Journal of Statistical Software

October 2020, Volume 95, Issue 10. doi: 10.18637/jss.v095.i10

Computing the Kolmogorov-Smirnov Distribution When the Underlying CDF is Purely Discrete, Mixed, or Continuous

Dimitrina S. Dimitrova
City, University of London

Vladimir K. Kaishev
City, University of London

Senren Tan
City, University of London

Abstract

The distribution of the Kolmogorov-Smirnov (KS) test statistic has been widely studied under the assumption that the underlying theoretical cumulative distribution function (CDF), $F(x)$, is continuous. However, there are many real-life applications in which fitting discrete or mixed distributions is required. Nevertheless, due to inherent difficulties, the distribution of the KS statistic when $F(x)$ has jump discontinuities has been studied to a much lesser extent and no exact and efficient computational methods have been proposed in the literature.

In this paper, we provide a fast and accurate method to compute the (complementary) CDF of the KS statistic when $F(x)$ is discontinuous, and thus obtain exact p values of the KS test. Our approach is to express the complementary CDF through the rectangle probability for uniform order statistics, and to compute it using fast Fourier transform (FFT). Secondly, we provide a C++ and an R implementation of the proposed method, which fills the existing gap in statistical software. We give also a useful extension of the Schmid's asymptotic formula for the distribution of the KS statistic, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. The numerical performance of the proposed FFT-based method, implemented both in C++ and in the R package **KSgeneral**, available from https://CRAN.R-project.org/package=KSgeneral, is illustrated when $F(x)$ is mixed, purely discrete, and continuous. The performance of the general asymptotic formula is also studied.

Keywords: Kolmogorov-Smirnov test statistic, discontinuous (discrete or mixed) distribution, fast Fourier transform, double boundary non-crossing, rectangle probability for uniform order statistics.

1. Introduction

The two-sided Kolmogorov-Smirnov (KS) statistic is one of the most popular goodness-of-fit test statistics that is used to measure how well the distribution of a random sample $\{X_1, \ldots, X_n\}$ agrees with a theoretical distribution. It is defined as

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{1}$$

where n is the sample size, $F_n(x)$ denotes the empirical (cumulative) distribution function (EDF) of $\{X_1, \ldots, X_n\}$, and $F(x)$ denotes the cumulative distribution function (CDF) of a pre-specified theoretical distribution under the null hypothesis (H_0) that the sample $\{X_1, \ldots, X_n\}$ comes from $F(x)$.

Many authors have studied the distribution of D_n, i.e., its CDF $\mathsf{P}(D_n \le q|H_0), q \in [0,1]$ under the assumption that $F(x)$ is continuous. Kolmogorov (1933), Smirnov (1939), Feller (1948), Doob (1949), and Smirnov (1948) considered the limiting distribution of D_n. Massey (1951) showed that the exact distribution of D_n is independent of $F(x)$ if $F(x)$ is continuous, and provided a table for exact critical levels of the KS test corresponding to certain significance levels for sample sizes $n \le 35$. Durbin (1968) studied the probability that the EDF of an ordered sample of n independent observations from the uniform $(0,1)$ distribution lies between two parallel straight lines. He also obtained the exact distribution of D_n for $F(x)$ continuous, when the two parallel straight lines are $ny = a + nx$ and $ny = -a + nx$. Durbin (1968) also noted the important link between this probability and the double-boundary non-crossing probability for a Poisson process that is easier to compute. Epanechnikov (1968), Steck (1971), Noé (1972), Niederhausen (1981) obtained the exact distribution of D_n when $F(x)$ is continuous, by studying the probability that the order statistics of n uniform $[0,1]$ random variables all lie within an n-dimensional rectangle. For brevity, we will further refer to this probability as the rectangle probability for uniform order statistics. Numerically computing the distribution of D_n when $F(x)$ is continuous is not easy and has been recently considered by Marsaglia, Tsang, and Wang (2003), Simard and L'Ecuyer (2011), Carvalho (2015), among others. Details related to these works and further references are provided in Section 3.3.

