Weighted Kolmogorov-Smirnov test: Accounting for the tails

Rémy Chicheportiche[1,2] and Jean-Philippe Bouchaud[1]
[1]*Capital Fund Management, 75 007 Paris, France*
[2]*Chaire de finance quantitative, Ecole Centrale Paris, 92 295 Châtenay-Malabry, France*
(Received 1 August 2012; published 10 October 2012)

Accurate goodness-of-fit tests for the extreme tails of empirical distributions is a very important issue, relevant in many contexts, including geophysics, insurance, and finance. We have derived exact asymptotic results for a generalization of the large-sample Kolmogorov-Smirnov test, well suited to testing these extreme tails. In passing, we have rederived and made more precise the approximate limit solutions found originally in unrelated fields, first in [L. Turban, J. Phys. A **25**, 127 (1992)] and later in [P. L. Krapivsky and S. Redner, Am. J. Phys. **64**, 546 (1996)].

DOI: 10.1103/PhysRevE.86.041115 PACS number(s): 02.50.Ey, 03.65.Ge, 05.10.Gg

I. INTRODUCTION AND MOTIVATION

The problem of testing whether a null-hypothesis theoretical probability distribution is compatible with the empirical probability distribution of a sample of observations is known as goodness-of-fit (GoF) testing and is ubiquitous in all fields of science and engineering. The best known theoretical result is due to Kolmogorov and Smirnov (KS) [1,2], and has led to the eponymous statistical test. Several specific cases have been studied (and/or are still under scrutiny), including univariate or multivariate samples [3–6], independent or dependent data [7], different choices of distance measures [8], investigation of different parts of the distribution domain [9,10], etc.

This class of problems has a particular appeal for physicists since the works of Doob [11] and Khmaladze [12], who showed how GoF testing is related to stochastic processes. Finding the law of a test often amounts to treating a Fokker-Planck problem, which in turn maps into a Schrödinger equation for a particle in a certain potential confined by walls.

The classical KS test suffers from an important flaw: the test is only weakly sensitive to the quality of the fit in the tails of the tested distribution, when it is often these tail events (corresponding to centennial floods, devastating earthquakes, financial crashes, etc.) that one is most concerned with (see, e.g., Ref. [13]). Here we focus on a GoF test for a univariate sample of size $N \gg 1$, with the Kolmogorov distance but equiweighted quantiles, which is equally sensitive to all regions of the distribution. We unify two earlier attempts at finding asymptotic solutions, one by Anderson and Darling in 1952 [9] and a more recent, seemingly unrelated one that deals with "life and death of a particle in an expanding cage" by Krapivsky and Redner [14,15]. We present here the exact asymptotic solution of the corresponding stochastic problem, and deduce from it the precise formulation of the GoF test, which is of a fundamentally different nature than the KS test.

II. EMPIRICAL CUMULATIVE DISTRIBUTION AND ITS FLUCTUATIONS

Let $\mathbf{X}$ be a latent random vector of N independent and identically distributed variables, with marginal cumulative distribution function (cdf) F. One realization of $\mathbf{X}$ consists of a time series $\{x_1, \ldots, x_n, \ldots, x_N\}$ that exhibits no persistence (see Ref. [7] when some nontrivial dependence is present). For a given number x in the support of F, let $\mathbf{Y}(x)$ be the random vector the components of which are the Bernoulli variables $\mathrm{Y}_n(x) = \mathbb{1}_{\{\mathrm{X}_n \leqslant x\}}$. The expected value and the covariance of $\mathrm{Y}_n(x)$ are given by

$$\mathbb{E}[\mathrm{Y}_n(x)] = F(x),$$
$$\mathbb{E}[\mathrm{Y}_n(x)\mathrm{Y}_m(x')] = \begin{cases} F(\min(x,x')), & n = m \\ F(x)F(x') \quad , & n \neq m \end{cases}.$$

The centered sample mean of $\mathbf{Y}(x)$ is

$$\overline{\mathrm{Y}}(x) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Y}_n(x) - F(x), \tag{1}$$

which measures the difference between the empirically determined cdf at point x and its true value. It is therefore the quantity on which any statistics for GoF testing is built. Denoting $u = F(x)$ and $v = F(x')$, the covariance function of $\overline{\mathrm{Y}}$ is easily shown to be

$$\mathrm{cov}(\overline{\mathrm{Y}}(u), \overline{\mathrm{Y}}(v)) = \frac{1}{N}[\min(u,v) - uv],$$

where now and in the following

$$\overline{\mathrm{Y}}(u) = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Y}_n(F^{-1}(u)) - u. \tag{1'}$$

A. Limit properties

One now defines the process $y(u)$ as the limit of $\sqrt{N}\,\overline{\mathrm{Y}}(u)$ when $N \to \infty$. For a given u, it represents the difference between the empirically determined cdf of the (infinitely many) X's and the theoretical one, evaluated at the uth quantile. According to the central limit theorem, it is Gaussian and its covariance function is given by

$$I(u,v) = \min(u,v) - uv, \tag{2}$$

which characterizes the so-called Brownian bridge, i.e., a Brownian motion $y(u)$ such that $y(u=0) = y(u=1) = 0$.

