We mention here for instance [10] and [14], where the standard entropy estimator was modified to deal with vanishing densities.

The main goal in this paper it to demonstrate that the standard entropy estimator has the same Gaussian limit for vanishing probability densities as well.

2. Root n consistency of the m-spacing entropy estimator

For a given sequence $r(n) \geqslant 1$ define the family of balls in $\mathbb{R}^1$ by

$$\mathbb{B}_r^n(x) = \left\{y : \ |F(x) - F(y)| \leqslant \frac{r(n)}{n}\right\}.$$

Denote also

$$D_r^n(x) = \sup_{y \in \mathbb{B}_r^n(x)} \left\{\frac{p'^2(y)}{p^2(y)} + \frac{|p''(y)|}{p(y)}\right\}. \tag{8}$$

If the second derivative of $p(x)$ does not exist, then we set $D_r^n(x) = \infty$. Let

$$\mathbb{Q}_{r,R}^n = \left\{x : \ p(x) \geqslant \sqrt{D_r^n(x)}\frac{R(n)}{n}\right\}, \tag{9}$$

where $R(n) \geqslant r(n)$ is a given sequence.

The main statistical fact in this section is provided by the following lemma.

Lemma 1. *Let $r(n) = 5\log(n)$ and $R(n) \geqslant 7r(n)$. Assume that*

- *the number of connected components of $\mathbb{Q}_{r,R}^n$ is bounded uniformly in n;*
- *for some $\varepsilon > 0$,*

$$\mathbf{E}|X_1 - X_2|^{\pm\varepsilon} < C, \quad \mathbf{E}p^{\pm\epsilon}(X_1) < C, \tag{10}$$

then

$$\mathbf{E}\epsilon_n^2 \leqslant C\log(n)\sqrt{n}\int\limits_{x\notin\mathbb{Q}_{r,R}^n} p(x)\,dx + \frac{C\log(n)}{n^{3/2}}\int\limits_{x\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(x)}{p(x)}\,dx + \frac{C\log(n)}{\sqrt{n}}. \quad (11)$$

The root-n consistency the m - spacing entropy estimator follows now immediately from Lemma 1.

Theorem 2. *Suppose that the conditions of Lemma* 1 *hold true. Assume also that*

$$\limsup_{n\to\infty}\left[\log(n)\sqrt{n}\int\limits_{x\notin\mathbb{Q}_{r,R}^n} p(x)\,dx + \frac{\log(n)}{n^{3/2}}\int\limits_{x\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(x)}{p(x)}\,dx\right] = 0. \quad (12)$$

Then

$$\limsup_{n\to\infty} n\mathbf{E}\left[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)\right]^2 \leqslant C.$$

2.1. Auxiliary lemmas

Lemma 2. *Assume that* $R(n) \geqslant 7r(n)$ *and* $x \in \mathbb{Q}_{r,R}^n$. *Then for any* $\xi_1, \xi_2 \in \mathbb{B}_r^n(x)$

$$\left|\frac{p(\xi_1)}{p(\xi_2)} - 1\right| < \frac{1}{4}.$$

Proof. Denote

$$\bar{d} = \sup_{\xi_1,\xi_2\in\mathbb{B}_r^n(x)}\frac{p(\xi_1)}{p(\xi_2)}.$$

With the help of the Taylor expansion we get for some $\xi_3, \xi_4 \in \mathbb{B}_r^n(x)$

$$\frac{p(\xi_1)}{p(\xi_2)} - 1 = \frac{p(\xi_1)-p(\xi_2)}{p(\xi_2)} = \frac{p'(\xi_3)(\xi_1-\xi_2)}{p(\xi_2)} = \frac{p'(\xi_3)[F(\xi_1)-F(\xi_2)]}{p(\xi_2)p(\xi_4)}$$
$$= \frac{1}{p(x)}\frac{p'(\xi_3)}{p(\xi_3)}[F(\xi_1)-F(\xi_2)]\frac{p(\xi_3)p(x)}{p(\xi_2)p(\xi_4)}. \quad (13)$$

Next we use that

$$\frac{p'(\xi_3)}{p(\xi_3)} \leqslant \sqrt{D_r^n(x)}, \quad \frac{p(\xi_3)p(x)}{p(\xi_2)p(\xi_4)} \leqslant \bar{d}^2$$

and therefore, in view of (9),

$$\bar{d} - 1 \leqslant \bar{d}^2 \frac{r(n)}{R(n)}.$$

It is easy to check with a simple algebra that

$$\bar{d} \leqslant \frac{R(n)}{2r(n)} - \sqrt{\frac{R^2(n)}{4r^2(n)} - \frac{R(n)}{r(n)}} = 2 \Big/ \left(1 + \sqrt{1 - \frac{4r(n)}{R(n)}}\right)$$

and therefore $\bar{d} \leqslant 1 + 0.21$, when $r(n)/R(n) \leqslant 1/7$.

Similar arguments can be used to get the lower bound for

$$\underline{d} = \inf_{\xi_1, \xi_2 \in \mathbb{B}_r^n(x)} \frac{p(\xi_1)}{p(\xi_2)}.$$

By (13), we obviously obtain

$$\underline{d} - 1 \geqslant -\bar{d}^2 \frac{r(n)}{R(n)} = 1 - \bar{d},$$

thus proving the lemma. □

Our principal idea to control the remainder term is related to

Lemma 3. *Assume that $R(n) \geqslant 7r(n)$, then for any $x \in \mathbb{Q}_{r,R}^n$ and any $y \in \mathbb{B}_r^n(x)$*

$$\left| 2 \log \frac{F(y) - F(x)}{p(x)(y-x)} - \log \frac{p(y)}{p(x)} \right| \leqslant C[F(x) - F(y)]^2 \frac{D_r^n(x)}{p^2(x)}. \tag{14}$$

Proof. It is based on the well-known formula

$$F(y) - F(x) = \int_x^y p(u)\, du = \frac{p(y) + p(x)}{2}(y-x) - \frac{p''(\xi)}{12}(y-x)^3,$$

where $\xi \in [x, y]$. Therefore we obviously obtain for some $\xi_1 \in [x, y]$

$$\begin{aligned} \left[\frac{F(y) - F(x)}{p(x)(y-x)}\right]^2 &= \left[1 + \frac{p(y) - p(x)}{2p(x)} - \frac{p''(\xi)}{12p(x)}(y-x)^2\right]^2 \\ &= 1 + \frac{p(y) - p(x)}{p(x)} - \frac{p''(\xi)}{6p(x)}(y-x)^2 + \frac{p'2(\xi_1)}{4p^2(x)}(y-x)^2 \\ &\quad + \left[\frac{p''(\xi)}{12p(x)}(y-x)^2\right]^2 - \frac{p'(\xi_1)p''(\xi)}{12p^2(x)}(x-y)^3. \end{aligned} \tag{15}$$

The last term at the right-hand side can be controlled with the help of $ab \leqslant a^2/2 + b^2/2$ as follows

$$\frac{|p'(\xi_1)||p''(\xi)|}{12p^2(x)}(x-y)^3 \leqslant \frac{p'^2(\xi_1)}{8p^2(x)}(y-x)^2 + \frac{1}{72}\left[\frac{p''(\xi)}{p(x)}(y-x)^2\right]^2.$$

Therefore substituting this in (15) and using that $F(y)-F(x) = p(\xi_3)(y-x)$, for some $\xi_3 \in [x,y]$, we get

$$\begin{aligned}\left|\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2 - \frac{p(y)}{p(x)}\right| \\ \leqslant \frac{\big[F(x)-F(y)\big]^2}{p^2(x)}\left\{\frac{|p''(\xi)|p(x)}{4p^2(\xi_3)} + \frac{p'^2(\xi_1)}{2p^2(\xi_3)}\right\}\end{aligned} \tag{16}$$

provided that

$$\big[F(x)-F(y)\big]^2\frac{|p''(\xi)|}{p(x)p^2(\xi_3)} \leqslant 1. \tag{17}$$

To finish the proof, notice that by the Taylor formula

$$\begin{aligned}\left|\log\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2 - \log\frac{p(y)}{p(x)}\right| &= \left|\log\frac{p(x)}{p(y)}\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2\right| \\ &= \left|\log\left\{1 + \frac{p(x)}{p(y)}\left[\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2 - \frac{p(y)}{p(x)}\right]\right\}\right| \\ &\leqslant 4\frac{p(x)}{p(y)}\left[\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2 - \frac{p(y)}{p(x)}\right]\end{aligned} \tag{18}$$

if

$$\frac{p(x)}{p(y)}\left|\left[\frac{F(y)-F(x)}{p(x)(y-x)}\right]^2 - \frac{p(y)}{p(x)}\right| \leqslant \frac{3}{4}.$$