While performing KS tests when $F(x)$ is continuous is widely applicable, there are many real-life applications, e.g., in biology, physics, engineering, finance, and insurance, in which fitting discrete or mixed distributions, i.e., with multiple jumps and continuous segments, to large samples of data is required. For example, Calabrese and Zenga (2010) modeled the bank loan recovery rates using mixed random variables, since empirical data suggest that loans are either not repaid at all (recovery rate = 0), partially repaid (recovery rate between 0 and 1), or fully repaid (recovery rate = 1). This leads to considering a mixed CDF $F(x)$ with jumps at 0 and 1 and a continuous segment in between. It is important to accurately model bank loan recovery rates, because this is required by the Basel II solvency framework. Mixed distributions with multiple jumps arise also in reinsurance, in relation to fitting claim amount data in multi-layer excess-of-loss treaties. We consider such an example in Section 3.1. Furthermore, numerous risk modeling applications in (general) insurance, e.g., car insurance and catastrophe insurance, require fitting appropriate discrete distributions to claim numbers data. The need to fit discrete distributions to data naturally arises also in almost any field of research in science and economics. In all such cases, the underlying CDF $F(x)$ has discontinuities at some points and it is important to be able to perform goodness-of-fit tests, such as the chi-squared test and the KS test. As demonstrated by Pettitt and

Stephens (1977), the KS test for discrete distributions can have greater power than the chi-squared test. On the other hand, Noether (1963), Slakter (1965), and Walsh (1963) showed that conducting a discontinuous KS test is more conservative than conducting a continuous KS test in terms of accepting/rejecting the null hypothesis. Thus, as we illustrate in Section 3.1, a null hypothesis that a sample comes from a discontinuous distribution will be accepted more often if one uses the continuous KS test, as opposed to using the discontinuous KS test. It should also be noted that the sample size in many applications can be substantial. Therefore, it is important to accurately and efficiently perform KS tests for $F(x)$ with discontinuities, when sample sizes are large. For this purpose, one needs to be able to efficiently and accurately compute probabilities of the type, $\mathsf{P}(D_n \geq q)$, known as the complementary CDF, for any values of n and q, $q \in [0, 1]$. Addressing this problem is the main objective of this paper.

The distribution of the KS test statistic D_n in this more general case, when $F(x)$ may have jump discontinuities (including purely discrete $F(x)$), has been studied to a much lesser extent. In an early paper, Schmid (1958) found the limiting distribution of D_n when $F(x)$ has countable number of jumps and is increasing between the jumps. Carnal (1962) has generalized Schmid (1958)'s formula by allowing constant segments between jumps. Conover (1972) provided an approach to finding the exact critical level for the one-sided KS test statistics $D_n^- = \sup_x(F(x) - F_n(x))$ and $D_n^+ = \sup_x(F_n(x) - F(x))$ for discontinuous $F(x)$. Approximated critical levels for the two-sided KS test statistic D_n were also provided. Gleser (1985) studied the exact power of two-sided KS tests. He showed that existing algorithms designed for KS tests with continuous $F(x)$ could be used (after some necessary adjustments) for KS tests when $F(x)$ is discontinuous. Specifically, Gleser (1985) showed that the power of the KS test when $F(x)$ has jump discontinuities could still be expressed as a rectangle probability with respect to uniform order statistics, but with modified non-linear boundaries. Therefore, the determinantal and recurrence formulae for the latter rectangle probability due to Steck (1971), Noé (1972) and Niederhausen (1981) could be applied in order to obtain the exact distribution of D_n when $F(x)$ is discontinuous. However, implementing these results is computationally expensive, especially when the sample size is large, and may lead to numerical instabilities, as noted by some authors and also illustrated in Section 3.2.

In summary, computing the distribution of D_n when $F(x)$ is discontinuous is even harder and much less explored than in the continuous case. To the best of our knowledge, no methods have been proposed in the literature to compute the exact distribution of D_n when $F(x)$ is mixed. Looking at the statistical software literature, all major packages implement the KS test only when $F(x)$ is continuous, see for example, the `ks.test` function of the package **stats** (R Core Team 2020) and the `ks.test.imp` function of the package **kolmim** (Carvalho 2015) in R (R Core Team 2020), SPSS (IBM Corp. 2017), the `ksmirnov` function in Stata (StataCorp. 2017), the `kstest` function in MATLAB (The MathWorks Inc. 2018), and the `KolmogorovSmirnovTest` function in Mathematica (Wolfram Research, Inc. 2018).

There is one exception, Arnold and Emerson (2011) provide the R function `ks.test` as part of the package **dgof** that calculates exact p values of the KS test assuming $F(x)$ is purely discrete. In their function `ks.test`, a one-sided KS p value is calculated by combining the approaches of Conover (1972) and Niederhausen (1981), while two-sided KS p values are calculated by combining the approaches of Gleser (1985) and Niederhausen (1981). However, the `ks.test` function due to Arnold and Emerson (2011) only provides exact p values for sample sizes less than or equal to 30, since as noted by the authors, when the sample size is large, numerical instabilities may occur. In the latter case, Arnold and Emerson (2011)

suggest using simulation to approximate p values, which as we show in Section 3.2, is rather slow and inaccurate.