Interestingly, F does not appear in Eq. (2) anymore, so the law of any functional of the limit process y is independent of the law of the underlying finite size sample. This property is important for the design of *universal* GoF tests.

B. Norms over processes and the Kolmogorov-Smirnov test

To measure a limit distance between distributions, a norm $||.||$ over the space of continuous bridges needs to be chosen. Typical such norms are the norm-2 (or Cramer–von Mises distance)

$$||y||_2 = \int_0^1 y(u)^2 du,$$

as the bridge is always integrable, or the norm-sup

$$||y||_\infty = \sup_{u\in[0,1]} |y(u)|,$$

as the bridge always reaches an extremal value (also called the Kolmogorov distance). Unfortunately, both these norms mechanically overweight the core values $u \approx 1/2$ and disfavor the tails $u \approx 0,1$: since the variance of $y(u)$ is zero at both extremes and maximal in the central value, the major contribution to $||y||$ indeed comes from the central region. To alleviate this effect, particularly when the GoF test is intended to investigate a specific region of the domain, it is preferable to introduce additional weights and study $||y\sqrt{\psi}||$ rather than $||y||$ itself. Anderson and Darling show in Ref. [9] that the solution to the problem with the Cramer–von Mises norm and arbitrary weights ψ is obtained by spectral decomposition of the covariance kernel, and use of Mercer's theorem. In this note we will rather focus on the case of the weights ψ being equal to the inverse variance $\psi(u) = 1/\mathbb{V}[y(u)]$, which equiweights all quantiles, and with the Kolmogorov distance.

Solutions for the distributions of such variance-weighted Kolmogorov-Smirnov statistics were studied by Noé, leading to the laws of the one-sided [16] and two-sided [17] finite sample tests. They were later generalized and tabulated numerically by Niederhausen [18–20]. However, although exact and appropriate for small samples, these solutions rely on recursive relations and are not in closed form. We instead come up with an analytic closed-form solution for large samples that relies on an elegant analogy from statistical physics.

III. THE WEIGHTED BROWNIAN BRIDGE: LAW OF THE SUPREMUM

So again $y(u)$ is a Brownian bridge, i.e. a centered Gaussian process on $u \in [0,1]$ with covariance function $I(u,v)$ given in Eq. (2). In particular, $y(0) = y(1) = 0$ with probability equal to 1, no matter how distant F is from the sample cdf around the core values. To zoom in on these tiny differences in the tails, we weight the Brownian bridge as follows: for given $a \in]0,1[$ and $b \in [a,1[$, we define

$$\tilde{y}(u) \equiv y(u)\sqrt{\psi(u;a,b)}, \tag{3}$$

with

$$\psi(u;a,b) = \begin{cases} \frac{1}{u(1-u)}, & a \leqslant u \leqslant b \\ 0, & \text{otherwise.} \end{cases}$$

We will characterize the law of the supremum $K(a,b) \equiv \sup_{u\in[a,b]} |\tilde{y}(u)|$:

$$\begin{aligned}\mathcal{P}_<(k|a,b) &\equiv \mathbb{P}[K(a,b) \leqslant k] \\ &= \mathbb{P}[|\tilde{y}(u)| \leqslant k, \forall\ u \in [a,b]].\end{aligned}$$

A. Diffusion in a cage with moving walls

Define the time change $t = \frac{u}{1-u}$. The variable $W(t) = (1+t)\, y(\frac{t}{1+t})$ is then a Brownian motion (Wiener process) on $[\frac{a}{1-a}, \frac{b}{1-b}]$, since one can check that

$$\mathrm{cov}(W(t), W(t')) = \min(t,t').$$

$\mathcal{P}_<(k|a,b)$ can be now written as

$$\mathcal{P}_<(k|a,b) = \mathbb{P}\left[|W(t)| \leqslant k\sqrt{t}, \quad \forall\, t \in \left[\frac{a}{1-a}, \frac{b}{1-b}\right]\right].$$

The problem with initial time $\frac{a}{1-a} = 0$ and horizon time $\frac{b}{1-b} = T$ has been treated by Krapivsky and Redner in Ref. [14] as the survival probability $S(T;k = \sqrt{\frac{A}{2D}})$ of a Brownian particle diffusing with constant D in a cage with walls expanding as $\sqrt{At}$. Their result is that for large T,

$$S(T;k) \equiv \mathcal{P}_<\left(k \middle| 0, \frac{T}{1+T}\right) \propto T^{-\theta(k)}.$$

They obtain analytical expressions for $\theta(k)$ in both limits $k \to 0$ and $k \to \infty$. The limit solutions of the very same differential problem were found earlier by Turban for the critical behavior of the directed self-avoiding walk in parabolic geometries [21].