In view of (17), the last display is equivalent to the following one

$$\frac{p(x)}{p(y)}\frac{\big[F(x)-F(y)\big]^2}{p^2(x)}\left\{\frac{|p''(\xi)|p(x)}{4p^2(\xi_3)} + \frac{p'^2(\xi_1)}{2p^2(\xi_3)}\right\} \leqslant \frac{3}{4} \tag{19}$$

and therefore it remains to check conditions (17) and (19). To do that, notice that by Lemma 2, for all $\xi_1, \xi_2 \in \mathbb{B}_r^n(x)$

$$\frac{p(\xi_1)}{p(\xi_2)} < \frac{5}{4}. \tag{20}$$

Therefore we have (see (17))

$$\big[F(x)-F(y)\big]^2\frac{|p''(\xi)|}{p(x)p^2(\xi_3)}=\frac{p(x)p(\xi)}{p^2(\xi_3)}\frac{[F(x)-F(y)]^2}{p^2(x)}\frac{|p''(\xi)|}{p(\xi)}$$
$$\leqslant\frac{25}{16}\frac{r^2(n)}{n^2}\frac{D_r^n(x)}{p^2(x)}\leqslant\frac{25}{16}\frac{r^2(n)}{R^2(n)}. \quad (21)$$

Thus (17) holds for all $R(n)\geqslant\sqrt{2}r(n)$.

To analyze (19) we use the same arguments. We have by (20)

$$\frac{p(x)}{p(y)}\frac{[F(x)-F(y)]^2}{p^2(x)}\left\{\frac{|p''(\xi)|p(x)}{4p^2(\xi_3)}+\frac{p'^2(\xi_1)}{2p^2(\xi_3)}\right\}$$
$$\leqslant\frac{5}{4}\frac{r^2(n)}{n^2p^2(x)}\left\{\frac{|p''(\xi)|}{p(\xi)}\frac{p(x)p(\xi)}{4p^2(\xi_3)}+\frac{p'^2(\xi_1)}{p^2(\xi_1)}\frac{p^2(\xi_1)}{2p^2(\xi_3)}\right\}$$
$$\leqslant\frac{125}{64}\frac{r^2(n)}{n^2}\frac{D_r^n(x)}{p^2(x)}\leqslant\frac{125}{64}\frac{r^2(n)}{R^2(n)}.$$

Noticing that this inequality holds true for $R(n)\geqslant\sqrt{2}r(n)$, we finish the proof. □

Lemma 4. *Let $q>0$ be a given integer. Suppose* (10) *holds, then there exists a constant $C(\varepsilon)$ such that*

$$\Big\{\mathbf{E}\max_{i=1,\dots,n}\log^{2q}[p(X_{(i)}]\Big\}^{1/(2q)}\leqslant C(\varepsilon)[\log(n)+q], \quad (22)$$

$$\Big\{\mathbf{E}\max_{1\leqslant i<j\leqslant n}\log^{2q}[X_{(j)}-X_{(i)}]\Big\}^{1/(2q)}\leqslant C(\varepsilon)[\log(n)+q] \quad (23)$$

and

$$\Big\{\mathbf{E}\max_{i=1,\dots,n-m}\log^{2q}[F(X_{(i+m)})-F(X_{(i)})]\Big\}^{1/(2q)}\leqslant C(m)[\log(n)+q]. \quad (24)$$

Proof. Note that for any integer $q\geqslant 1$, function $L(x)=\log^{2q}(x+\mathrm{e}^{2q-1})$, $x>0$ is concave, since

$$L''(x)=\frac{2q\log^{2q-2}\big(x+e^{2q-1}\big)}{\big(x+e^{2q-1}\big)^2}\big[2q-1-\log\big(x+e^{2q-1}\big)\big]\leqslant 0.$$

Therefore by the Jensen inequality we get

$$\begin{aligned}\mathbf{E}\max_{i=1,\dots,n}\log^{2q}[p(X_{(i)})] &\leqslant \frac{1}{\varepsilon^{2q}}\mathbf{E}\log^{2q}\left[\sum_{i=1}^{n}[p^{\varepsilon}(X_i)+p^{-\varepsilon}(X_i)]\right]\\ &\leqslant \frac{1}{\varepsilon^{2q}}\log^{2q}\left[\mathbf{E}\sum_{i=1}^{n}[p^{\varepsilon}(X_i)+p^{-\varepsilon}(X_i)]+\mathrm{e}^{2q-1}\right]\\ &= \frac{1}{\varepsilon^{2q}}\log^{2q}\left[n\mathbf{E}p^{\varepsilon}(X_1)+n\mathbf{E}p^{-\varepsilon}(X_1)+\mathrm{e}^{2q-1}\right],\end{aligned}$$

thus proving (22).

The proof of (23) is quite similar and therefore it is omitted. Finally, notice that (24) is a particular case of (23) when X_i are i.i.d. uniformly distributed on [0.1]. □

Proof of Lemma 1

To simplify technical details, we focus on the case $m = 1$. Decompose ϵ_n as follows

$$\epsilon_n = \epsilon_{1n} + \epsilon_{2n},$$

where

$$\begin{aligned}\epsilon_{1n} &= \frac{1}{\sqrt{n}}\sum_{i<n:X_{(i)}\in\mathbb{Q}^n_{r,R}}\log\frac{F(X_{(i+1)}-F(X_{(i)})}{p(X_{(i)})(X_{(i+1)}-X_{(i)})},\\ \epsilon_{2n} &= \frac{1}{\sqrt{n}}\sum_{i<n:X_{(i)}\notin\mathbb{Q}^n_{r,R}}\log\frac{F(X_{(i+1)}-F(X_{(i)})}{p(X_{(i)})(X_{(i+1)}-X_{(i)})}.\end{aligned}$$

Let us first bound from above ϵ_{2n}. We obviously have

$$\begin{aligned}|\epsilon_{2n}| &\leqslant \frac{1}{\sqrt{n}}\sum_{i<n}\left|\log\frac{F(X_{(i+1)}-F(X_{(i)})}{p(X_{(i)})(X_{(i+1)}-X_{(i)})}\right|\mathbf{1}\{X_{(i)}\notin\mathbb{Q}^n_{r,R}\}\\ &\leqslant \max_{i<n}\left|\log\frac{F(X_{(i+1)}-F(X_{(i)})}{p(X_{(i)})(X_{(i+1)}-X_{(i)})}\right|\frac{1}{\sqrt{n}}\sum_{i\leqslant n}\mathbf{1}\{X_{(i)}\notin\mathbb{Q}^n_{r,R}\}.\end{aligned}$$

Therefore by the Cauchy–Schwarz inequality and by Lemma 4 we obtain

$$\begin{aligned}\mathbf{E}^{1/2}\epsilon_{2n}^2 &\leqslant C\log(n)\ \mathbf{E}^{1/4}\left[\frac{1}{\sqrt{n}}\sum_{i\leqslant n}\mathbf{1}\{X_i\notin\mathbb{Q}_{r,R}^n\}\right]^4\\ &= C\log(n)\ \mathbf{E}^{1/4}\Bigg[\sqrt{n}\mathbf{P}\{X_1\notin\mathbb{Q}_{r,R}^n\}\\ &+\frac{1}{\sqrt{n}}\sum_{i\leqslant n}\big(\mathbf{1}\{X_i\notin\mathbb{Q}_{r,R}^n\}-\mathbf{P}\{X_i\notin\mathbb{Q}_{r,R}^n\}\big)\Bigg]^4\\ &\leqslant C\log(n)\sqrt{n}\mathbf{P}\{X_1\notin\mathbb{Q}_{r,R}^n\}+C\log(n)\mathbf{P}^{1/2}\{X_1\notin\mathbb{Q}_{r,R}^n\}\\ &\leqslant C\log(n)\sqrt{n}\mathbf{P}\{X_1\notin\mathbb{Q}_{r,R}^n\}+\frac{C\log(n)}{\sqrt{n}}.\end{aligned}\tag{25}$$

In the last line we used

$$\begin{aligned}\mathbf{P}^{1/2}\{X_1\notin\mathbb{Q}_{r,R}^n\} &= \frac{1}{n^{1/4}}n^{1/4}\mathbf{P}^{1/2}\{X_1\notin\mathbb{Q}_{r,R}^n\}\\ &\leqslant\frac{\sqrt{n}}{2}\mathbf{P}\{X_1\notin\mathbb{Q}_{r,R}^n\}+\frac{1}{2\sqrt{n}}\end{aligned}$$

since $ab\leqslant a^2/2+b^2/2$.