Our aim in this paper is two-fold. The first goal is to provide a fast and accurate method to compute $\mathsf{P}(D_n \geq q)$ when $F(x)$ is discontinuous (i.e., mixed or purely discrete), and thus obtain exact p values of the KS test for any (small or large) sample size n, and any $q \in [0,1]$, possibly close to 1. Our second goal is to give the C++ code and an R package **KSgeneral** (Dimitrova, Kaishev, and Tan 2020), based on the C++ code that implements this fast and accurate method, which we believe fills the gap in the existing statistical software. As we will see, the proposed method is also applicable and highly competitive when $F(x)$ is continuous. The approach we take, described in Section 2.1, is to express $\mathsf{P}(D_n \geq q)$ as an appropriate rectangle probability for uniform order statistics, as noted by Gleser (1985), and to compute the latter probability using the fast Fourier transform (FFT) method. FFT has been recently utilized by Moscovich and Nadler (2017) to calculate this rectangle probability when $F(x)$ is continuous. Furthermore, in Section 2.2, we provide a useful extension (cf., (15) and (20)) of Schmid (1958)'s asymptotic formula, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. A similar formula has been obtained by Carnal (1962), but the embedded implicit index structure makes its numerical implementation prohibitive. In Section 3, we illustrate the C++ and the R implementation in the package **KSgeneral** of the proposed FFT-based method. In particular, in Section 3.1, we study its numerical properties based on some mixed (inflated) distributions and also illustrate the performance of the general asymptotic formula (15). We show in Section 3.2 that when $F(x)$ is purely discrete, our approach to computing $\mathsf{P}(D_n \geq q)$, based on FFT and the asymptotic formula (22), outperforms in terms of speed and accuracy the R function of Arnold and Emerson (2011), especially for large sample sizes. Finally, in Section 3.3, we consider the case of continuous $F(x)$ and compare with the state-of-the-art procedures of Simard and L'Ecuyer (2011) and Carvalho (2015).

2. Distribution of D_n when $F(x)$ is discontinuous

It is well known that the distribution of D_n does not depend on $F(x)$ when the latter CDF is continuous. To see this, note that

$$\begin{aligned} D_n &= \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = \sup_{0\leq t\leq 1} \left|F_n(F^{-1}(t)) - F(F^{-1}(t))\right|, \\ &= \sup_{0\leq t\leq 1} \left|F_n(F^{-1}(t)) - t\right| = \sup_{0\leq t\leq 1} |U_n(t) - t|, \end{aligned} \tag{2}$$

where $F^{-1}(t) \equiv \inf\{x : F(x) \geq t\}, t \in [0,1]$, and $U_n(t)$ is the empirical CDF of the uniform random sample $\{U_i = F(X_i), i = 1, \ldots, n\}$. In this section, we relax the assumption of continuity of $F(x)$ and assume that $F(x)$ is non-decreasing and right-continuous, with countable (possibly infinite) number of jumps. From the right-continuity of $F(x)$, it follows that $F(F^{-1}(t)) \geq t$ and $F^{-1}(F(x)) \leq x$ and hence, the distribution-free property, illustrated by (2) is no longer valid. Therefore, it becomes difficult to compute the exact and asymptotic distributions of D_n. This problem is addressed in the next two sections.

2.1. The exact distribution of D_n

Our approach to computing the exact distribution of D_n is based on the following four major steps:

Step 1. It is not difficult to show (see Appendix A) that the complementary CDF $\mathsf{P}(D_n \geq q), q \in [0,1]$, can be expressed in terms of a rectangle probability for the vector of n uniform order statistics as

$$\mathsf{P}(D_n \geq q) = 1 - \mathsf{P}\left(A_i \leq U_{(i)} \leq B_i,\ 1 \leq i \leq n\right), \tag{3}$$

where

$$\begin{aligned} A_i &= \lim_{\epsilon \downarrow 0} F\Big(\Big(F^{-1}\big(\frac{i}{n} - q + \epsilon\big)\Big) - \Big), \\ F(x-) &= \lim_{z \uparrow x} F(z) = \mathsf{P}(X < x), \\ B_i &= \lim_{\epsilon \downarrow 0} F\Big(F^{-1}\big(\frac{i-1}{n} + q - \epsilon\big)\Big),\ i = 1, 2, \ldots, n, \end{aligned} \tag{4}$$

and where $U_{(i)}$, $i = 1, \ldots, n$, are the order statistics of n independent and identically distributed uniform $(0,1)$ random variables $U_i, i = 1, 2, \ldots, n$.