We take here a slightly different route suggested by Anderson and Darling in Ref. [9] but where the authors did not come to a conclusion. Our contributions are (i) we treat the general case $a > 0$ for *any* k; (ii) we explicitly compute the k dependence of both the exponent *and* the prefactor of the power-law decay; and (iii) we provide the link with the theory of GoF tests and compute the preasymptotic distribution when $]a,b[\to]0,1[$ of the weighted Kolmogorov-Smirnov test statistics.

Choosing a constant weight function ψ instead of the one above corresponds to the usual KS case and leads, after appropriate change of variable and time change, to a similar problem of a Brownian diffusion inside a box with walls moving at *constant* velocity. Since the walls now expand as Vt faster than the diffusive particle can move, the survival probability clearly decays to a positive value. The resulting survival probability turns out to be the usual Kolmogorov-Smirnov distribution. Other choices of ψ generally result in much harder problems.

Still, a simple and elegant GoF test for the tails *only* can be designed starting with digital weights in the form $\psi(u;a) = \mathbb{1}_{\{u\geqslant a\}}$ or $\psi(u;b) = \mathbb{1}_{\{u\leqslant b\}}$ for upper and lower tail, respectively. The corresponding test laws can be read off Eq. (5.9) in Ref. [9].[1] Investigation of both tails is attained with $\psi(u;q) = \mathbb{1}_{\{u\leqslant 1-q\}} + \mathbb{1}_{\{u\geqslant q\}}$ (where $q > \frac{1}{2}$).

B. An Ornstein-Uhlenbeck process with fixed walls

Introducing now the new time change $\tau = \ln\sqrt{\frac{1-a}{a}}\, t$, the variable $Z(\tau) = W(t)/\sqrt{t}$ is a stationary Ornstein-Uhlenbeck

[1]The quantity M appearing there is the volume under the normal bivariate surface between specific bounds, and it takes a very convenient form in the unilateral cases $\frac{1}{2} \leqslant a \leqslant u \leqslant 1$ and $0 \leqslant u \leqslant b \leqslant \frac{1}{2}$. Mind the missing j exponentiating the alternating (-1) factor.

process on $[0,T]$ where

$$T = \ln\sqrt{\frac{b(1-a)}{a(1-b)}}, \tag{4}$$

and

$$\mathrm{cov}(Z(\tau),Z(\tau')) = e^{-|\tau-\tau'|}.$$

Its dynamics is described by the stochastic differential equation

$$dZ(T) = -Z(T)dT + \sqrt{2}\,dB(T), \tag{5}$$

with $B(T)$ an independent Wiener process. The initial condition for $T=0$ (corresponding to $b=a$) is $Z(0) = y(a)/\sqrt{\mathbb{V}[y(a)]}$, a random Gaussian variable of zero mean and unit variance. The distribution $\mathcal{P}_<(k|a,b)$ can now be understood as the unconditional survival probability of a mean-reverting particle in a cage with fixed absorbing walls:

$$\begin{aligned}\mathcal{P}_<(k|T) &= \mathbb{P}[-k \leqslant Z(\tau) \leqslant k, \forall\, \tau \in [0,T]] \\ &= \int_{-k}^{k} f_T(z;k)dz,\end{aligned}$$

where

$$f_T(z;k)dz = \mathbb{P}[Z(T) \in [z, z+dz[\,|\,\{Z(\tau)\}_{\tau<T}]$$

is the density probability of the particle being at z at time T, when walls are in $\pm k$. Its dependence on k, although not explicit on the right-hand side, is due to the boundary condition associated with the absorbing walls (it will be dropped in the following for the sake of readability).[2]

The Fokker-Planck equation governing the evolution of the density $f_T(z)$ reads

$$\partial_\tau f_\tau(z) = \partial_z\left[z\, f_\tau(z)\right] + \partial_z^2\left[f_\tau(z)\right], \quad 0 < \tau \leqslant T.$$

Calling $\mathcal{H}_{\mathrm{FP}}$ the second-order differential operator $-[\mathbb{1} + z\partial_z + \partial_z^2]$, the full problem thus amounts to finding the general solution of

$$\begin{cases} -\partial_\tau f_\tau(z) = \mathcal{H}_{\mathrm{FP}}(z) f_\tau(z), \\ f_\tau(\pm k) = 0, \forall\, \tau \in [0,T]. \end{cases}$$