Our final step is to find an upper bound for $\mathbf{E}\epsilon_{1n}^2$. Consider the following set

$$\mathcal{A}_r^n=\left\{\mathbf{X}^n:\ \max_i[F(X_{(i+1)})-F(X_{(i)})]\leqslant\frac{r(n)}{n}\right\}$$

and notice that with a very high probability $\mathbf{X}^n$ belongs to $\mathcal{A}_r^n$. To see this one may combine Pyke's theorem together with Lemma 7. Indeed, recalling that $r(n)\geqslant 5\log(n)$, we arrive at

$$\begin{aligned}&\mathbf{P}\left\{\mathbf{X}^n\notin\mathcal{A}_r^n\right\}=\mathbf{P}\left\{\max_i[U_{(i+1)}-U_{(i)}]\geqslant\frac{r(n)}{n}\right\}\\ =&\mathbf{P}\left\{\max_i e_i\geqslant\frac{r(n)}{n}\sum_{k=1}^{n+1}e_k\right\}\leqslant Cn\exp[-r(n)]\leqslant\frac{C}{n^4}.\end{aligned}\tag{26}$$

We begin with a rough upper bound for ϵ_{1n}. By Lemma 4, we obtain

$$\mathbf{E}^{1/4}\epsilon_{1n}^4\leqslant C\sqrt{n}\log n.$$

Hence using (26) and the Cauchy–Schwarz inequality, we arrive at

$$\mathbf{E}^{1/2}\epsilon_{1n}^2\mathbf{1}\Big\{\mathbf{X}^n\notin\mathcal{A}_r^n\Big\}\leqslant\mathbf{E}^{1/4}\epsilon_{1n}^4\mathbf{P}^{1/4}\Big\{\mathbf{X}^n\notin\mathcal{A}_r^n\Big\}\leqslant\frac{C\log(n)}{\sqrt{n}}. \tag{27}$$

In order to control $\mathbf{E}^{1/2}|\epsilon_{1n}|^2\mathbf{1}\{\mathbf{X}^n\in\mathcal{A}_r^n\}$, we apply Lemma 3. We have

$$|\epsilon_{1n}|\mathbf{1}\Big\{\mathbf{X}^n\in\mathcal{A}_r^n\Big\}\leqslant\mathbf{1}\Big\{\mathbf{X}^n\in\mathcal{A}_r^n\Big\}\frac{C}{\sqrt{n}}\Big|\sum_{i<n:X_{(i)}\in\mathbb{Q}_{r,R}^n}\log\frac{p(X_{(i+1)})}{p(X_{(i)})}\Big|$$

$$+\mathbf{1}\Big\{\mathbf{X}\in\mathcal{A}_r^n\Big\}\frac{C\log(n)}{n^{3/2}}\sum_{i<n:X_{(i)}\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(X_{(i)})}{p^2(X_{(i)})}[F(X_{(i+1)})-F(X_{(i)})]. \tag{28}$$

In view of the definition of $D_r^n(x)$ and Lemma 2 it is clear that

$$\mathbf{1}\Big\{\mathbf{X}^n\in\mathcal{A}_r^n\Big\}\sum_{i:X_{(i)}\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(X_{(i)})}{p^2(X_{(i)})}[F(X_{(i+1)})-F(X_{(i)})]$$

$$\leqslant C\int\limits_{x\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(x)}{p(x)}\,dx. \tag{29}$$

Next, since $\mathbb{Q}_{r,R}^n$ may have only a finite number of connected components, say N, we get

$$\Big|\sum_{i:X_{(i)}\in\mathbb{Q}_{r,R}^n}\log\frac{p(X_{(i+1)})}{p(X_{(i)})}\Big|\leqslant\sum_{k=1}^N\Big|\log\frac{p(X_{i_2(k)})}{p(X_{i_1(k)})}\Big|$$

$$\leqslant\sum_{k=1}^N\Big[|\log p(X_{i_2(k)})|+|\log p(X_{i_1(k)})|\Big]\leqslant 2N\max_i|\log p(X_i)|,$$

where $i_1(k)$ and $i_2(k)$ are some indices from $\{1,\dots,n\}$. Therefore the above display and Lemma 4 obviously yield

$$\mathbf{E}\Bigg[\sum_{i:X_{(i)}\in\mathbb{Q}_{r,R}^n}\log\frac{p(X_{(i+1)})}{p(X_{(i)})}\Bigg]^2\leqslant C\log^2(n).$$

Finally, combining this inequality with (27)–(29) we arrive at

$$\mathbf{E}\epsilon_{1n}^2\leqslant\frac{C\log(n)}{\sqrt{n}}+\frac{C\log(n)}{n^{3/2}}\int\limits_{x\in\mathbb{Q}_{r,R}^n}\frac{D_r^n(x)}{p(x)}\,dx,$$

thus finishing the proof (see also (25)). □

3. Controlling the correlation term

In the previous section, it was shown (see Lemma 1) that under some conditions the m-spacing entropy estimator admits the following decomposition

$$\sqrt{n}\left[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)\right] = -E^n + S^n + \epsilon_n, \qquad (30)$$

where $\lim_{n\to\infty} \mathbf{E}\,\epsilon_n^2 = 0$ and

$$E^n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[\log p(X_i) - \mathbf{E}\log p(X_1)\big],$$

$$S^n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n-m} \Big\{\log[F(X_{(i+m)}) - F(X_{(i)})] - \mathbf{E}\log[F(X_{(i+m)}) - F(X_{(i)})]\Big\}.$$

The main goal in this section is to show that E^n and S^n are weakly correlated.

Theorem 3. *Assume that the conditions of Lemma* 1 *and* (12) *hold true. Then*

$$\limsup_{n\to\infty} \big|\mathbf{E}E^n S^n\big| = 0. \qquad (31)$$

To simplify technical detail we assume that $m = 1$ and that $\mathbb{Q}_{r,R}^n$ has the only one connected component. Denote

$$P^{(0)} = \mathbf{P}\Big\{X_1 \in \mathbb{Q}_{r,R}^n\Big\}, \quad P^{(1)} = 1 - P^{(0)}.$$

Notice that the sample $\mathbf{X}^n$ can be generated as follows. Let $\chi_i \in \{0,1\}$ be i.i.d. such that

$$\mathbf{P}\{\chi_i = k\} = P^{(k)} \quad k = 0,1.$$

Denote

$$\tau_k = \sum_{i=1}^{n} \mathbf{1}\{\chi_i = k\} \quad k = 0,1.$$

Then $\mathbf{X}^n$ can be represented as $(X_1^{(0)}, \ldots, X_{\tau_0}^{(0)}, X_1^{(1)}, \ldots, X_{\tau_1}^{(1)})$, where $X_k^{(0)}$ and $X_k^{(1)}$ are i.i.d. random variables with the densities

$$p^{(0)}(x) = \frac{p(x)\mathbf{1}\{x \in \mathbb{Q}_{r,R}^n\}}{P^{(0)}} \quad \text{and} \quad p^{(1)}(x) = \frac{p(x)\mathbf{1}\{x \notin \mathbb{Q}_{r,R}^n\}}{P^{(1)}},$$

respectively. Denote also for brevity

$$F^{(k)}(x) = \int_{-\infty}^{x} p^{(k)}(u)\,du, \quad k = 0, 1.$$

With the samples $(X_1^{(0)}, \dots, X_{\tau_0}^{(0)})$ and $(X_1^{(1)}, \dots, X_{\tau_1}^{(1)})$ we associate ($k = 0, 1$) the following statistics

$$\begin{aligned} E_k^n &= \frac{1}{\sqrt{n}} \sum_{i=1}^{\tau_k} \big[\log p^{(k)}(X_i^{(k)}) - \mathbf{E} \log p^{(k)}(X_1^{(k)})\big], \\ S_k^n &= \frac{1}{\sqrt{n}} \sum_{i=1}^{\tau_k} \Big\{ \log[F^{(k)}(X_{(i+m)}^{(k)}) - F^{(k)}(X_{(i)}^{(k)})] \\ &\qquad - \mathbf{E} \log[F^{(k)}(X_{(i+m)}^{(k)}) - F^{(k)}(X_{(i)}^{(k)})] \Big\}. \end{aligned} \tag{32}$$

The proof of Lemma 3 is essentially based the following fact.