Step 2. Express the rectangle probability on the right hand side of (3) in terms of the double-boundary non-crossing probability with respect to the empirical process $\eta_n(t) = nU_n(t) = \sum_{i=1}^n \mathbb{1}(U_i \leq t), 0 \leq t \leq 1$, where $U_n(t)$ is the EDF of the sample $\{U_1, \ldots, U_n\}$. In particular, it can be directly verified that (3) can be rewritten as

$$\begin{aligned} \mathsf{P}(D_n \geq q) &= 1 - \mathsf{P}\left(A_i \leq U_{(i)} \leq B_i,\ 1 \leq i \leq n\right), \\ &= 1 - \mathsf{P}\left(g(t) \leq \eta_n(t) \leq h(t), \forall\ 0 \leq t \leq 1\right), \end{aligned} \tag{5}$$

where the upper and lower boundary functions $h(t)$, $g(t)$ are defined as

$$h(t) = \sum_{i=1}^{n} \mathbb{1}_{(A_i < t)}, \quad g(t) = \sum_{i=1}^{n} \mathbb{1}_{(B_i \leq t)}. \tag{6}$$

Let us note that $h(t)$ and $g(t)$ are correspondingly left and right continuous functions which equivalently satisfy the following conditions

$$\sup\{t \in [0,1] : h(t) < i\} = A_i, \text{ and } \inf\{t \in [0,1] : g(t) > i - 1\} = B_i, \tag{7}$$

with A_i, B_i defined in (4)[1]. The last equality in (5) is illustrated in Figure 1, where one can see that considering the rectangle probability with respect to the uniform order statistics, $\mathsf{P}\left(A_i \leq U_{(i)} \leq B_i,\ 1 \leq i \leq n\right)$ is equivalent to considering the non-exit probability, $\mathsf{P}\left(g(t) \leq \eta_n(t) \leq h(t), \forall\ 0 \leq t \leq 1\right)$.

Step 3. Use the fact that the process $\eta_n(t)$, $t \in [0,1]$, has the same distribution as the conditional distribution of a Poisson process with intensity n, denoted by $\xi_n(t) : [0,1] \mapsto \{0, 1, 2 \ldots\}$, given $\xi_n(1) = n$, (see, e.g., Shorack and Wellner 2009, Chapter 8, Proposition 2.2). Therefore, the non-crossing probability in (5) can be re-expressed as

$$\begin{aligned} &\mathsf{P}(g(t) \leq \eta_n(t) \leq h(t), \forall\ 0 \leq t \leq 1) \\ &\quad = \mathsf{P}(g(t) \leq \xi_n(t) \leq h(t) | \xi_n(1) = n, \forall\ 0 \leq t \leq 1) \\ &\quad = \frac{\mathsf{P}(g(t) \leq \xi_n(t) \leq h(t) \text{ and } \xi_n(1) = n, \forall\ 0 \leq t \leq 1)}{\mathsf{P}(\xi_n(1) = n)} = \frac{Q(1,n)}{e^{-n} n^n / n!}, \end{aligned} \tag{8}$$

[1] An expression similar to (5) for the case of $\mathsf{P}(D_n > q)$ has been obtained by Gleser (1985) (cf., Theorem 2 therein).

