The empirical identity process: Asymptotics and applications

The Canadian Journal of Statistics
Vol. xx, No. yy, 2017, Pages 1–25
La revue canadienne de statistique

The empirical identity process: asymptotics and applications

Enrico Bibbona [1]*, Giovanni Pistone [2] and Mauro Gasparini [1]

[1] *Dipartimento di Scienze Matematiche "G.L. Lagrange", Politecnico di Torino, Italy*
[2] *de Castro Statistics, Collegio Carlo Alberto, Moncalieri , Italy*

Key words and phrases: strong approximations, uniform empirical process, uniform quantile process

MSC 2010: Primary 62G30 ; secondary 60F17, 62G30

Abstract:

When sampling independent observations drawn from the uniform distribution on the unit interval, the asymptotic behavior of both the empirical distribution function and empirical quantile function (as the sample size gets large) is well known. In this paper we study analogous asymptotic results for the function that is obtained by composing the empirical quantile function with the empirical distribution function. Since the former is the generalized inverse of the latter, the result will approximate the identity function. We define a scaled and centered version of this function – the *empirical identity process* – and prove it converges to a highly irregular limit process whose trajectories are not right continuous and impossible to study using standard probability in metric spaces. However, when this process is integrated over time, and appropriately rescaled and centered, it becomes possible to define a functional limit theorem for it, which then converges to a randomly pinned Brownian motion. By applying these theoretical results, a new goodness-of-fit test is derived. We demonstrate that this test is very efficient when it is applied to data which come from a multimodal or mixture distribution, like the classic *old faithful* dataset.

The Canadian Journal of Statistics xx: 1–25; 2017

Résumé: Insérer votre résumé ici. We will supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–25; 2017

1. INTRODUCTION

The asymptotic behavior of both the uniform empirical process and the uniform quantile process is well known (e.g., Csörgő and Révész (1981), Csörgő (1983), Shorack and Wellner (1986)). Such processes are defined as centered and scaled versions of the empirical distribution function and the empirical quantile function,respectively. To the best of our knowledge this is the first investigation of the process resulting from applying the empirical quantile function to the empirical distribution function itself. Intuitively, such a back-and-forth operation approximates the identity function. We introduce a centered and scaled version of this process which we define as the *empirical identity process* (EIP). The asymptotic properties of the EIP are somewhat unexpected: the EIP converges in distribution to a white noise process whose finite dimensional distributions are products of exponential distributions (or Laplace distributions, see the discussion around Theorem 1). This limiting process is very irregular and in particular it is not right-continuous. It is therefore impossible to build a proper weak convergence theory in any metric space. In the hope of gaining regularity, we study the integral of the EIP. The resulting limit theorem is a functional

* *Author to whom correspondence may be addressed.*
E-mail: enrico.bibbona@polito.it
mauro.gasparini@polito.it
giovanni.pistone@gmail.com

version of a classical result by Moran (1947) on the asymptotic behavior of the sum of squared spacings.

Our results provide some unexpected asymptotic properties of a process related to the uniform empirical and quantile function. Moreover, we demonstrate that these asymptotic results have interesting applications from a statistical point of view. Based on the asymptotics of the EIP, we propose a new test statistic that can be used in goodness-of-fit problems. The new methodology is not always superior to other methods based on the empirical distribution function, but it performs better when the true distribution is multimodal or a mixture. Using simulations, we identify cases in which this new statistic outperforms existing alternatives and also provide a relevant application to the popular *old faithful* dataset.

2. EMPIRICAL IDENTITY PROCESSES, AND ASYMPTOTIC RESULTS

Let $U_1,\ U_2 \ldots, U_n$ be independent random variables from a uniform distribution on $[0,1]$. Let $U_{n,1} \leq\ \ldots\ \leq U_{n,n}$ be their order statistics, $U_{n,0} = 0$ and $U_{n,n+1} = 1$. Let $\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n}(U_i \leq t)$, $0 \leq t \leq 1$, denote the empirical distribution function and $\mathbb{Q}_n(u) = \inf\{t \in [0,1] :\ \mathbb{F}_n(t) \geq u\}$ the empirical quantile function, $0 < u \leq 1$. $\mathbb{Q}_n(\cdot)$ is the left-continuous generalized inverse function of $\mathbb{F}_n(\cdot)$.

Define the *lower empirical identity function* as

$$R_n^L(t) = U_{n,n\mathbb{F}_n(t)} = \begin{cases} 0 & \text{if } 0 \leq t < U_{n,1} \\ \mathbb{Q}_n(\mathbb{F}_n(t)) & \text{if } U_{n,1} \leq t \leq 1\ , \end{cases}$$

the *upper empirical identity function* as

$$R_n^U(t) = U_{n,n\mathbb{F}_n(t)+1} = \begin{cases} \mathbb{Q}_n\left(\mathbb{F}_n(t) + \frac{1}{n}\right) & \text{if } 0 \leq t < U_{n,n} \\ 1 & \text{if } U_{n,n} \leq t \leq 1 \end{cases}$$

and the *empirical identity function* as their average

$$R_n(t) = \frac{R_n^L(t) + R_n^U(t)}{2}.$$

The trajectories of $R_n(t)$, $R_n^L(t)$, and $R_n^U(t)$ for a specific sample of size $n = 2$ are shown in Figure 1. By the Glivenko-Cantelli theorem, as $n \to \infty$, the three random sequences, $R_n(t)$, $R_n^L(t)$, and $R_n^U(t)$, converge almost surely in the uniform norm to the identity function. It is therefore interesting to study their second order asymptotics by defining the *lower* and *upper empirical identity process*

$$Y_n^L(t) = (n+1)(R_n^L(t) - t) \quad \text{and} \quad Y_n^U(t) = (n+1)(R_n^U(t) - t),$$

and the *empirical identity process* (EIP)

$$Y_n(t) = (n+1)(R_n(t) - t) = \frac{1}{2}(Y_n^L(t) + Y_n^U(t))$$

for $0 \leq t \leq 1$. We use the scaling factor $n+1$ instead of n to make notation easier in the following sections.