Lemma 5. *Under the conditions of Lemma* 1 *and* (12),

$$\limsup_{n\to\infty} \mathbf{E} E_0^n S_0^n = 0.$$

Proof. To simplify notations, we denote

$$X_{(k)}^{(0)} = X_{(k)}, \quad p^{(0)}(x) = p(x), \quad F^{(0)}(x) = F(x).$$

It clear that

$$E_0^n = \frac{1}{\sqrt{n}} \sum_{i=1}^{\tau_0} \big[\log p(X_{(i)}) - \mathbf{E} \log p(X_1)\big],$$

and to evaluate the correlation of E_0^n and S_0^n we can use Pyke's theorem (see Theorem 1) which permits to represent these random variables in terms of $e_1, \dots, e_{\tau_0+1}$.

$$X_{(i)} = F^{-1}(U_{(i)}) = F^{-1}\Big(\sum_{k=1}^{i} e_k \Big/ \sum_{k=1}^{\tau_0+1} e_k \Big).$$

Let us first look at S_0^n. We have

$$\log[F(X_{(i+1)}) - F(X_{(i)})] = \log(e_{i+1}) - \log\left[1 + \frac{1}{\tau_0+1}\sum_{k=1}^{\tau_0+1}(e_k - 1)\right] + \log(\tau_0+1).$$

Next expanding $\log(1+\cdot)$ (see for details Lemma 10), we can write

$$\log\left[1 + \frac{1}{\tau_0+1}\sum_{k=1}^{\tau_0+1}(e_k-1)\right] = \frac{1}{\tau_0+1}\sum_{k=1}^{\tau_0+1}(e_k-1) + O\left(\frac{1}{\tau_0}\right),$$

and, therefore,

$$S_0^n = \frac{1}{\sqrt{n}}\sum_{i=1}^{\tau_0-1}[\log(e_{i+1}) - \mathbf{E}\log(e_1)] + \frac{\tau_0-1}{(\tau_0+1)\sqrt{n}}\sum_{k=1}^{\tau_0+1}[e_k - 1] + O\left(\frac{1}{\sqrt{\tau_0}}\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=2}^{\tau_0}[\log(e_i) + e_i - \mathbf{E}\log(e_1) - 1] + O\left(\frac{1}{\sqrt{\tau_0}}\right) = \widetilde{S}_0^n + O\left(\frac{1}{\sqrt{\tau_0}}\right), \quad (33)$$

with

$$\widetilde{S}_0^n = \frac{1}{\sqrt{n}}\sum_{i=1}^{\tau_0}[\log(e_i) + e_i - \mathbf{E}\log(e_1) - 1].$$

In view of (33) and the Cauchy–Schwarz inequality, it is clear that

$$\mathbf{E}E_0^n S_0^n = \mathbf{E}E_0^n\widetilde{S}_0^n + O\left(\mathbf{E}\frac{1}{\sqrt{\tau_0}}\right). \quad (34)$$

The representation of E_0^n in terms of $e_1, \ldots, e_{\tau_0+1}$ is quite similar. Denote for brevity

$$G(u) = \log p[F^{-1}(u)].$$

Notice that instead of E_0^n we can deal with

$$\widetilde{E}_0^n \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^{\tau_0}\log p(X_{(i)}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\tau_0}G[U_{(i)}]$$

because

$$\mathbf{E}E_0^n\widetilde{S}_0^n = \mathbf{E}\widetilde{E}_0^n\widetilde{S}_0^n. \quad (35)$$

So, in view of (34), it remains to compute $\mathbf{E}\widetilde{E}_0^n\widetilde{S}_0^n$. We do that with the help of Lemma 8. Denote

$$U_{(i)}^{-k}=\Big[\sum_{s=1}^{i}e_i-\mathbf{1}\{k\leqslant i\}e_k\Big]\Big/\Big[\sum_{s=1}^{n}e_i-e_k\Big]$$

and notice that e_k and $U_{(i)}^{-k}$ are independent. Therefore we have

$$\begin{aligned}\mathbf{E}\widetilde{E}_0^n\widetilde{S}_0^n&=\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G[U_{(i)}][\log(e_k)+e_k-\mathbf{E}\log(e_1)-1]\\&=\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G[U_{(i)}^{-k}][\log(e_k)+e_k-\mathbf{E}\log(e_1)-1]\\&+\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}\Big\{G[U_{(i)}]-G[U_{(i)}^{-k}]\Big\}[\log(e_k)+e_k-\mathbf{E}\log(e_1)-1]\\&=\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}\Big\{G[U_{(i)}]-G[U_{(i)}^{-k}]\Big\}[\log(e_k)+e_k-\mathbf{E}\log(e_1)-1].\end{aligned}\tag{36}$$

To control the right-hand side at the above display, we use the following Tailor expansion

$$G[U_{(i)}]-G[U_{(i)}^{-k}]=G'\big(U_{(i)}^{-k}\big)\big[U_{(i)}-U_{(i)}^{-k}\big]+\frac{1}{2}G''(\xi_i)\big[U_{(i)}-U_{(i)}^{-k}\big]^2,\tag{37}$$

where

$$\begin{aligned}G'(u)&=\frac{p'(F^{-1}(u))}{p^2(F^{-1}(u))},\\G''(u)&=\frac{1}{p^2(F^{-1}(u))}\left[\frac{p''(F^{-1}(u))}{p(F^{-1}(u))}-2\left(\frac{p'(F^{-1}(u))}{p(F^{-1}(u))}\right)^2\right]\end{aligned}$$

and $\xi_i\in[\min(U_{(i)},U_{(i)}^{-k}),\max(U_{(i)},U_{(i)}^{-k})]$. Substituting (37) in (36), we obtain the following formula

$$\mathbf{E}\widetilde{E}_0^n\widetilde{S}_0^n=\mathcal{R}_1^n+\mathcal{R}_2^n,\tag{38}$$

with

$$\mathcal{R}_1^n=\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G'\big(U_{(i)}^{-k}\big)\big[U_{(i)}-U_{(i)}^{-k}\big][\log(e_k)+e_k-\mathbf{E}\log(e_1)-1],$$

$$\mathcal{R}_2^n = \frac{1}{2n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G''(\xi_i)\big[U_{(i)} - U_{(i)}^{-k}\big]^2[\log(e_k) + e_k - \mathbf{E}\log(e_1) - 1],$$

and our goal is to show that $\mathcal{R}_1^n$ and $\mathcal{R}_2^n$ are small.