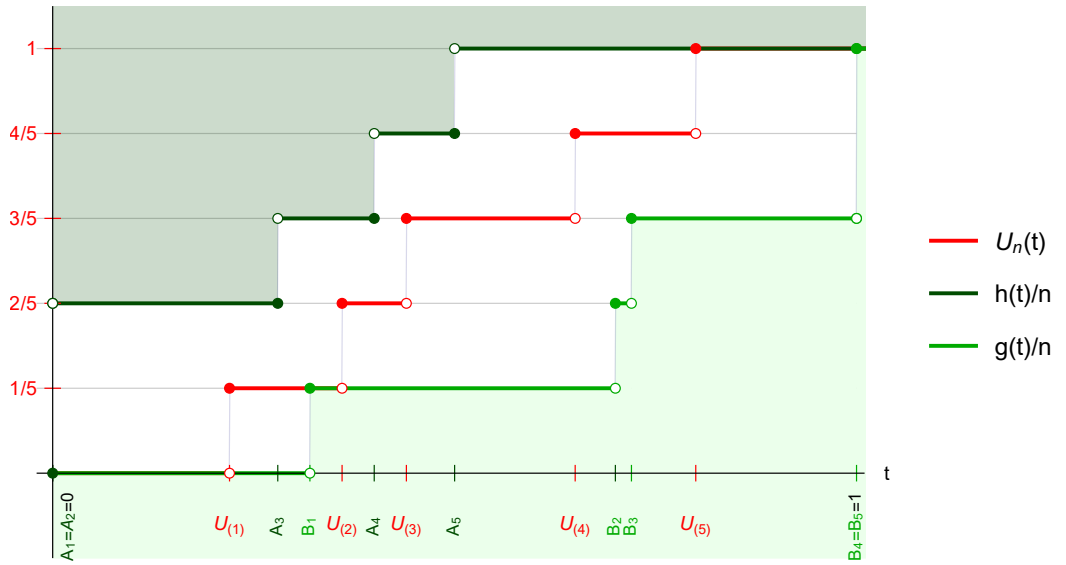

Figure 1: Illustration of the fact that the non-exit probability, $\mathsf{P}\left(A_i \le U_{(i)} \le B_i,\ 1 \le i \le n\right)$ is equivalent to the non-exit probability, $\mathsf{P}\left(g(t) \le \eta_n(t) \le h(t), \forall\ 0 \le t \le 1\right)$, where $g(t)$ and $h(t)$ are defined as in (6) using $F(x)$ given in (23) (cf., Example 2.8), with $n = 5$.

where $\xi_n(1)$ follows a Poisson(n) distribution and $Q(1, n)$ is defined as in (9). It is not difficult to see that in order to compute the non-crossing probability $\mathsf{P}(g(t) \le \xi_n(t) \le h(t)$ and $\xi_n(1) = n, \forall\ 0 \le t \le 1)$ on the right-hand-side of (8), defined on a continuum of times $t \in [0, 1]$, it suffices to consider the events of non-crossing only over some fixed times, $0 = t_0 < t_1 < t_2 < \ldots < t_N = 1$, which are the ordered set of all distinct points in $\{1, A_i, B_i, i = 1, \ldots, n\}$, where A_i and B_i are specified in (4) (and (7)). Based on this discretization, similarly as done by Khmaladze and Shinjikashvili (2001) and Moscovich and Nadler (2017) in the continuous case, the non-crossing probability in (8) can be calculated by solving recursively an appropriate system of Chapman-Kolmogorov forward equations[2]. In order to introduce these equations, for any $s \in [0, 1]$ and $m \in \{0, 1, 2, \ldots\}$, let

$$Q(s, m) = \mathsf{P}(g(t) \le \xi_n(t) \le h(t),\ \forall t \in [0, s] \text{ and } \xi_n(s) = m), \tag{9}$$

where $g(s) \le m \le h(s)$ and $Q(0, 0) = \mathsf{P}(g(0) \le 0 \le h(0)) = 1$ by assumption. For any $j \in \{0, 1, \ldots, N-1\}$ and any $m \in \{0, 1, 2, \ldots\}$, the Chapman-Kolmogorov equations are

$$Q(t_{j+1}, m) = \begin{cases} \sum_{g(t_j) \le l \le m} Q(t_j, l)\mathsf{P}(Y_j = m - l), & \text{if } g(t_{j+1}) \le m \le h(t_{j+1}), \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

where Y_j denotes a Poisson random variable with parameter $n(t_{j+1} - t_j)$. The required non-crossing probability is obtained by computing $Q(1, n)$ following (10). This is illustrated by Figure 2, where $g(t)$ and $h(t)$ are obtained based on (6), with $F(x)$ defined

[2]Both Khmaladze and Shinjikashvili (2001) and Moscovich and Nadler (2017) assume $F(x)$ is continuous and consider strict inequalities in (8), i.e., they do not allow the process to touch the boundaries.