Theorem 1. *For any positive integer k and points $0 < u_1 < \cdots < u_k < 1$, the random vector $\left(Y_n^U(u_1), \cdots, Y_n^U(u_k), -Y_n^L(u_1), \cdots, -Y_n^L(u_k)\right)$ converges in distribution to a vector of $2k$ independent exponential random variables, as $n \to \infty$.*

In other words, the joint finite dimensional distributions (FDDs) of the bivariate processe $\big(Y_n^U(\cdot), -Y_n^L(\cdot)\big)$ converge to those of two independent exponential variates. As a consequence, for the EIP itself we can conclude that

Corollary 1. *For any positive integer k and points $0 < u_1 < \cdots < u_k < 1$, the random vector $\big(Y_n(u_1), \cdots, Y_n^U(u_k)\big)$ converges in distribution, as $n \to \infty$, to a vector of k independent random variables with the so called Laplace density $f(z) = \exp(-2|z|)$.*

The proof of Theorem 1 is long and requires some preliminary lemmas. It is sketched in the Appendix. A process with independent finite-dimensional distributions cannot be right-continuous (see the criterion in Theorem 13.6 of Billingsley (1999)). Therefore the theory of weak convergence in the usual space of cadlag paths $D(0,1)$ does not apply. The conclusion is that such limit processes are fairly intractable objects and for any statistical application we need to regularize them.

3. THE INTEGRATED EMPIRICAL IDENTITY PROCESS

The anti-derivative of a function is always more regular then the function itself. Therefore, in the hope of obtaining a more regular limit process, it is natural to look at the asymptotic behavior of the integrals of the processes defined in the previous section.

Since it turns out that the asymptotic behaviors of the lower and upper EIPs are equivalent (see section 3.2 for further details), it is simpler and notationally convenient to study only the integrated lower EIP defined as

$$I_n(t) = -\int_0^t Y_n^L(u)du = (n+1)\int_0^t [u - R_n^L(u)]du, \quad t \in [0,1] \tag{1}$$

(the minus sign is to make it non-negative), called simply the *integrated process* hereafter.

3.1. The integrated process an its relation with spacings

A simple geometric inspection of Figure 1 shows that the integrated process is strictly related to the uniform *spacings*, defined as

$$D_{n,i} = U_{n,i} - U_{n,i-1}, \qquad i = 1 \ldots n+1. \tag{2}$$

We obtain

$$\begin{aligned} I_n(t) =& (n+1)\int_0^t [u - R_n^L(u)]\, du \\ =& (n+1)\sum_{i=1}^{nF_n(t)} \int_{U_{n,i-1}}^{U_{n,i}} (u - U_{n,i-1})\, du + (n+1)\int_{U_{n,nF_n(t)}}^{t} [u - U_{n,nF_n(t)}]du \\ =& \frac{n+1}{2}\sum_{i=1}^{nF_n(t)} D_{n,i}^2 + \frac{n+1}{2}(t - R_n^L(t))^2 \end{aligned} \tag{3}$$

In particular, the integrated process evaluated at $t = 1$ equals

$$I_n(1) = (n+1)\int_0^1 [u - R_n^L(u)]\, du = \frac{n+1}{2}\sum_{i=1}^{n+1} D_{n,i}^2.$$

This is the well-known Greenwood statistic, for which a classical theorem due to Moran states convergence to normality in the following way:

Theorem 2. (Moran 1947.) *The following convergence in law holds*

$$M_n = \sqrt{n+1}(I_n(1)-1) = \sqrt{n+1}\left(\frac{n+1}{2}\sum_{i=1}^{n+1} D_{n,i}^2 - 1\right) \Rightarrow N(0,1). \tag{4}$$

Theorem 2 refers to the convergence of $I_n(1)$, the value taken by the integrated process at the final coordinate $t=1$. More can be said about the convergence of $I_n(\cdot)$ as a process in the functional space $D(0,1)$. In other words, we are in a position to extend Moran's theorem into a *functional version*. To do so, we need some technical steps which involve so called strong approximation theorems. The rest of this section can be skipped by those not interested in probabilistic details; the important result is Theorem 4 of the next subsection.

Recall first the following strong approximation results of Aly (1983) and Aly (1988):

Theorem 3. *(Aly 1983, 1988) There exists a probability space on which both*

- *a two dimensional Wiener process $(W_1(\cdot), W_2(\cdot))$ with zero mean and autocovariance matrix*

$$\mathbb{E}\left[\begin{pmatrix} W_1(s) \\ W_2(s)) \end{pmatrix} (W_1(t)\ W_2(t))\right] = \min\{s,t\}\begin{pmatrix} 1 & 4 \\ 4 & 20 \end{pmatrix}, \quad t,s>0$$

- *and a vector $\{D_{n,i}\}_{i=1\ldots,n+1}$ of random variables of arbitrary size n with the same law as the uniform spacings (2)*

are defined, such that the two processes $E_n(t)$ and $V_n(t)$, $0 \le t \le 1$, defined as

$$E_n(t) = \begin{cases} 0 & \text{if} \quad 0 \le t < \frac{2}{n+1} \\ \sqrt{n+1}\left((n+1)\sum_{i=1}^{\lfloor (n+1)t \rfloor} D_{n,i}^2 - 2t\right) & \text{if} \quad \frac{2}{n+1} \le t \le 1 \end{cases}$$

$$V_n(t) = \frac{1}{\sqrt{n+1}}\left[W_2((n+1)t) - 4tW_1(n+1)\right] \tag{5}$$

are so close to each other that the following condition holds: for every $\varepsilon > 0$ there are constants A,B such that

$$P\left[\sup_{0\le t\le 1} |E_n(t) - V_n(t)| > A\frac{\log(n+1)}{\sqrt{n+1}}\right] \le B(n+1)^{-\varepsilon}. \tag{6}$$

Theorem 3 was first stated in Aly (1983), but a minor step of the proof was not fully justified. This led the author to write a second paper (Aly (1988)) with a rigorous proof of Theorem 3, based on a multivariate Hungarian construction due to Einmahl (1989). However, the rate of convergence in Einmahl (1989) is suboptimal and not fast enough to allow for the rate $\log(n+1)/\sqrt{n+1}$ in Equation (6). Theorem 3 is then provided in Aly (1988) with a rate slowed to $(\log(n+1))^2/\sqrt{n+1}$. We here conclude that the statement in Aly (1983) is actually correct. The reason is that a stronger multivariate Hungarian construction is now available that allows us to prove Theorem 3 with the same proof as in Aly (1988), as long as Zaitsev (1998) is cited in place of Einmahl (1989). The process $V_n(\cdot)$ defined in Equation (5) is constructed by applying such bivariate Hungarian construction. Different processes arise for different n, but their law is

actually the same irrespective of n, since they are all centered Gaussian processes with covariance function given by

$$\mathbb{E}(V_n(t)V_n(s)) = 20\,\min(t,s) - 16\,s\,t.$$

Moreover, each of the V_n has the same law as the process

$$V(t) = 2\sqrt{5}W(t) - 2(\sqrt{5}-1)\,tW(1), \quad 0 \le t \le 1, \tag{7}$$

where $W(\cdot)$ is a standard one-dimensional Wiener process. A consequence of Theorem 3 and of the previous remarks is the following corollary

Corollary 2. *The process $E_n(\cdot)$ converges weakly in $D(0,1)$ to the process $V(\cdot)$ defined by Equation (7).*

3.2. A new functional version of Moran's theorem

Since the statistic M_n introduced in Equation (4) equals $E_n(1)/2$, the previous corollary may already be considered a functional version of Moran's theorem. However, it involves the process $E_n(\cdot)$ and not the process $I_n(\cdot)$ which is of interest here. We therefore need another functional version of Moran's theorem which provides the asymptotics for the process $I_n(\cdot)$ directly.