We have explicitly introduced a minus sign since we expect that the density decays with time in an absorption problem. Because of the term $z\partial_z$, $\mathcal{H}_{\mathrm{FP}}$ is not Hermitian and thus cannot be diagonalized. However, as is well known, one can define $f_\tau(z) = e^{-\frac{z^2}{4}}\phi_\tau(z)$ and the Fokker-Planck equation becomes

$$\begin{cases} -\partial_\tau \phi_\tau(z) = \left[-\partial_z^2 + \frac{1}{4}z^2 - \frac{1}{2}\mathbb{1}\right]\phi_\tau(z), \\ \phi_\tau(\pm k) = 0, \quad \forall\, \tau \in [0,T], \end{cases}$$

and its Green's function, i.e., the (separable) solution *conditionally on the initial position* (z_i,T_i), is the superposition of all modes

$$G_\phi(z,T \mid z_i,T_i) = \sum_\nu e^{-\theta_\nu(T-T_i)}\widehat{\varphi}_\nu(z)\widehat{\varphi}_\nu(z_i),$$

where $\widehat{\varphi}_\nu$ are the normalized solutions of the stationary Schrödinger equation

$$\begin{cases} \left[-\partial_z^2 + \frac{1}{4}z^2\right]\varphi_\nu(z) = \left(\theta_\nu + \frac{1}{2}\right)\varphi_\nu(z), \\ \varphi_\nu(\pm k) = 0, \end{cases}$$

each decaying with its own energy θ_ν, where ν labels the different solutions with increasing eigenvalues, and the set of eigenfunctions $\{\widehat{\varphi}_\nu\}$ defines an orthonormal basis of the Hilbert space on which $\mathcal{H}_{\mathrm{S}}(z) = [-\partial_z^2 + \frac{1}{4}z^2]$ acts. In particular,

$$\sum_\nu \widehat{\varphi}_\nu(z)\widehat{\varphi}_\nu(z') = \delta(z-z'), \tag{6}$$

so that indeed $G(z,T_i \mid z_i,T_i) = \delta(z-z_i)$, and the general solution writes

$$f_T(z_T;k) = \int_{-k}^{k} e^{\frac{z_i^2 - z_T^2}{4}} G_\phi(z_T,T \mid z_i,T_i)\, f_0(z_i)dz_i,$$

where $T_i = 0$, which corresponds to the case $b=a$ in Eq. (3), and f_0 is the distribution of the initial value z_i which is here, as noted above, Gaussian with unit variance.

$\mathcal{H}_{\mathrm{S}}$ figures out an harmonic oscillator of mass $\frac{1}{2}$ and frequency $\omega = \frac{1}{\sqrt{2}}$ within an infinitely deep well of width $2k$: its eigenfunctions are parabolic cylinder functions [22,23]

$$\begin{aligned} y_+(\theta;z) &= e^{-\frac{z^2}{4}}\,{}_1F_1\left(-\frac{\theta}{2},\frac{1}{2},\frac{z^2}{2}\right), \\ y_-(\theta;z) &= z\,e^{-\frac{z^2}{4}}\,{}_1F_1\left(\frac{1-\theta}{2},\frac{3}{2},\frac{z^2}{2}\right), \end{aligned}$$

properly normalized. The only acceptable solutions for a given problem are the linear combinations of y_+ and y_- which satisfy orthonormality (6) and the boundary conditions: for periodic boundary conditions, only the integer values of θ would be allowed, whereas with our Dirichlet boundaries $|\widehat{\varphi}_\nu(k)| = -|\widehat{\varphi}_\nu(-k)| = 0$, real noninteger eigenvalues θ are allowed.[3] For instance, the fundamental level $\nu = 0$ is expected to be the symmetric solution $\widehat{\varphi}_0(z) \propto y_+(\theta_0;z)$ with θ_0 the smallest possible value compatible with the boundary condition:

$$\theta_0(k) = \inf_{\theta>0}\{\theta : y_+(\theta;k) = 0\}. \tag{7}$$

In what follows, it will be more convenient to make the k dependence explicit, and a hat will denote the solution with the normalization relevant to our problem, namely, $\widehat{\varphi}_0(z;k) = y_+(\theta_0(k);z)/||y_+||_k$, with the norm

$$||y_+||_k^2 \equiv \int_{-k}^{k} y_+(\theta_0(k);z)^2 dz,$$

so that $\int_{-k}^{k}\widehat{\varphi}_\nu(z;k)^2dz = 1$.

[2] In particular, $\mathcal{P}_<(k|0) = \mathrm{erf}(\frac{k}{\sqrt{2}})$.