We begin with $\mathcal{R}_2^n$. By Lemma 8,

$$\begin{aligned} U_{(i)} - U_{(i)}^{-k} &= U_{(i)}^{-k} e_k \Big/ \sum_{s\neq k} e_s - \mathbf{1}\{k \leqslant i\} e_k \Big/ \sum_{s\neq k} e_s + O\left(\frac{1}{\tau_0^2}\right), \\ U_{(i)} - U_{(i)}^{-k} &= U_{(i)} e_k \Big/ \sum_{s} e_s - \mathbf{1}\{k \leqslant i\} e_k \Big/ \sum_{s} e_s + O\left(\frac{1}{\tau_0^2}\right). \end{aligned} \tag{39}$$

This obviously yields that $|U_{(i)} - U_{(i)}^{-k}| \leqslant O(1/\tau_0)$, and with the Cauchy–Schwarz inequality we get

$$|\mathcal{R}_2^n| \leqslant \frac{C}{n}\mathbf{E}\frac{1}{\tau_0^2}\sum_{i,k=1}^{\tau_0} \mathbf{E}^{1/2}[G''(\xi_i)]^2 \leqslant \frac{C}{R(n)}. \tag{40}$$

In the above inequality we used (see (9)) that $|G''(\xi_i)| \leqslant n/R(n)$.

Let us look at $\mathcal{R}_1^n$. Denoting for brevity $\rho = \mathbf{E}\big[\log(e_k) + e_k - \mathbf{E}\log(e_1) - 1\big]e_k$, we obtain with the help of (39)

$$\begin{aligned} \mathcal{R}_1^n &= \frac{\rho}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}^{-k}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k \leqslant i\}\big] \\ &\quad + O\left(\frac{1}{n}\mathbf{E}\frac{1}{\tau_0^2}\sum_{i,k=1}^{\tau_0} \mathbf{E}^{1/2}G'^2\big(U_{(i)}^{-k}\big)\right) \\ &= \frac{\rho}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}^{-k}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k \leqslant i\}\big] + O\left(\frac{1}{R(n)}\right). \end{aligned} \tag{41}$$

Consider

$$\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G'\big(U_{(i)}\big)\bigg[\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\big[U_{(i)}-\mathbf{1}\{k\leqslant i\}\big]$$
$$=\frac{1}{n}\mathbf{E}\sum_{i=1}^{\tau_0}\mathbf{E}G'\big(U_{(i)}\big)\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]$$
$$=\frac{1}{n}\mathbf{E}\sum_{i=1}^{\tau_0}G'\bigg(\frac{i}{\tau_0}\bigg)\mathbf{E}\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]$$
$$+\frac{1}{n}\mathbf{E}\sum_{i=1}^{\tau_0}\mathbf{E}G''\big(\xi_i\big)\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]^2. \tag{42}$$

Since $|G''\big(\xi_i\big)|\leqslant n/R(n)$, the last term can be controlled very easily

$$\bigg|\frac{1}{n}\mathbf{E}\sum_{i=1}^{\tau_0}\mathbf{E}G''\big(\xi_i\big)\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]^2\bigg|$$
$$\leqslant\frac{1}{R(n)}\mathbf{E}\sum_{i=1}^{\tau_0}\mathbf{E}^{1/2}\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-2}\mathbf{E}^{1/2}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]^4\leqslant\frac{C}{R(n)}. \tag{43}$$

The same upper bound holds true for the first therm at the right-hand side of (42). By Lemma 11, we have

$$\frac{1}{n}\mathbf{E}\sum_{i=1}^{\tau_0}\bigg|G'\bigg(\frac{i}{\tau_0}\bigg)\bigg|\times\bigg|\mathbf{E}\bigg[\frac{1}{\tau_0}\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\bigg[U_{(i)}-\frac{i}{\tau_0}\bigg]\bigg|$$
$$\leqslant\frac{C}{n}\mathbf{E}\frac{1}{\tau_0}\sum_{i=1}^{\tau_0}\bigg|G'\bigg(\frac{i}{\tau_0}\bigg)\bigg|\leqslant\frac{C}{R(n)}.$$

Therefore, combining the above inequalities, we obtain

$$\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G'\big(U_{(i)}\big)\bigg[\sum_{s=1}^{\tau_0+1}e_s\bigg]^{-1}\big[U_{(i)}-\mathbf{1}\{k\leqslant i\}\big]=O\bigg(\frac{1}{R(n)}\bigg). \tag{44}$$

Recall that we need to control a little bit different term, namely

$$\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0}\mathbf{E}G'\big(U_{(i)}^{-k}\big)\bigg[\sum_{s\neq k}e_s\bigg]^{-1}\big[U_{(i)}^{-k}-\mathbf{1}\{k\leqslant i\}\big].$$

However, noticing that

$$\left|U_{(i)}^{-k} - U_{(k)}\right| \leqslant O\left(\frac{1}{\tau_0}\right) \quad \text{and} \quad \left|\left[\sum_{s\neq k} e_s\right]^{-1} - \left[\sum_{s=1}^{\tau_0+1} e_s\right]^{-1}\right| \leqslant O\left(\frac{1}{\tau_0^2}\right),$$

we can do that very easily. Indeed, applying several times the Taylor formula, we obtain

$$\begin{aligned}
&\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}^{-k}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] \\
=&\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] \\
&+ O(1)\frac{1}{n}\mathbf{E}\frac{1}{\tau_0}\sum_{i,k=1}^{\tau_0} \mathbf{E}G''\big(\xi_i\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] \\
=&\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] + O\left(\frac{1}{R(n)}\right) \\
=&\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}\big)\left[\sum_{s=1}^{\tau_0+1} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] + O\left(\frac{1}{R(n)}\right) \\
=&\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}\big)\left[\sum_{s=1}^{\tau_0+1} e_s\right]^{-1}\big[U_{(i)} - \mathbf{1}\{k\leqslant i\}\big] + O\left(\frac{1}{R(n)}\right).
\end{aligned}$$

Therefore, by (44),

$$\frac{1}{n}\mathbf{E}\sum_{i,k=1}^{\tau_0} \mathbf{E}G'\big(U_{(i)}^{-k}\big)\left[\sum_{s\neq k} e_s\right]^{-1}\big[U_{(i)}^{-k} - \mathbf{1}\{k\leqslant i\}\big] \leqslant O\left(\frac{1}{R(n)}\right).$$

This inequality, together with (34)–(36), (40), and (41), finish the proof. □

Lemma 6. *Under the conditions of Lemma* 1 *and* (12)

$$\limsup_{n\to\infty} \mathbf{E}\big[E_1^n\big]^2 = 0, \quad \limsup_{n\to\infty} \mathbf{E}\big[S_1^n\big]^2 = 0$$

(see (32) for definitions of E_1^n and S_1^n).

Proof. It remains to notice that in view of condition (12)

$$\mathbf{E}\tau_1 = nP^{(1)} = n \int\limits_{x \notin \mathbb{Q}^n_{r,R}} p(x)\,dx \leqslant C\sqrt{n}.$$

Therefore, by the independence of $X_k^{(1)}$, we get

$$\mathbf{E}\big[E_1^n\big]^2 = \frac{1}{n}\mathbf{E}\bigg[\sum_{i=1}^{\tau_1}\big[\log p^{(1)}(X_i^{(1)}) - \mathbf{E}\log p^{(1)}(X_1^{(1)})\big]\bigg]^2$$
$$\leqslant \frac{\mathbf{E}\tau_1}{n} \int\limits_{x \notin \mathbb{Q}^n_{r,R}} \log^2(p(x))p(x)\,dx \leqslant CP^{(1)} \leqslant \frac{C}{\sqrt{n}}.$$

To compute $\mathbf{E}\big[S_1^n\big]^2$, we use Pyke's theorem. So, we have

$$\mathbf{E}\big[S_1^n\big]^2 = \frac{1}{n}\mathbf{E}\bigg\{\sum_{i=1}^{\tau_1-m}\log\bigg[\frac{1}{m}\sum_{k=i}^{i+m-1}e_k\bigg] - \mathbf{E}\log\bigg[\frac{1}{m}\sum_{k=i}^{i+m-1}e_k\bigg]$$
$$-\big(\tau_1-m\big)\log\bigg[\frac{1}{\tau_1}\sum_{k=1}^{\tau_1+1}e_k\bigg] + \mathbf{E}\big(\tau_1-m\big)\log\bigg[\frac{1}{\tau_1}\sum_{k=1}^{\tau_1+1}e_k\bigg]\bigg\}^2 \leqslant \frac{C\mathbf{E}\tau_1}{n} \leqslant \frac{C}{\sqrt{n}}.$$