Theorem 4. *The process $2\sqrt{n+1}\,(\,I_n(\cdot) - \mathbb{F}_n(\cdot)\,)$ converges weekly in $D(0,1)$ to the Gaussian process $V(\cdot)$ of Equation (7).*

A full proof is given in the Appendix.

Due to the continuous mapping theorem and to the continuity of the $\sup$ operator and of the absolute value, we can deduce the following corollary, which will be used in the next sections to build a new goodness-of-fit test.

Corollary 3. *The random variable $2\sup_{0\le t\le 1}\sqrt{n+1}\,|I_n^L(t) - \mathbb{F}_n(t)|$ converges weakly to $\sup_{0\le t\le 1}|V(t)|$.*

At the beginning of Section 3 we introduced the integrated process focusing on the integral of the lower EIP. However it is legitimate to consider what would happen if we instead used the integral of the upper EIP

$$I_n^U(t) = \int_0^t Y_n^U(u)du = (n+1)\int_0^t [R_n^U(u) - u]du, \quad t \in [0,1]$$

or if we considered the joint distribution of the two. To give a satisfactory answer a more formal proof would be required, but it can be seen from Figure 1 that the difference between $I_n(t)$ and $I_n^U(t)$ is uniformly bounded by $(n+1)/2$ times the squared maximal spacing. Therefore one can apply the classical results in Slud (1978) on the almost sure rate of converence to zero of the maximal spacing to show that the difference between $\sqrt{n+1}\,(I_n(t) - \mathbb{F}_n(t))$ and $\sqrt{n+1}\,(I_n^U(t) - \mathbb{F}_n(t))$ is almost surely vanishing when the sample size tends to infinity. Consequently, the couple $(\sqrt{n+1}\,(I_n(t) - \mathbb{F}_n(t)),\ \sqrt{n+1}\,(I_n^U(t) - \mathbb{F}_n(t)))$ converges weakly to a couple of identical copies of the process $V(\cdot)$ defined in Equation (7).

3.3. Characterization of the limit process and the asymptotic distribution of the maximum

To use Corollary 3, we need to explicitely compute the distribution of the $\sup$ of the limit process. We are able to derive an explicit distribution in Theorem 5.

Let $W(\cdot)$ be standard Brownian motion and $B(t) = W(t) - tW(1),\ t \geq 0$ be a Brownian bridge. For every t, $B(t)$ is independent of $W(1)$ and it has the distribution of Brownian motion which is constrained to visit zero (or "pinned to 0") at time 1. A process $B(t) + ty,\ t \geq 0$ also represents Brownian motion pinned to y when $t = 1$ (cf. Revuz and Yor (1999), page 37). The limit process $V(\cdot)$ defined in Equation (7) admits therefore the equivalent representation $V(t) = 2\sqrt{5}\left(B(t) + t\frac{W(1)}{\sqrt{5}}\right),\ t \geq 0$. It can be seen as Brownian motion which at time 1 is pinned to a random position $W(1)/\sqrt{5}$, scaled by a factor $2\sqrt{5}$. We can derive the so-called two-sided maximal probability distribution for $V(t)$ from that of the pinned Brownian motion (cf. Beghin and Orsingher (1999), Equation 4.12 or Borodin and Salminen (2002), Part II, Chapter 1, Equation 1.15.8(1)).

Theorem 5. *The distribution function of the maximum of the absolute value of the stochastic process $V(\cdot)$ is given by*

$$\mathbb{P}\left(\sup_{0\leq t\leq 1}|V(t)| < b\right) = \sum_{h=-\infty}^{\infty}(-1)^h \mathrm{e}^{-\frac{4}{50}h^2b^2} \tag{8}$$

A detailed proof is given in the Appendix.

4. A NEW GOODNESS-OF-FIT TEST

Let $X_1 \dots X_n$ be a sample of independent continuous random variables, not necessarily with identical distributions. Under the null hypothesis H_0 that the distribution functions of the X_i are some given $F_i(x), i = 1, \dots, n$, the transformed sample $\{\hat{U}_i = F_i(X_i)\}_{i=1\cdots n}$ is composed of independent uniform random variables. The integrated process of the transformed sample can be used to construct a new goodness-of-fit test of H_0 in the same spirit as the Kolmogorov-Smirnov goodness-of-fit test.

4.1. A statistic derived from the sup of the integrated EIP

Let $\hat{\mathbb{F}}_n(t)$ be the empirical distribution function of the transformed sample and let $\hat{I}_n(t)$ be the related integrated process (Equation (1)). We define the test statistic

$$d_n = 2\sup_{0\leq t\leq 1}\sqrt{n+1}\,|\hat{I}_n(t) - \hat{\mathbb{F}}_n(t)|. \tag{9}$$

Under H_0, the sequence d_n converges weakly to $\sup_{0\leq t\leq 1}|V(t)|$, whose distribution is given by Equation (8). Let us remark that the distribution of d_n is the same, irrespective of the distributions $F_i(x)$ of the single observations.

Define now a new goodness-of-fit test which rejects the null hypothesis if the value of d_n is larger then a critical value. An asymptotic critical value can be derived by numerically inverting Equation (8), e.g. the 95% percentile of $\sup_{0\leq t\leq 1}|V(t)|$ equals $b = 6.790494$. Numerical simulations show that the convergence is slow. For intermediate values of n (e.g $n = 100$) the 0.95 quantiles of d_n are not well approximated by the asymptotic values, but they can easily be derived using Monte Carlo methods. Results are summarized in Table 1.