[3] A similar problem with a *one-sided* barrier leads to a continuous spectrum; this case was studied originally in Ref. [22] and more recently in Ref. [24] (it was shown that there exists a quasistationary distribution for any θ) and generalized in Ref. [25].

C. Asymptotic survival rate

Denoting by $\Delta_\nu(k) \equiv [\theta_\nu(k) - \theta_0(k)]$ the gap between the excited levels and the fundamental, the higher energy modes $\widehat{\varphi}_\nu$ cease to contribute to Green's function when $\Delta_\nu T \gg 1$, and their contributions to the above sum die out exponentially as T grows. Eventually, only the lowest energy mode $\theta_0(k)$ remains, and the solution tends to

$$f_T(z;k) = A(k)\,\mathrm{e}^{-\frac{z^2}{4}}\,\widehat{\varphi}_0(z;k)\,e^{-\theta_0(k)T},$$

when $T \gg (\Delta_1)^{-1}$, with

$$A(k) = \int_{-k}^{k} \mathrm{e}^{\frac{z_i^2}{4}}\,\widehat{\varphi}_0(z_i;k)\,f_0(z_i)dz_i. \tag{8}$$

Let us come back to the initial problem of the weighted Brownian bridge reaching its extremal value in $[a,b]$. If we are interested in the limit case where a is arbitrarily close to 0 and b close to 1, then $T \to \infty$ and the solution is thus given by

$$\begin{aligned}\mathcal{P}_<(k|T) &= A(k)\,\mathrm{e}^{-\theta_0(k)T}\int_{-k}^{k} e^{-\frac{z^2}{4}}\widehat{\varphi}_0(z;k)dz \\ &= \widetilde{A}(k)\,\mathrm{e}^{-\theta_0(k)T},\end{aligned}$$

with $\widetilde{A}(k) \equiv \sqrt{2\pi}A(k)^2$.

We now compute explicitly the limit behavior of both $\theta_0(k)$ and $\widetilde{A}(k)$.

1. $k \to \infty$

As k goes to infinity, the absorption rate $\theta_0(k)$ is expected to converge toward 0: intuitively, an infinitely far barrier will not absorb anything. At the same time, $\mathcal{P}_<(k|T)$ must tend to 1 in that limit. So $\widetilde{A}(k)$ necessarily tends to unity. Indeed,

$$\begin{aligned}\theta_0(k) &\overset{k\to\infty}{\longrightarrow} \sqrt{\frac{2}{\pi}}k\,e^{-\frac{k^2}{2}} \to 0, \\ \widetilde{A}(k) &\overset{k\to\infty}{\longrightarrow} \left(\int_{-\infty}^{\infty}\widehat{\varphi}_0(z;\infty)^2dz\right)^2 = 1.\end{aligned} \tag{9}$$

In principle, we see from Eq. (8) that corrections to the latter arise both (and jointly) from the functional relative difference of the solution $\epsilon(z;k) = y_+(\theta_0(k);z)/y_+(0;z) - 1$, and from the finite integration limits ($\pm k$ instead of $\pm\infty$). However, it turns out that the correction of the first kind is of second order in ϵ.[4] The correction to $A(k)$ is thus dominated by the finite integration limits $\pm k$, so that

$$\widetilde{A}(k \to \infty) \approx \mathrm{erf}\left(\frac{k}{\sqrt{2}}\right)^2. \tag{10}$$

[4]From Eq. (8) we have, when $k \to \infty$,

$$A(k) = (2\pi)^{-1/2}\frac{\int_{-k}^{k} e^{-z^2/2}[1+\epsilon(z;k)]\,dz}{\sqrt{\int_{-k}^{k} e^{-z^2/2}[1+\epsilon(z;k)]^2\,dz}}.$$

The result follows by keeping only the dominant terms in the expansion in powers of $\epsilon(z;k)$. A similar computation for the asymptotic analysis by expanding the wave function in θ was performed in Ref. [14]. Alternatively, algebraic arguments allow us to understand that to first order in the energy correction $\theta_0(k) - \theta_0(\infty)$, the perturbation of the wave function is orthogonal to $\widehat{\varphi}_0(z;\infty)$.