□

Proof of Theorem 3. It is clear that

$$E^n = E_0^n + E_1^n.$$

On the other hand, the decomposition of S^n is not so banal because of the chaining. It means that ordering $(X_1^{(0)}, \dots, X_{\tau_0}^{(0)}, X_1^{(1)}, \dots, X_{\tau_1}^{(1)})$, we get a vector with the following structure

$$(X_{(1)}^{(1)}, \dots, X_{(t)}^{(1)}, X_{(1)}^{(0)}, \dots, X_{(\tau_0)}^{(0)}, X_{(t+1)}^{(1)}, \dots, X_{(\tau_1)}^{(1)})$$

with some integer $t \in [0, \tau_1]$. This results in the following decomposition

$$S^n = S_0^1 + S_1^n$$
$$+ \frac{1}{\sqrt{n}}\big\{\log[F(X_{(1)}^{(0)}) - F(X_{(t)}^{(1)})] - \mathbf{E}\log[F(X_{(1)}^{(0)}) - F(X_{(t)}^{(1)})]\big\}$$
$$+ \frac{1}{\sqrt{n}}\big\{\log[F(X_{(t+1)}^{(1)}) - F(X_{(\tau_0)}^{(0)})] - \mathbf{E}\log[F(X_{(t+1)}^{(1)}) - F(X_{(\tau_0)}^{(0)})]\big\}$$
$$- \frac{1}{\sqrt{n}}\big\{\log[F(X_{(t+1)}^{(1)}) - F(X_{(t)}^{(1)})] - \mathbf{E}\log[F(X_{(t+1)}^{(1)}) - F(X_{(t)}^{(1)})]\big\}.$$

By Pyke's theorem and Lemma 7, three remainder terms at the right-hand side of the above display are bounded by $C\log(n)/\sqrt{n}$. This remark and Lemmas 5 and 6 finish the proof. □

The following theorem summarizes the principal facts of this paper.

Theorem 4. *Let $r(n) = 5\log(n)$ and $R(n) \geqslant 7r(n)$. Assume that*

- *the number of connected components of $\mathbb{Q}^n_{r,R}$ is bounded uniformly in n;*
- *for some $\varepsilon > 0$*

$$\mathbf{E}|X_1 - X_2|^{\pm\varepsilon} < \infty, \quad \mathbf{E}p^{\pm\epsilon}(X_1) < \infty,$$

-

$$\limsup_{n\to\infty} \log(n)\left[\sqrt{n} \int\limits_{x\notin\mathbb{Q}^n_{r,R}} p(x)\,dx + \frac{1}{n^{3/2}} \int\limits_{x\in\mathbb{Q}^n_{r,R}} \frac{D^n_r(x)}{p(x)}\,dx\right] = 0. \tag{46}$$

Then

$$\lim_{n\to\infty} \sqrt{n}[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2 + \sigma^2(p)),$$

where

$$\Sigma^2 = (2m^2 - 2m + 1)\Psi'(m) - 2m + 1,$$

$$\sigma^2(p) = \int\limits_{-\infty}^{\infty} \log^2(p(x))p(x)\,dx - H^2(p),$$

and

$$\lim_{n\to\infty} n\mathbf{E}[\widehat{H}_m(\mathbf{X}^n) - H(p) - \Psi(m) + \log(m)]^2 = \sigma^2(p) + \Sigma^2.$$

4. Examples

In this section, we show that the conditions of Theorem 4 can be checked for standard densities. Our attention is focused on two distinct densities such as Gaussian and Cauchy but our arguments can be easily extended to very general density families. Notice however that the main difficulties are related to checking the condition (46).

1. The Gaussian density. We start with controlling $\mathbb{Q}^n_{r,R}$. Since

$$D_n(x) \geqslant \frac{|p'^2(x)|}{p^2(x)} + \frac{|p''(x)|}{p(x)} \geqslant x^2 + |x^2 - 1| \geqslant 1,$$

we obviously have

$$\mathbb{Q}^n_{r,R} = \left\{x : \ \frac{\sqrt{D^n_r(x)}}{p(x)} \leqslant \frac{n}{R(n)}\right\} \subseteq \mathbb{Q}^n_+ \stackrel{\text{def}}{=} \left\{x : \ |x| \leqslant \sqrt{2\log\frac{n}{R(n)\sqrt{2\pi}}}\right\}. \quad (47)$$

On the other hand, it is easy to check that

$$\frac{|p'^2(x)|}{p^2(x)} + \frac{|p''(x)|}{p(x)} \leqslant 2x^2 + 1,$$

and therefore for all sufficiently large n

$$\begin{aligned} D_n(x) &\leqslant \sup_{x\geqslant 0:\, x\in\mathbb{Q}^n_+} \left\{\left\{F^{-1}\left[F(x) + \frac{r(n)}{n}\right]\right\}^2 + 1\right\} \\ &\leqslant \left\{\left\{F^{-1}\left[1 - \frac{R(n)}{n\sqrt{2\log(n)}} + \frac{r(n)}{n}\right]\right\}^2 + 1\right\} \\ &\leqslant -2\log\left[\frac{R(n)}{n\sqrt{2\log(n)}} - \frac{r(n)}{n}\right]_+ + 1, \quad x \in \mathbb{Q}^n_{r,R}. \end{aligned} \quad (48)$$

In the above display we used two well-known inequalities:

- $$F(x) \leqslant 1 - \frac{1}{\sqrt{2\pi}x}\exp(-x^2/2)$$

- $F(x) \geqslant 1 - \exp(-x^2/2)$, or equivalently $F^{-1}(x) \leqslant \sqrt{-2\log(1-x)}$.

Thus if $R(n) \geqslant 2\sqrt{2\log(n)}r(n)$, then with (48) we arrive at

$$\mathbb{Q}^n_{r,R} = \left\{x : \ \frac{\sqrt{D^n_r(x)}}{p(x)} \leqslant \frac{n}{R(n)}\right\} \supseteq \mathbb{Q}^n_- \stackrel{\text{def}}{=} \left\{x : \ |x| \leqslant \sqrt{2\log\frac{n}{R(n)\sqrt{8\pi\log(2n)}}}\right\}. \quad (49)$$

Finally, integrating by parts, we obtain by (49)

$$\int\limits_{x\notin\mathbb{Q}^n_{r,R}} p(x)\,dx \leqslant \frac{CR(n)\sqrt{\log(n)}}{n}$$

and, by (47),

$$\int\limits_{x\in\mathbb{Q}^n_{r,R}} \frac{\sqrt{D^n_r(x)}}{p(x)}\,dx \leqslant \frac{Cn}{R(n)}.$$

Thus, in order to check condition (46), we choose $R(n) \geqslant 7\log(n)$ and obtain the following equivalent form of this condition

$$\lim_{n\to\infty} \log(n)\left[\frac{R(n)\sqrt{\log(n)}}{\sqrt{n}} + \frac{1}{R(n)\sqrt{n}}\right] = 0.$$

2. Cauchy density. In this case, it is easy to see that

$$\frac{|p'^2(x)|}{p^2(x)} + \frac{|p''(x)|}{p(x)} \leqslant \frac{C}{1+x^2}.$$

Therefore since

$$\mathbb{B}^\tau_r(x) \subseteq \left\{y : y \geqslant x - \frac{r(n)}{np(x)}\right\},$$

we get that if for some $\alpha \in (0,1)$

$$\frac{r(n)}{np(x)} \leqslant \alpha|x|,$$

then

$$D_n(x) \leqslant \frac{C}{x^2}, \quad |x| \geqslant 1.$$

These facts immediately imply that for some $C_1 < C_2$

$$\left\{x : |x| \leqslant \frac{C_1 n}{R(n)}\right\} \subseteq \mathbb{Q}^n_{r,R} \subseteq \left\{x : |x| \leqslant \frac{C_2 n}{R(n)}\right\}.$$

Thus we easily obtain

$$\int\limits_{x\in\mathbb{Q}^n_{r,R}} \frac{\sqrt{D^n_r(x)}}{p(x)}\,dx \leqslant \left[\frac{Cn}{R(n)}\right]^2$$

and

$$\int\limits_{x\notin\mathbb{Q}^n_{r,R}} p(x)\,dx \leqslant \frac{CR(n)}{n}.$$

Therefore the principal condition (46) is fulfilled if

$$\lim_{n\to\infty} \log(n)\left[\frac{R(n)}{\sqrt{n}} + \frac{\sqrt{n}}{R^2(n)}\right] = 0.$$

It is easy to see than $R(n) = n^{1/3}$ provides the optimal choice guaranteing this property.