4.2. Numerical simulations

We have run some numerical experiments to benchmark the performance of these new tests just against other well-known goodness-of-fit tests like that based on the original statistic M_n of Moran (Equation (4)). Other competitors are the classical Kolmogorov-Smirnov test, which is

based on the statistic

$$D_n = \sup_t \sqrt{n}|\mathbb{F}_n(t) - F(t)| = \sup_t \sqrt{n}|\hat{\mathbb{F}}_n(t) - t|$$

and the Anderson-Darling test, which belongs to the Cramer-von Mises family of tests and is based on the statistic

$$A_n = n \int_0^1 w(u) \left(\hat{\mathbb{F}}_n(u) - u\right)^2 du$$

with the weight function $w(x) = 1/[x(1-x)]$, designed to pick up possible dissimilarities in the tails.

For goodness-of-fit tests of this kind, the alternative hypothesis is completely nonparametric. For data not generated according to the null distribution, the power of the test is strongly influenced by the choice of the *alternative distribution* from which they are drawn. For example, if we test normality of a sample which was generated from a Student-t distribution with the same mean, we expect a test based on A_n to have higher power than one based on D_n, due to the difference between the null and alternative distribution in the tails.

Now, if the dataset is generated from a mixture of normal distributions having the same variance σ^2 but different means μ_1 and μ_2, we expect that a test based on the empirical distribution may not easily find significant dissimilarities between the data and a normal distribution with mean equal to a linear combination of μ_1 and μ_2, and variance somewhat larger then σ^2 (say, the variance of the mixture). However a test based on the spacings, such as the one using d_n, could achieve greater power. The intuitive justification is that the gap in the uniformized data between the two modes would easily give rise to some large spacings that could make d_n significantly larger than in the uniform case.

We therefore check the power of the different tests by simulation in situations where the data come from mixture distributions. This is first to highlight situations where d_n performs better than existing tests. For completeness, we also study a second set of examples in less favorable conditions.

In all cases the power of the test is approximated by simulation as the ratio between the number of rejections and the number of simulated samples, since the generating distribution is always different from the null.

All numerical experiments are done using the R environment for statistical computing (R Core Team (2017)). We use built-in functions to generate samples and to compute the value of the test statistic D_n. For A_n we use the `goftest` package. To compute M_n and d_n we use our own code. Our R scripts are included as Supplementary Material in order to ensure reproducibility and to allow the reader to try new cases.

We perform Monte Carlo calculations of the 0.95 quantiles for the different test statistics over 10^6 uniform samples of lengths $30, 50, 100, 200,$ and 272 respectively. Such values are then used as critical values for the tests, in order to make a fair comparison of the methodologies. The quantiles are displayed in Table 1. Note that while the quantiles of A_n and D_n are already very close to their asymptotic values when $n = 30$, this is not true for d_n and M_n, their convergence is much slower.

We next evaluate the power of the different tests for some examples.

In particular we choose a first set of examples whose common feature is that data are simulated from a mixture distribution, while the null hypothesis is that they come from a single component:

- The null hypothesis is that our sample is taken from the standard normal distribution. The 10000 simulated samples of size 100 are generated from a mixture of two normal distributions with equal weights, common standard deviation 0.45 and means 0.88 and -0.88. Numbers are chosen so that the mixture distribution has approximatively zero mean and unit variance. We call this alternative distribution a symmetric normal mixture.
- The null hypothesis is that our sample is taken from the standard normal distribution. The 10000 simulated samples of size 30 are generated from a mixture of two normal distributions. The weights of the mixture are $p_1 = 1/5$ and $p_2 = 4/5$, the means are $\mu_1 = 1.68$ and $\mu_2 = 0.42$ and the standard deviations $\sigma_1 = 0.2$ and $\sigma_2 = 0.6$. Again, numbers are chosen so that the mixture distribution has approximatively zero mean and unit variance. We call such an alternative distribution an asymmetric normal mixture.
- The null hypothesis is that our sample is taken from the uniform distribution on $(0, 1)$. The 10000 simulated samples of size 100 are generated from a mixture of two beta distributions with parameters $(2, 8)$ and $(8, 2)$ and equal weights.

The results are summarized in Table 2. Both the d_n and M_n statistics, which are based on spacings, outperform the tests based on the empirical distribution. Moreover, the test based on d_n has the best power in all three examples.

Of course no test is uniformly best again all types of alternatives. In more standard situations we do not expect our methodology to be superior to the classical methods based on the empirical distribution function. A second group of examples follows. The null hypothesis is always that our sample is taken from the standard normal distribution while the alternative is listed below:

- the 10000 simulated samples of size 100 are drawn from a Cauchy distribution with scale parameter 0.5;
- the 10000 simulated samples of size 100 are generated from a Student's t distribution with 2 degrees of freedom;
- the 10000 simulated samples of size 100 are generated from a normal distribution with mean 0.3.

The results are summarized in Table 3. Both the d_n and M_n statistics, based on spacings, are inferior to A_n which has the best power in all three examples.

The critical values in Table 1 are obtained by simulating from the uniform distribution. There might be a loss in accuracy when data are generated from other distributions and transformed through the cumulative distribution function. To check for such a possibility we perform the following consistency check: we test the goodness of fit with respect to the true distribution of the data, looking for the rate of occurrence of type I errors (see the code in the supplementary material). We do not report the values here, but we noted no relevant discrepancies with respect to the nominal level of 0.05.

4.3. An application to the Old Faithful dataset

Old faithful is a geyser in Yellowstone National Park, Wyoming, USA. For centuries it has been erupting several times a day, spewing streams of hot water high into the sky. A popular dataset, consisting of records of waiting times between eruptions and of their durations, is distributed with R (R Core Team (2017)). The dataset contains 272 observations of both waiting times and durations. We focus on the waiting times, which show a bimodal distribution as illustrated in Figure 2. The data are in minutes (integers) and in order to avoid ties we jitter the data by adding Gaussian noise with zero mean and standard deviation 0.4. The sample mean (before jittering) is 70.90 and the sample standard deviation is 13.59. All tests reject normality of the sample if the null mean is fixed to 71 and the null standard deviation to 14. We do not report p-values, but the

code is available as Supplementary Material. If subsamples of size 50 are taken from the dataset, then rejection of the same null hypothesis is not obvious anymore. We have selected at random 10000 subsamples of size 50 and run all four tests on each one. The results are summarized in Table 4. The test based on d_n seems to be able to reject the null hypothesis of normality with a higher power.