2. $k \to 0$

For small k, the system behaves like a free particle in a sharp and infinitely deep well, since the quadratic potential is almost flat around 0. The fundamental mode becomes then

$$\widehat{\varphi}_0(z;k \to 0) = \frac{1}{\sqrt{k}}\cos\left(\frac{\pi z}{2k}\right),$$

and consequently

$$\theta_0(k) \overset{k\to\infty}{\longrightarrow} \frac{\pi^2}{4k^2} - \frac{1}{2}, \tag{11}$$

$$\begin{aligned}\widetilde{A}(k) &\overset{k\to\infty}{\longrightarrow} \left(\int_{-k}^{k}\frac{e^{-\frac{z^2}{4}}}{(2\pi)^{\frac{1}{4}}}\frac{1}{\sqrt{k}}\cos\left(\frac{\pi z}{2k}\right)dz\right)^2 \\ &\approx \frac{1}{\sqrt{2\pi k}}\left(\frac{4k}{\pi}\right)^2 = \frac{16}{\pi^2\sqrt{2\pi}}k.\end{aligned} \tag{12}$$

We show in Fig. 1 the functions $\theta_0(k)$ and $\widetilde{A}(k)$ computed numerically from the exact solution, together with their asymptotic analytic expressions. In intermediate values of k (roughly between 0.5 and 3) these limit expressions fail to reproduce the exact solution.

D. Higher modes and validity of the asymptotic ($N \gg 1$) solution

Higher modes $\nu > 0$ with energy gaps $\Delta_\nu \lesssim 1/T$ must in principle be kept in the preasymptotic computation. This, however, is irrelevant in practice since the gap $\theta_1 - \theta_0$ is never small. Indeed, $\widehat{\varphi}_1(z;k)$ is proportional to the asymmetric solution $y_-(\theta_1(k);z)$ and its energy

$$\theta_1(k) = \inf_{\theta > \theta_0(k)}\{\theta : y_-(\theta;k) = 0\}$$

is found numerically to be very close to $1 + 4\theta_0(k)$. In particular, $\Delta_1 > 1$ (as we illustrate in Fig. 2) and thus $T\Delta_1 \gg 1$ will always be satisfied in cases of interest.

IV. BACK TO GoF TESTING AND CONCLUSION

Let us now come back to GoF testing. In the case of a constant weight, corresponding to the classical KS test, the probability $\mathcal{P}_<(k|a=0,b=1)$ is well defined and has the well-known KS form [1]

$$\mathcal{P}_<(k|a=0,b=1) = 1 - 2\sum_{n=1}^{\infty}(-1)^{n-1}e^{-2n^2k^2},$$

which, as expected, grows from 0 to 1 as k increases. The value k^* such that this probability is 95% is $k^* \approx 1.358$ [2]. This can be interpreted as follows: if, for a data set of size N, the maximum value of $\overline{Y}(u)$ is larger than $\approx 1.358/\sqrt{N}$, then the hypothesis that the proposed distribution is a "good fit" can be rejected with 95% confidence.

To convert the above calculations into a meaningful test, one must specify values of a and b. The natural choice is $a = 1/N$ and $b = 1 - a$, corresponding to the min and max of the sample series. Indeed, $a = F(\min z) \approx \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{\{z_n \leqslant \min z\}} = \frac{1}{N}$, and similarly for b. Correspondingly,