5. Appendix

In this section, we collect some simple technical facts. They are well-known, and we provide them only for reader convenience. Let e_i be i.i.d. standard exponentially distributed random variables.

Lemma 7. *Uniformly in* $x \in [0, \sqrt{n}]$

$$\mathbf{P}\left\{\max_{k=1,n} e_k \geqslant \frac{\log(n)+x}{n}\sum_{k=1}^{n} e_k\right\} \leqslant C\exp(-x). \tag{50}$$

Proof. By the Markov inequality and the Taylor formula, we obtain

$$\begin{aligned}
&\mathbf{P}\left\{\max_{k=1,n} e_k \geqslant \frac{\log(n)+x}{n}\sum_{k=1}^{n} e_k\right\} \leqslant \sum_{i=1}^{n}\mathbf{P}\left\{e_i \geqslant \frac{\log(n)+x}{n}\sum_{k=1}^{n} e_k\right\}\\
&\leqslant n\mathbf{P}\left\{e_1\left(1-\frac{\log(n)+x}{n}\right) \geqslant \frac{\log(n)+x}{n}\sum_{k=2}^{n} e_k\right\}\\
&\quad = n\mathbf{E}\exp\left(-\frac{\log(n)+x}{n-\log(n)-x}\sum_{k=2}^{n} e_k\right)\\
&= n\exp\left[-(n-1)\log\left(1-\frac{\log(n)+x}{n-\log(n)-x}\right)\right]\\
&= n\exp\left[-(\log(n)+x)\big(1+O(n^{-1/2})\big)+O(1)\right] \leqslant \exp[-x+O(1)].
\end{aligned}$$

□

Lemma 8. *For some $C < \infty$*

$$\mathbf{E}^{1/2}\Big[U_{(i)}^{-k} - U_{(i)} - U_{(i)}^{-k}e_k \Big/ \sum_{s\neq k} e_s + \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s\Big]^2 \leqslant \frac{C}{n^2},$$

where

$$U_{(i)} = \sum_{s=1}^{i} e_i \Big/ \sum_{s=1}^{n} e_i \quad \textit{and} \quad U_{(i)}^{-k} = \Big[\sum_{s=1}^{i} e_i - \mathbf{1}\{k \leqslant i\}e_k\Big] \Big/ \Big[\sum_{s=1}^{n} e_i - e_k\Big].$$

Proof. Define the following subset in $\mathbb{R}^n$

$$\mathcal{A}^n = \Big\{x_i \geqslant 0 : \sum_{i=1}^{n+1} x_i \geqslant \frac{n}{2}\Big\}.$$

If $\mathbf{e}^n \in \mathcal{A}^n$, it follows immediately from the Taylor formula that

$$U_{(i)}^{-k} = \sum_{s=1}^{i} e_s \Big/ \sum_{s\neq k} e_s - \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s$$

$$= U_{(i)} + U_{(i)}e_k \Big/ \sum_{s=1}^{n+1} e_s - \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s + O\Big(\frac{e_k^2 + e_k}{n^2}\Big) \tag{51}$$

$$= U_{(i)} + U_{(i)}^{-k}e_k \Big/ \sum_{s\neq k} e_s - \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s + O\Big(\frac{e_k^2 + e_k}{n^2}\Big).$$

Next note that using the Chernov inequality with $\lambda = 1$, we get

$$\mathbf{P}\{\mathbf{e}^n \notin \mathcal{A}^n\} = \mathbf{P}\Big\{\sum_{i=1}^{n+1} e_i \leqslant \frac{n}{2}\Big\} = \mathbf{P}\Big\{\sum_{i=1}^{n}(1 - e_i) \geqslant \frac{n}{2}\Big\} \tag{52}$$

$$\leqslant \exp\{-n \max_{\lambda}[\log(1+\lambda) - \lambda/2]\} = \exp\{-n[\log(2) - 1/2]\} \leqslant \exp[-0.19\,n].$$

Combining this with (51) and using the Cauchy–Schwarz inequality, we obtain

$$\mathbf{E}\Big[U_{(i)}^{-k} - U_{(i)} - U_{(i)}^{-k}e_k \Big/ \sum_{s\neq k} e_s + \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s\Big]^2$$

$$= \mathbf{E}\Big[U_{(i)}^{-k} - U_{(i)} - U_{(i)}^{-k}e_k \Big/ \sum_{s\neq k} e_s + \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s\Big]^2 \mathbf{1}\{\mathbf{e}^n \in \mathcal{A}^n\}$$

$$+\mathbf{E}\Big[U_{(i)}^{-k} - U_{(i)}^{-k}e_k \Big/ \sum_{s\neq k} e_s + \mathbf{1}\{k \leqslant i\}e_k \Big/ \sum_{s\neq k} e_s\Big]^2 \mathbf{1}\{\mathbf{e}^n \notin \mathcal{A}^n\} \leqslant \frac{C}{n^4}.$$

□

Lemma 9. *For some $C > 0$, uniformly in $x \in [0, \sqrt{n}/2]$*

$$\mathbf{P}\left\{\max_i \sqrt{n}\left|U_{(i)} - \frac{i}{n}\right| \geqslant x\right\} \leqslant \exp(-Cx^2). \tag{53}$$

Proof. Let $\mathcal{F}_i = \sigma(e_1, \dots, e_i)$. Since for any given $\lambda \in (0,1)$,

$$X_\lambda(t) = \exp\left[\lambda \sum_{k=1}^{t}(e_k - 1) + t[\log(1-\lambda) + \lambda]\right]$$

is martingale such that $\mathbf{E}X_\lambda(t) = 1$, by the Doob inequality we obtain

$$\mathbf{P}\left\{\max_{1 \leqslant t \leqslant n} X_\lambda(t) \geqslant x^2\right\} \leqslant \exp(-x^2).$$

Therefore choosing $\lambda = x/\sqrt{n}$, we obtain

$$\mathbf{P}\left\{\max_{1 \leqslant t \leqslant n} \frac{1}{\sqrt{n}} \sum_{k=1}^{t}(e_k - 1) \geqslant x - \frac{n}{x}\left[\log\left(1 - \frac{x}{\sqrt{n}}\right) + \frac{x}{\sqrt{n}}\right]\right\} \leqslant \exp(-x^2).$$

Similarly

$$\mathbf{P}\left\{\max_{1 \leqslant t \leqslant n} \frac{1}{\sqrt{n}} \sum_{k=1}^{t}(1 - e_k) \geqslant x - \frac{n}{x}\left[\log\left(1 + \frac{x}{\sqrt{n}}\right) - \frac{x}{\sqrt{n}}\right]\right\} \leqslant \exp(-x^2).$$

Next notice that

$$F(z) = \frac{\log(1-z) + z}{z^2}, \quad z \in (-\infty, 1]$$

is decreasing and $F(0.5) = -0.7726$. Therefore, for any $x \leqslant \sqrt{n}/2$

$$\mathbf{P}\left\{\max_{1 \leqslant t \leqslant n} \frac{1}{\sqrt{n}} \left|\sum_{k=1}^{t}(1 - e_k)\right| \geqslant x\right\} \leqslant 2\exp(-x^2/4).$$

Combining this inequality with

$$\begin{aligned}
&\sqrt{n}\max_i \left|U_{(i)} - \frac{i}{n}\right| \\
&\quad = \max_i \left|\frac{1}{\sqrt{n}} \sum_{k=1}^{i}(e_k - 1) - \frac{i}{n^{3/2}} \sum_{k=1}^{n}(e_k - 1)\right| \Big/ \left[1 + \frac{1}{n}\sum_{k=1}^{n}(e_k - 1)\right] \\
&\leqslant \left[\max_i \frac{1}{\sqrt{n}}\left|\sum_{k=1}^{i}(e_k - 1)\right| + \frac{1}{\sqrt{n}}\left|\sum_{k=1}^{n}(e_k - 1)\right|\right] \Big/ \left[1 + \frac{1}{n}\sum_{k=1}^{n}(e_k - 1)\right],
\end{aligned}$$

we finish the proof. □

Lemma 10.