5. CONCLUSIONS

We have defined the lower EIP and found that it converges to a process which has highly irregular trajectories. In the hope of gaining regularity we have studied the limiting behavior of its integral, obtaining Theorem 4 as our main result. We have also computed the explicit limiting distribution of the running sup of the integrated process given in Theorem 5. This result has an important statistical application in the construction of a new goodness-of-fit test based on spacings, as illustrated in Section 4. An application to the classic *old faithful* dataset supports the conclusion that this new test is useful for multimodal data, coming for example from mixtures of distributions.

6. FIGURES AND TABLES

TABLE 1: Approximate values of the 0.95 quantile of the distributions of d_n, M_n, D_n and A_n for different sample sizes n computed by Monte Carlo simulations. In the last column their asymptotic values.

stat \ n	30	50	100	200	272	∞
d_n	5.857	6.127	6.349	6.493	6.536	6.790
M_n	1.449	1.559	1.637	1.679	1.684	1.645
A_n	2.493	2.497	2.495	2.494	2.490	2.492
D_n	1.322	1.332	1.341	1.346	1.346	1.358

TABLE 2: First group of examples (with mixtures). Power of the test at 5% significance level for the new test based on $d_n = 2\sup_{0\le t\le 1}\sqrt{n+1}\,|\hat{I}_n(t) - \hat{\mathbb{F}}_n(t)|$ in comparison with the Moran test based on M_n, the Kolmogorov Smirnov test based on D_n and the Anderson Darling test based on A_n.

H_0	true distribution	power of d_n	power of M_n	power of D_n	power of A_n
Normal	symmetric Normal mixture	0.698	0.644	0.618	0.521
Normal	asymmetric Normal mixture	0.553	0.527	0.232	0.179
Uniform	Beta mixture	0.858	0.788	0.792	0.753

ACKNOWLEDGEMENTS

G.P. acknowledges the support of Collegio Carlo Alberto, Torino. He is a member of INdAM-GNAMPA. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and from the project DISMA - Dipartimento di Eccellenza. We thank Ivo Stoepker for having carefully read the manuscript and eliminated a few important errors.

TABLE 3: Second group of examples (without mixtures). Power of the test at 5% significance level for the new test based on $d_n = 2\sup_{0\le t\le 1}\sqrt{n+1}\,|\hat{I}_n(t) - \hat{\mathbb{F}}_n(t)|$ in comparison with the Moran test based on M_n, the Kolmogorov Smirnov test based on D_n and the Anderson Darling test based on A_n.

H_0	true distribution	power of d_n	power of M_n	power of D_n	power of A_n
Normal	Cauchy	0.657	0.769	0.261	0.998
Normal	Student t	0.346	0.383	0.252	0.982
Normal	Shifted Normal	0.554	0.207	0.732	0.824

TABLE 4: Results of the analysis of subsamples of the waiting times between the eruptions of the *old faithful* geyser.

Total samplesize	272
Number of subsamples	10000
Size of subsamples	50
Rejections by d_n	6692
Rejections by M_n	6282
Rejections by D_n	4369
Rejections by A_n	2976

BIBLIOGRAPHY

Aly, E. (1983). Some limit theorems for uniform and exponential spacings. *Canadian Journal of Statistics*, 11(3):211–219.

Aly, E. (1988). Strong approximations of quadratic sums of uniform spacings. *Canadian Journal of Statistics*, 16(2):201–207.

Beghin, L., & Orsingher, E. (1999). On the maximum of the generalized Brownian bridge. *Liet. Mat. Rink.*, 39(2):200–213.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons., New York, second edition.

Borodin, A. N., & Salminen, P. (2002). *Handbook of Brownian motion—facts and formulae*. Probability and its Applications. Birkhäuser Verlag, Basel, second edition.

Csörgő, M., & Révész, P. (1981). *Strong approximations in probability and statistics*, series *Probability and Mathematical Statistics*. Academic Press, Inc., New York-London.

Csörgő (1983). *Quantile processes with statistical applications*, volume 42 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Einmahl, U. (1989). Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *J. Multivariate Anal.*, 28(1):20–68.

Moran, P. A. P. (1947). The random division of an interval. *Suppl. J. Roy. Statist. Soc., 9(1): 92–98.*

R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017

Revuz, D., & Yor, M. (1999). Continuous martingales and Brownian motion, volume 293 of Fundamental Principles of Mathematical Sciences. Springer-Verlag, Berlin, third edition.

Shorack, Galen R., & Wellner, Jon A. (1986). Empirical processes with applications to statistics, volume 59 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. 2009 reprint of the 1986 original.

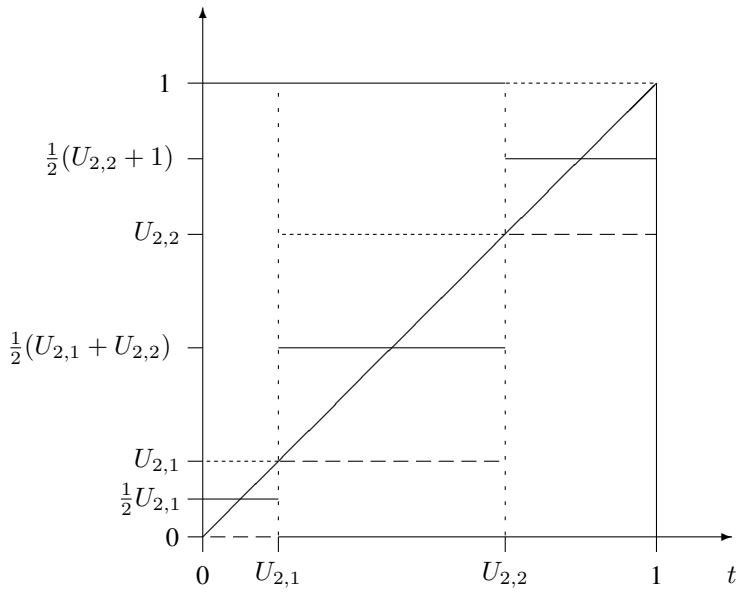

FIGURE 1: Trajectories of the EIP (solid), of the lower EIP (dashed) and of the upper EIP (dotted) with $n = 2$, $U_{2,1} = 1/6$ and $U_{2,2} = 4/6$.

Slud, Eric (1978). Entropy and maximal spacings for random partitions, Probability Theory and Related Fields, 41 (4): 341–352.

Zaitsev, A. Y. (1998). Multidimensional version of the results of Komlós, Major and Tusnády for vectors with finite exponential moments. ESAIM Probab. Statist., 2:41–108.