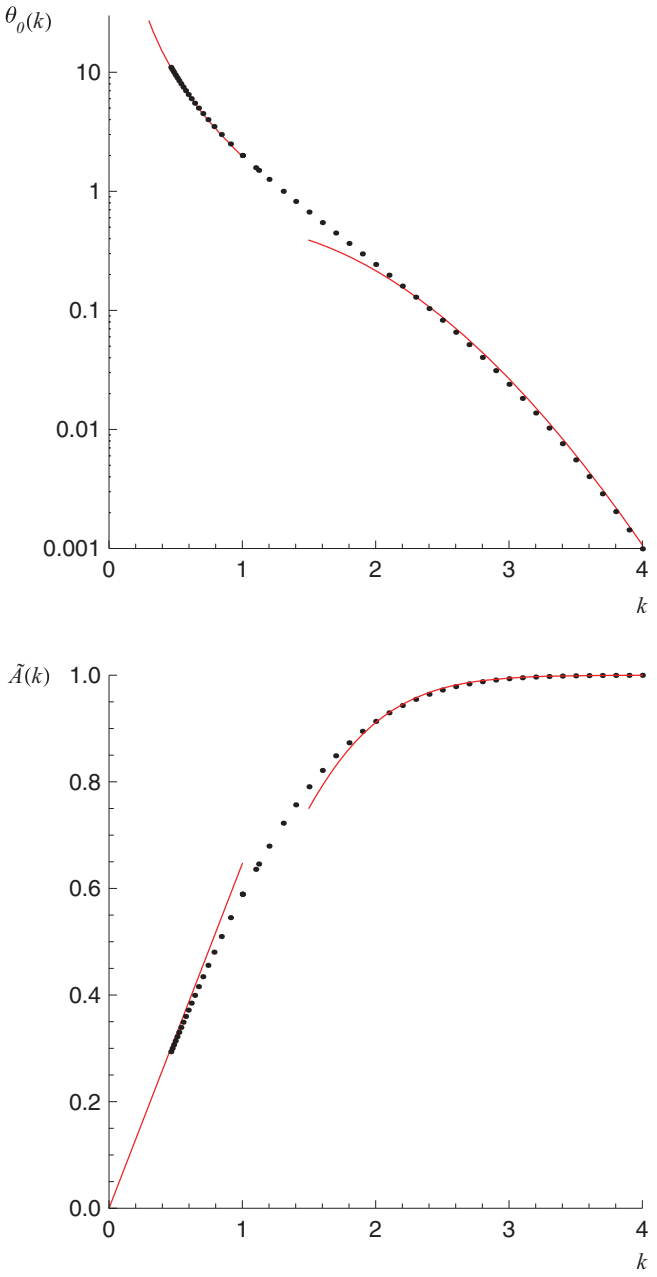

FIG. 1. (Color online) Top: Dependence of the exponent θ_0 on k; similar to Fig. 2 in Ref. [14], but in linear-log$_{10}$ scale; see in particular Eqs. (9b) and (12) there. Bottom: Dependence of the prefactor $\widetilde{A}$ on k. The red solid lines illustrate the analytical behavior in the limiting cases $k \to 0$ and $k \to \infty$.

the relevant value of T is given, according to Eq. (4) above, by

$$T = \ln \sqrt{\frac{b(1-a)}{a(1-b)}} \approx \ln N, \quad N \gg 1.$$

This leads to our central result for the cdf of the weighted maximal Kolmogorov distance $K(\frac{1}{N+1}, \frac{N}{N+1})$ under the hypothesis that the tested and the true distributions coincide:

$$S(N;k) = \mathcal{P}_<(k|\ln N) = \widetilde{A}(k) N^{-\theta_0(k)}, \qquad (13)$$

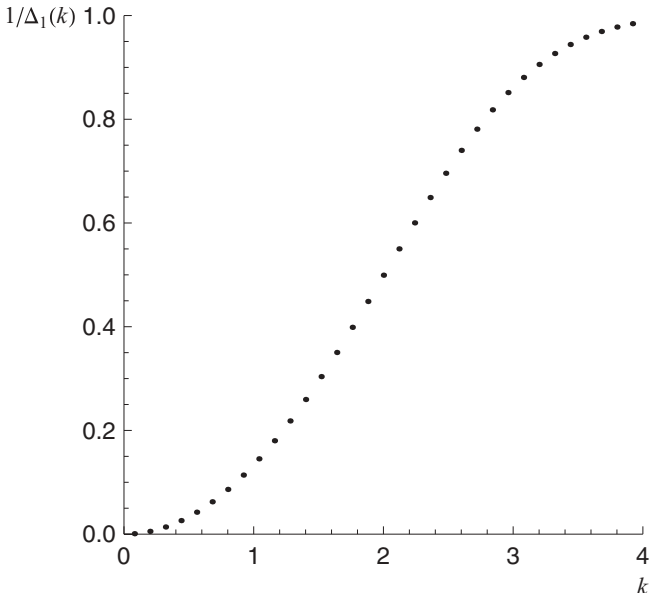

FIG. 2. $1/\Delta_1(k)$ saturates to 1, so that the condition $N \gg \exp[1/\Delta_1(k)]$ is virtually always satisfied.

which is valid whenever $N \gg 1$ since, as we discussed above, the energy gap Δ_1 is greater than unity.

The final cumulative distribution function (the test law) is depicted in Fig. 3 for different values of the sample size N. Contrary to the standard KS case, this distribution *still depends on* N. In particular, the threshold value k^* corresponding to a 95% confidence level increases with N. Since for large N, $k^* \gg 1$ one can use the asymptotic expansion above, which

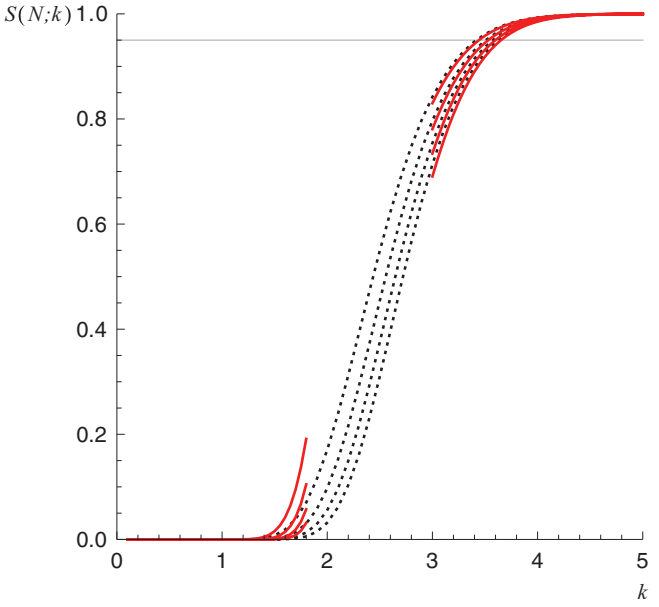

FIG. 3. (Color online) Dependence of $S(N;k)$ on k for $N = 10^3, 10^4, 10^5, 10^6$ (from left to right). As N grows toward infinity, the curve is shifted to the right, and eventually $S(\infty;k)$ is zero for any k. The red solid lines illustrate the analytical behavior in the limiting cases $k \to 0$ and $k \to \infty$. The horizontal grey line corresponds to a 95% confidence level.