$$\mathbf{E}\left[n\log\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)-\sum_{i=1}^{n}(e_i-1)\right]^2\leqslant C.$$

Proof. Note that since $\log(1+x)\leqslant x$,

$$\mathbf{E}\left[\log\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)\right]_+^4\leqslant\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(e_i-1)\right)^4\leqslant\frac{C}{n}.$$

On the other hand,

$$\mathbf{E}\left[-\log\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)\right]_+^4\leqslant\mathbf{E}[-\log(\min_{i=1,n}e_i)]_+^4$$
$$=\mathbf{E}[-\log(e_1/n)]_+^4\leqslant C\log^4(n)$$

and therefore

$$\mathbf{E}\log^4\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)\leqslant C\log^4(n).$$

Let

$$\mathcal{A}^n=\left\{x_i\geqslant 0:\ \left|\sum_{i=1}^{n}(x_i-1)\right|\leqslant\frac{n}{2}\right\}.$$

It was already proved (see (52)) that for some $C>0$

$$\mathbf{P}\{\mathbf{e}^n\notin\mathcal{A}^n\}\leqslant\exp(-Cn).$$

Therefore using the Taylor formula and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}
&\mathbf{E}\left[n\log\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)-\sum_{i=1}^{n}(e_i-1)\right]^2\\
&=\mathbf{E}\left[n\log\left(1+\frac{1}{n}\sum_{i=1}^{n}(e_i-1)\right)-\sum_{i=1}^{n}(e_i-1)\right]^2\mathbf{1}\{\mathbf{e}^n\in\mathcal{A}^n\}\\
&+\mathbf{E}\left[n\log\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)-\sum_{i=1}^{n}(e_i-1)\right]^2\mathbf{1}\{\mathbf{e}^n\notin\mathcal{A}^n\}\\
&\leqslant C\mathbf{E}\frac{1}{n}\left(\sum_{i=1}^{n}(e_i-1)\right)^2+2n^2\mathbf{E}\log^2\left(\frac{1}{n}\sum_{i=1}^{n}e_i\right)\mathbf{1}\{\mathbf{e}^n\notin\mathcal{A}^n\}\\
&+\mathbf{E}\left[\sum_{i=1}^{n}(e_i-1)\right]^2\mathbf{1}\{\mathbf{e}^n\notin\mathcal{A}^n\}\leqslant C.\qquad\square
\end{aligned}$$

Lemma 11.

$$\mathbf{E}\left[\frac{1}{n+1}\sum_{k=1}^{n+1}e_k\right]^{-1}\left[U_{(i)}-\frac{i}{n+1}\right]=O\left(\frac{1}{n}\right).$$

Proof. We use the same argument as in the proof of Lemma 10. First of all notice that

$$\sum_{k=1}^{i}e_k-\frac{i}{n+1}\sum_{k=1}^{n+1}e_k=\sum_{k=1}^{i}(e_k-1)-\frac{i}{n+1}\sum_{k=1}^{n+1}(e_k-1)=O\big(\sqrt{n}\big).$$

Next, by the Taylor formula,

$$\left[\frac{1}{n+1}\sum_{k=1}^{n+1}e_k\right]^{-2}=1-\frac{2}{n+1}\sum_{k=1}^{n+1}(e_k-1)+\frac{3}{(n+1)^2}\left[\sum_{k=1}^{n+1}(e_k-1)\right]^2+O\left(\frac{1}{n^{3/2}}\right).$$

Therefore

$$\begin{aligned}\mathbf{E}&\left[\frac{1}{n+1}\sum_{k=1}^{n+1}e_k\right]^{-1}\left[U_{(i)}-\frac{i}{n+1}\right]\\&=\mathbf{E}\left[\frac{1}{n+1}\sum_{k=1}^{n+1}e_k\right]^{-2}\left[\sum_{k=1}^{i}e_k-\frac{i}{n+1}\sum_{k=1}^{n+1}e_k\right]\\&=\mathbf{E}\left\{-\frac{2}{n+1}\sum_{k=1}^{n+1}(e_k-1)+\frac{3}{(n+1)^2}\left[\sum_{k=1}^{n+1}(e_k-1)\right]^2\right\}\\&\qquad\left[\sum_{k=1}^{i}e_k-\frac{i}{n+1}\sum_{k=1}^{n+1}e_k\right]+O\left(\frac{1}{n}\right).\end{aligned}$$

Finally, it is easy to see with a simple algebra that

$$\mathbf{E}\sum_{k=1}^{n+1}(e_k-1)\sum_{k=1}^{i}e_k=i,\quad\mathbf{E}\left[\sum_{k=1}^{n+1}(e_k-1)\right]^2\sum_{k=1}^{i}e_k=ni+3i.$$

So, substituting this in (54), we finish the proof. □

References

1. J. Beirlant, *Limit theory for spacings og order statistics from general univariate distribution.* — Publ. Inst. Stat. Univ. Paris XXXI, fasc. 1, (1986) 27–57.
2. J. Beirlant, M. C. A. van Zuijlen, *The impirical distribution function and strong laws for functions of order statistics of uniform spacings.* — J. Multivar. Anal. **16** (1985), 300–317.
3. N. Cressie, *On the logarithm of high-order spasings.* — Biometrica **63**, no 2 (1976), 343–355.
4. D. A. Darling, *On a class of problems related to the random division of an interval.* — Ann. Math. Statist. **24** (1953), 239–253.
5. P. Deheuvels, G. Derzko, *Exact laws for sums of logarithms of uniform spacings* — Austr. J. Statist. **32** no. 1& 2 (2003), 29–47.
6. F. El Haje, Yu. Golubev, P.-Y. Liardet, Ya. Teglia, *On statistical testing of random numbers generators.* — In: Security and Criptography for Networks. De Prisco and Yung (eds.) Springer-Verlag Berlin Heidelberg (2006), pp. 271–287.
7. R. Hasminskii, I. Ibragimov, *On the nonparametric estimation of functionals.* — In: Proc. II Prague Symp. Asymptotic Statistics (1978), pp. 41–51.
8. I. Ibragimov, R. Khasminskii, *Statistical Estimation: Asymptotic Theory*, Springer Verlag (1981).
9. I. Ibragimov, A. Nemirovski, R. Khasminskii, *Some problems of nonparametric estimation in Gaussian white noise.* — Theory Probab. Appl. **31** (1986), no. 3, 391–406.
10. L. F. Kozachenko, N. N. Leonenko, *On statistical estimation of entropy of a random vector.* — Probl. Inform. Transmission **23** (1987), 95–101.
11. B. Levit, *On the efficiency of a class of nonparametric estimates.* — Theor. Probab. and Applic. **20** (1975), 723–74.
12. A. Nemirovski, *Topics in Nonparametric Statistics.* — In: M. Emery, A. Nemirovski, D. Voiculescu, Lectures on Probability Theory and Statistics, Ecole d'ete de Probabilities de Saint-Flour XXVIII (1998), (Editor: P. Bernard), Lect. Notes Math. 1738, Springer Verlag (2000).
13. R. Pyke, *Spacings (with discussion).* — J. R. Statist. Soc. **7** (1965), 395–449.
14. A. B. Tsybakov, E. C. Van der Meulen, *Root-n consistent estimators of entropy for densities with unbounded supprot.* — Discussion paper 9206, Universié Catholique de Louvain (1992).
15. A. W. Van der Vaart, *Asymptotic Statistics* Cambridge University Press (1998).

Université Rennes 2,
Rennes, France

E-mail: fidahh@hotmail.fr

CNRS, Université de Provence,
Marseille, France
Institute for Problems of Information
Transmission, Moscow, Russia

E-mail: golubev.yuri@gmail.com

Поступило 16 декабря 2008 г.