APPENDIX

This Appendix contains proofs of the theorems found in the body of the paper.

Proof of Theorem 1.

Theorem 1 states that the FDDs of the bivariate process $\left(Y_n^U(\cdot), -Y_n^L(\cdot)\right)$ converge weakly to the FDDs of two independent exponential white noise processes. Before proving the theorem, two useful lemmas are proven.

Consider k fixed distinct numbers $u_1 < u_2 < \ldots < u_k$ in the interval $(0, 1)$ and let $u_0 = 0$ and $u_{k+1} = 1$ be the extreme points. We will be working with the $k + 1$ bins induced by these points and with the order statistics from the uniform i.i.d. process $U_1, U_2, \ldots$ falling into the different bins. In particular, let

$$\begin{aligned} C_n = C_n(u_1, u_2, \ldots u_k) &= (C_{n,1}, C_{n,2}, \ldots, C_{n,k}, C_{n,k+1})' \\ &= n \cdot (F_n(u_1), F_n(u_2) - F_n(u_1), \ldots, F_n(u_k) - F_n(u_{k-1}), 1 - F_n(u_k))' \end{aligned}$$

be the sequence of the vectors of counts of i.i.d. uniform observations $U_1, U_2, \ldots, U_n$ falling into the different bins, $n = 1, 2, \ldots,$. In order to keep the notation simple, the dependence of C_n on $u_1, u_2, \ldots u_k$ will be understood. It is well known that the distribution of C_n is multinomial with parameters $n, u_1, u_2 - u_1, \ldots, u_k - u_{k-1}, 1 - u_k$; i.e. the probability mass function of the vector $(C_{n,1}, C_{n,2}, \ldots, C_{n,k})'$, evaluated at the vector of

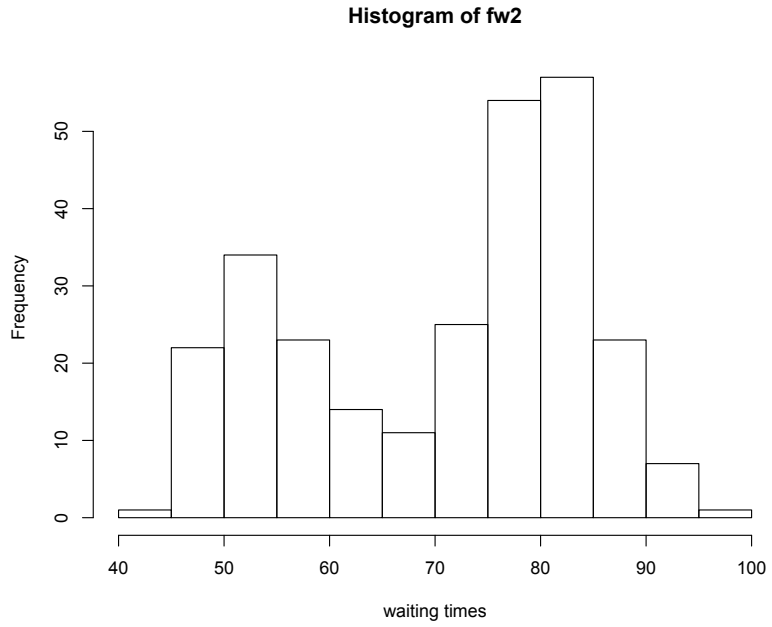

FIGURE 2: Estimated density of the Old Faithful dataset. Bimodality is clearly visible.

non-negative integers $(c_1, \dots, c_k)$, with $\sum_{j=1}^{k} c_j \leq n$, is

$$n! \prod_{j=1}^{k+1} \frac{(u_j - u_{j-1})^{c_j}}{c_j!} \tag{1}$$

where $c_{k+1} = n - \sum_{j=1}^{k} c_j$.

Appendix Lemma 1 *Let $U_1, \dots, U_n$, be a sequence of i.i.d. uniform random variables on $[0,1]$. Let $U_{n,1} \leq \dots \leq U_{n,n}$ be their order statistics. The constants $0 = u_0 < u_1 < u_2 < \dots < u_k < u_{k+1} = 1$ induce a partition of the interval $(0,1)$ into bins. The order statistics $U_{n,1}, \dots, U_{n,n}$ may be subdivided into groups belonging to the different bins. The counts of the order statistics falling into each of the bins are summarized into the multinomial vector C_n. The vectors of order statistics falling into each bin are conditionally independent given C_n Each of these vectors has a conditional distribution equal to the distribution of the order statistics of $C_{n,j}$ i.i.d. uniform observations on the interval (u_{j-1}, u_j).*

Proof.. The density of the order statistics $U_{n,1} \leq \dots \leq U_{n,n}$ evaluated at $0 < x_1, \dots, x_n < 1$ is $n! \cdot (0 < x_1 < x_2 \dots < x_n < 1)$. In order to write the joint density of the order statistics and the vector $C_n = (c_1, \dots, c_k, c_{k+1})'$, a compatibility factor $\prod_{j=1}^{k} (x_{c_1+\dots+c_j} < u_j < x_{c_1+\dots+c_{j+1}})$ has to be included. The conditional density of $U_{n,1} \leq U_{n,2} \leq \dots U_{n,n}$ given

$C_n = (c_1, \ldots, c_k, c_{k+1})'$ is, letting $c_0 = 0$,

$$\frac{n!(0 < x_1 < \ldots < x_n < 1) \prod_{j=1}^{k} (x_{c_1+\ldots+c_j} < u_j < x_{c_1+\ldots+c_{j+1}})}{n! \prod_{j=1}^{k+1} (u_j - u_{j-1})^{c_j}/c_j!}$$
$$= \prod_{j=1}^{k+1} \frac{c_j!}{(u_j - u_{j-1})^{c_j}} (u_{j-1} < x_{c_0+\ldots+c_{j-1}+1} < \ldots < x_{c_0+\ldots+c_j} < u_j)$$

which is seen to factor into the $k+1$ marginal densities of the vectors of adjacent order statistics falling into the $k+1$ bins. It is apparent that their distributions are as described in the statement of the theorem. ■

The following lemma is a corollary of the previous one:

Appendix Lemma 2 *The conditional density of the bidimensional vector* $(R_n^U(u_{j-1}), R_n^L(u_j))'$ *given* $C_n = (c_1, \ldots, c_k, c_{k+1})'$*, evaluated at* (x, y)*, is*

$$\frac{c_j(c_j - 1)(y - x)^{c_j - 2}}{(u_j - u_{j-1})^{c_j}} (u_j < x < y < u_{j-1}), \tag{2}$$

provided $c_j > 1$*, for each* $j = 1, \ldots, k+1$*.*

Proof.. Given C_n, by the previous lemma $R_n^U(u_{j-1})$ and $R_n^L(u_j)$ have the same distribution as the minimum and the maximum, respectively, of c_j uniform observations on the interval (u_{j-1}, u_j). It is easy to check that their density is given by formula (2). ■

Proof of Theorem 1. Consider the moment generating function

$$\phi_{Y_n^U(u_1),\ldots,Y_n^U(u_k),-Y_n^L(u_1),\ldots,-Y_n^L(u_k)}(v_1, \ldots, v_k, w_1, \ldots, w_k)$$

of the vector $(Y_n^U(u_1), \ldots, Y_n^U(u_k), -Y_n^L(u_1), \ldots, -Y_n^L(u_k))'$ evaluated at $(v_1, \ldots, v_k, w_1, \ldots, w_k)'$. Its asymptotic expression as $n \to \infty$ will be discussed. For each n an explicit form is found by conditioning on C_n:

$$\phi_{Y_n^U(u_1),\ldots,Y_n^U(u_k),-Y_n^L(u_1),\ldots,-Y_n^L(u_k)}(v_1, \ldots, v_k, w_1, \ldots, w_k) =$$
$$= \mathbb{E}\left[\exp\left(\sum_{j=1}^{k} \left[v_j Y_n^U(u_j) - w_j Y_n^L(u_j)\right]\right)\right]$$
$$= \mathbb{E}\left\{\mathbb{E}\left[\exp\left(\sum_{j=1}^{k} \left[v_j Y_n^U(u_j) - w_j Y_n^L(u_j)\right]\right) \,\middle|\, C_n\right]\right\}$$
$$= \mathbb{E}\left\{\prod_{j=1}^{k+1} \mathbb{E}\left[\exp\left(v_{j-1} Y_n^U(u_{j-1}) - w_j Y_n^L(u_j)\right) | C_n\right]\right\}$$

where $v_0 = w_{k+1} = 0$ for the sake of obtaining a compact notation. To derive the last line, the conditional independence property of Appendix Lemma 1 has been used and random variables relative to the different bins have been grouped.

Now we apply Appendix Lemma 2 to obtain an explicit form for the random variable $\mathbb{E}\left[\exp(v_{j-1}Y_n^U(u_{j-1}) - w_j Y_n^L(u_j))|C_n\right]$.

By the strong law of large numbers, for every ε there exist a set A_ε such that $\mathbb{P}(A_\varepsilon) = 1$ and for all $\omega \in A_\varepsilon$ there exist an $N(\omega, \varepsilon)$ such that for all $n > N(\omega, \varepsilon)$

$$\min_{j=1\cdots k+1} \frac{C_{n,j}}{n}(\omega) > u_j - u_{j-1} - \varepsilon.$$

On any such set for all $n > \max\{N(\omega), \frac{1}{u_j - u_{j-1} - \varepsilon}\}$ we have

$$\min_{j=1\cdots k+1} C_{n,j}(\omega) > 1.$$

Without loss of generality, we can work on the set A_ε as long as only asymptotic results are of interest. Then, on this set A_ε, for any large enough n, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp(v_{j-1}Y_n^U(u_{j-1}) - w_j Y_n^L(u_j))|C_n\right] \\
&\qquad = \int_{u_{j-1}}^{u_j}\int_{x}^{u_j} \exp(v_{j-1}(n+1)(x-u_{j-1}) - w_j(n+1)(y-u_j)) \\
&\qquad\qquad \frac{C_{n,j}(C_{n,j}-1)(y-x)^{C_{n,j}-2}}{(u_j-u_{j-1})^{C_{n,j}}} dydx \\
&\qquad = \int_0^1\int_0^{1-s} \exp(v_{j-1}(n+1)(u_j-u_{j-1})s + w_j(n+1)(u_j-u_{j-1})t) \\
&\qquad\qquad C_{n,j}(C_{n,j}-1)(1-s-t)^{C_{n,j}-2} dtds \\
&\qquad\qquad = 1 + I_{j1} + I_{j2} + I_{j3}
\end{aligned}$$

after the change of variable $s = (x - u_{j-1})/(u_j - u_{j-1})$ and $t = (u_j - y)/(u_j - u_{j-1})$ and a few integrations by parts, where

$$I_{j1} = v_{j-1}(n+1)(u_j - u_{j-1}) \int_0^1 \exp(v_{j-1}(n+1)(u_j-u_{j-1})s)(1-s)^{C_{n,j}} ds$$

$$I_{j2} = w_j(n+1)(u_j - u_{j-1}) \int_0^1 \exp(w_j(n+1)(u_j-u_{j-1})t)(1-t)^{C_{n,j}} dt$$

$$I_{j3} = v_{j-1}\, w_j\, (n+1)^2 (u_j - u_{j-1})^2$$
$$\int_0^1\int_0^{1-t} \exp\left(v_{j-1}(n+1)(u_j-u_{j-1})s + w_j(n+1)(u_j-u_{j-1})t\right)(1-s-t)^{C_{n,j}} ds\, dt.$$

Now, by the law of large numbers we have $C_{n,j}/(n+1) \to u_j - u_{j-1}$ a.s., as $n \to \infty$, for each $j = 1, \ldots, k+1$. Thus on the set A_ε introduced before

$$\left(\left(1 - \frac{s}{n+1}\right)^{n+1}\right)^{C_{n,j}/(n+1)} \leq \exp(-s)^{u_j - u_{j-1} - \varepsilon}. \tag{3}$$

After the changes of variable $s = \frac{z}{n}$ and $t = \frac{r}{n}$, we can apply the dominated convergence theorem and it follows that, as $n \to \infty$ and for $|v_j| < 1, j = 0, 1, \ldots, k+1$:

$$\begin{aligned}
I_{j1} &= v_{j-1}(u_j - u_{j-1}) \int_0^n \exp[v_{j-1}(u_j - u_{j-1})z] \left[\left(1 - \frac{z}{n+1}\right)^{n+1} \right]^{\frac{C_{n,j}}{n+1}} dz \\
&\to v_{j-1}(u_j - u_{j-1}) \int_0^\infty \exp[(v_{j-1} - 1)(u_j - u_{j-1})z] dz \\
&= \frac{v_{j-1}}{1 - v_{j-1}} \quad \text{almost surely (a.s.)} \\
I_{j2} &\to \frac{w_j}{1 - w_j} \quad \text{a.s.} \\
I_{j3} &= v_{j-1}\, w_j\, n^2 (u_j - u_{j-1})^2 \\
&\int_0^n \int_0^{n(1-r)} \exp[v_{j-1}(u_j - u_{j-1})z + w_j(u_j - u_{j-1})r] \left[\left(1 - \frac{z+r}{n+1}\right)^{n+1} \right]^{\frac{C_{n,j}}{n+1}} dz\, dr \\
&\to \frac{v_{j-1}}{1 - v_{j-1}} \frac{w_j}{1 - w_j} \quad \text{a.s.}
\end{aligned}$$