soon becomes quite accurate, as shown in Fig. 3. This leads to

$$\theta_0(k^*) \approx -\frac{\ln 0.95}{\ln N} \approx \sqrt{\frac{2}{\pi}} k^* e^{-\frac{k^{*2}}{2}},$$

which gives $k^* \approx 3.439, 3.529, 3.597, 3.651$ for, respectively, $N = 10^3, 10^4, 10^5, 10^6$. For exponentially large N and to logarithmic accuracy, one has $k^* \sim \sqrt{2 \ln(\ln N)}$. This variation is very slow, but one sees that as a matter of principle, the "acceptable" maximal value of the weighted distance is much larger (for large N) than in the KS case.

In conclusion, we believe that accurate GoF tests for the extreme tails of empirical distributions is a very important issue, relevant in many contexts. We have derived exact asymptotic results for a generalization of the Kolmogorov-Smirnov test, well suited to testing the whole domain up to these extreme tails. Our final results are summarized in Eq. (13) and Fig. 3. In passing, we have rederived and made more precise the result of Krapivsky and Redner [14] concerning the survival probability of a diffusive particle in an expanding cage. It would be interesting to exhibit other choices of weight functions that lead to soluble survival probabilities. It would also be interesting to extend the present results to multivariate distributions and to dependent observations, along the lines of Ref. [7].

ACKNOWLEDGMENTS

We want to thank Sid Redner for a useful discussion and his inspiring work, and Loïc Turban for bringing Ref. [21] to our attention.

[1] A. N. Kolmogorov, Giornale dell'Istituto Italiano degli Attuari **4**, 83 (1933).
[2] N. Smirnov, Ann. Math. Statistics **19**, 279 (1948).
[3] G. Fasano and A. Franceschini, Mon. Not. R. Astron. Soc. **225**, 155 (1987).
[4] A. Cabaña and E. M. Cabaña, Ann. Stat. **22**, 1447 (1994).
[5] A. Cabaña and E. M. Cabaña, Ann. Stat. **25**, 2388 (1997).
[6] J.-D. Fermanian, J. Multivar. Anal. **95**, 119 (2005).
[7] R. Chicheportiche and J.-P. Bouchaud, J. Stat. Mech.: Theory Exp. (2011) P09003.
[8] D. A. Darling, Ann. Math. Stat. **28**, 823 (1957).
[9] T. W. Anderson and D. A. Darling, Ann. Math. Stat. **23**, 193 (1952).
[10] P. Deheuvels, Afrika Stat. **1**, 1:14 (2009).
[11] J. L. Doob, Ann. Math. Stat. **20**, 393 (1949).
[12] E. V. Khmaladze, Theory Probab. Its Appl. **26**, 240 (1982).
[13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[14] P. L. Krapivsky and S. Redner, Am. J. Phys. **64**, 546 (1996).
[15] S. Redner, *A Guide to First-Passage Processes* (Cambridge University, Cambridge, UK, 2001).
[16] M. Noé and G. Vandewiele, Ann. Math. Stat. **39**, 233 (1968).
[17] M. Noé, Ann. Math. Stat. **43**, 58 (1972).
[18] H. Niederhausen, Ann. Stat. **9**, 923 (1981).
[19] H. Niederhausen, Tech. Rep., Stanford University, Department of Statistics, 1981 (unpublished), http://statistics.stanford.edu/~ckirby/techreports/ONR/SOL%20ONR%20298.pdf.
[20] R. R. Wilcox, Comm. Stat. Simulat. Comput. **18**, 237 (1989).
[21] L. Turban, J. Phys. A: Math. Gen. **25**, L127 (1992).
[22] W. N. Mei and Y. C. Lee, J. Phys. A: Math. Gen. **16**, 1623 (1983).
[23] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products. Corrected and Enlarged Edition* (Academic, New York, 1980).
[24] M. Lladser and J. San Martín, J. Appl. Probab. **37**, 511 (2000).
[25] O. O. Aalen and H. K. Gjessing, Lifetime Data Anal. **10**, 407 (2004).