Now, after some algebra we have that as $n \to \infty$:

$$\prod_{j=1}^{k+1} (1 + I_{j1} + I_{j2} + I_{j3}) \to \prod_{j=1}^{k+1} \frac{1}{1 - v_{j-1}} \frac{1}{1 - w_j} = \prod_{j=1}^{k} \frac{1}{1 - v_j} \frac{1}{1 - w_j} \qquad a.s.$$

since we had set $v_0 = w_{k+1} = 0$. Moreover, since (3) also ensures uniform integrability, we have that

$$\begin{aligned}
&\phi_{Y_n^U(u_1),\ldots,Y_n^U(u_k),-Y_n^L(u_1),\ldots,-Y_n^L(u_k)}(v_1, \ldots, v_k, w_1, \ldots, w_k) = \\
&\qquad = \mathbb{E}\left(\prod_{j=1}^{k+1} (1 + I_{j1} + I_{j2} + I_{j3}) \right) \to \prod_{j=1}^{k} \frac{1}{1 - v_j} \frac{1}{1 - w_j}
\end{aligned}$$

which concludes the proof of Theorem 1. ■

Proof of Theorems 4 and 5.

Theorem 4 can be seen as a consequence of Theorem 3, by a suitable application of the continuous mapping theorem. The detailed proof is below.

Proof of Theorem 4. For every t, there are $nF_n(t)$ observations in our sample of size n which are smaller then t and the value of the last one is exactly equal to

$$R_n^L(t) = U_{n,n\mathbb{F}_n(t)} = \mathbb{Q}_n(\mathbb{F}_n(t)).$$

By Equation (3), we have $I_n(t) = \frac{n+1}{2} \sum_{i=1}^{nF_n(t)} D_{n,i}^2 + \frac{n+1}{2}(t - R_n^L(t))^2$. Then

$$\frac{n+1}{2} \sum_{i=1}^{nF_n(t)} D_{n,i}^2 \le I_n(t) \le \frac{n+1}{2} \sum_{i=1}^{(n+1)F_n(t)} D_{n,i}^2.$$

Since $\lfloor (n+1)F_n(t) \rfloor = nF_n(t)$ for all $0 \leq t < U_{n,n}$ and $\lfloor (n+1)F_n(t) \rfloor = nF_n(t)+1$ for all $U_{n,n} \leq t \leq 1$, we have that the difference between $2\sqrt{n+1}\ (\ I_n(t) - \mathbb{F}_n(t)\ )$ and $E_n(\mathbb{F}_n(t))$ is uniformly bounded by $(n+1)(\max_i D_{n,i})/2$. By Slud (1978) we have that $\max_i D_{n,i} = O\left(\frac{\log n}{n}\right)$ almost surely as $n \to \infty$, and

$$2\sqrt{n}\ (\ I_n^L(t) - \mathbb{F}_n(t)\ ) = E_{n+1}(\mathbb{F}_n(t)) + O\left[\left(\frac{(\log n)^2}{\sqrt{n}}\right)\right]$$

with probability one. Accordingly $E_n(\mathbb{F}_n(t))$ and $2\sqrt{n+1}\ (\ I_n(t) - \mathbb{F}_n(t)\ )$ have the same limit distribution. Now, $\mathbb{F}_n(\cdot)$ converges to the identity function $\mathrm{id}(\cdot)$ (which of course is deterministic) and $E_n(\cdot)$ converges weakly to $V(\cdot)$. As a result, the pair $(E_n(\cdot), \mathbb{F}_n(\cdot))$ converges weakly in $D^2(0,1)$ to $(V(\cdot), \mathrm{id}(\cdot))$, and by the continuity of the composition map, applying the continuous mapping theorem (see Billingsley (1999), page 151), we can conclude that

$$2\sqrt{n+1}\ (\ I_n(t) - \mathbb{F}_n(t)\ ) \Rightarrow V(t), \ \ t \in [0,1].$$

■

Proof of Theorem 5. We can derive the two-sided maximal probability distribution for $V(t)$ from that of the pinned Brownian motion (cf. Beghin and Orsingher (1999), Equation 4.12 or Borodin and Salminen (2002), Part II, Chapter 1, Equation 1.15.8(1) page 174) as follows

$$\mathbb{P}\left(\sup_{0\leq t\leq 1}|V(t)| < b\right) = \mathbb{E}\left[\mathbb{P}\left(\sup_{0\leq t\leq 1}\left|B(t) + t\frac{W(1)}{\sqrt{5}}\right| < \frac{b}{2\sqrt{5}}\right)\Bigg| W(1)\right].$$

Based on Equation 4.12 in Beghin and Orsingher (1999) or Part II, Chapter 1, Equation 1.15.8(1) page 174, in Borodin and Salminen (2002), we know that

$$\mathbb{P}\left(\sup_{0\leq t\leq 1}|B(t)+ty| < a\right) = \sum_{h=-\infty}^{\infty}(-1)^h \mathrm{e}^{-2ha(ha-y)}$$

then

$$\begin{aligned}\mathbb{P}\left(\sup_{0\leq t\leq 1}|V(t)| < b\right) =& \mathbb{E}\left[\sum_{h=-\infty}^{\infty}(-1)^h \mathrm{e}^{-2h\frac{b}{2\sqrt{5}}\left(h\frac{b}{2\sqrt{5}} - \frac{W(1)}{\sqrt{5}}\right)}\right] = \\ &= \sum_{h=-\infty}^{\infty}(-1)^h \mathrm{e}^{-\frac{h^2b^2}{10}}\mathbb{E}\left(\mathrm{e}^{\frac{hb}{5}W(1)}\right) \\ &= \sum_{h=-\infty}^{\infty}(-1)^h \mathrm{e}^{-\frac{4}{50}h^2b^2}.\end{aligned}$$

■

Received 31 December 2017
Accepted 31 December 2017