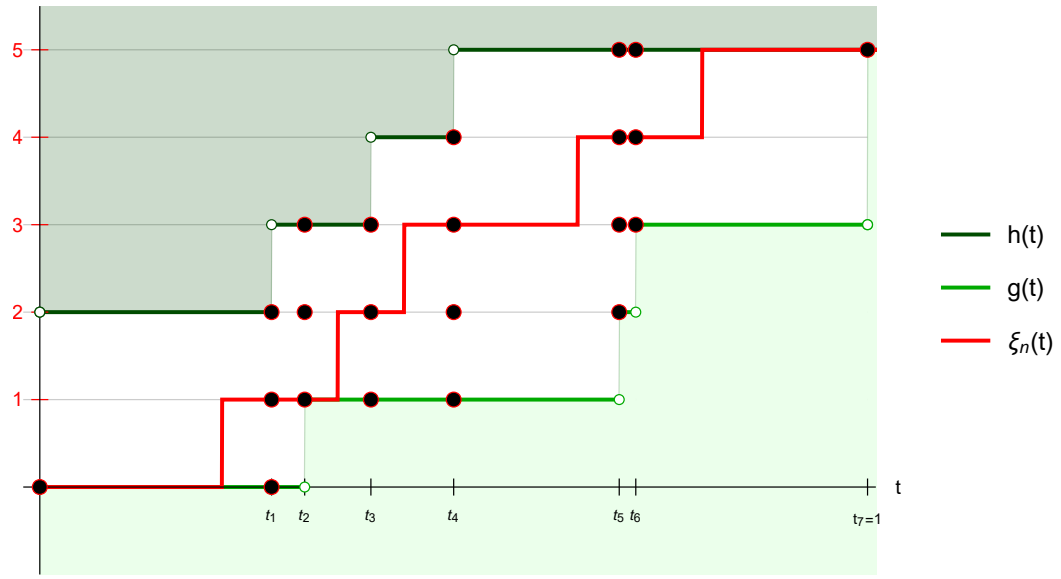

Figure 2: Illustration of a trajectory of the Poisson process $\xi_n(t)$ staying in the corridor between the boundaries $h(t)$ and $g(t)$ defined as in (6) using $F(x)$ given in (23) (cf., Example 2.8). The black dots illustrate the mesh of points (t_{j+1}, m), $j = 0, 1, \ldots, 6$, $m = 0, 1, 2, \ldots, 5$, at which non-crossing of the trajectory of $\xi_n(t)$ with the boundaries $g(t)$, $h(t)$ may occur and the corresponding probabilities, $Q(t_{j+1}, m)$ need to be computed, following (10).

in (23) as part of Example 2.8. The black dots illustrate the mesh of points (t_{j+1}, m), $j = 0, 1, \ldots, 6$, $m = 0, 1, 2, \ldots, 5$, at which non-crossing of the trajectory of $\xi_n(t)$ with the boundaries $g(t)$, $h(t)$ may occur and the corresponding probabilities, $Q(t_{j+1}, m)$ need to be computed, following (10).

As shown by Khmaladze and Shinjikashvili (2001), the recurrent computation following (10) requires total running time of order at most $\mathcal{O}(n^3)$. In the next step we employ FFT in order to improve this rate.

Step 4. Apply FFT to compute the truncated linear convolution of the vectors $\mathbf{Q}_{t_j} = (Q(t_j, 0), Q(t_j, 1), \ldots, Q(t_j, n))$ and $\boldsymbol{\pi}_{n(t_{j+1}-t_j)} = (\mathsf{P}(Y_j = 0), \mathsf{P}(Y_j = 1), \ldots, \mathsf{P}(Y_j = n))$ in order to solve (10), as proposed by Moscovich and Nadler (2017), see Section 2 therein. As shown by these authors, the total running time of this method is of order at most $\mathcal{O}(n^2 \log n)$, which is faster than $\mathcal{O}(n^3)$ especially for large n.

In summary, our approach to computing the exact $\mathsf{P}(D_n \geq q)$ when $F(x)$ is discontinuous is outlined in the following procedure (procedure exact-KS-FFT).

(i) Specify a discontinuous CDF $F(x)$, a sample size n, and a quantile q.

(ii) As detailed in Step 1, compute A_i and B_i for $i = 1, \ldots, n$, based on (4), where the limits are coded using a very small ϵ, e.g., $\epsilon = 10^{-10}$.

(iii) As detailed in Step 2, compute the upper and lower boundaries $g(t)$, $h(t)$ using (6).

(iv) Following Steps 3 and 4, apply FFT to compute $Q(1,n)$ defined in (10). Hence, calculate the double-boundary non-crossing probability with respect to the Poisson process on the right-hand-side of (8) and respectively obtain the double-boundary non-crossing probability with respect to $\eta_n(t)$ on the left-hand-side of (8).

(v) Finally, compute the exact $\mathsf{P}(D_n \geq q)$ using (5) (cf., Steps 2 and 3).

Remark 2.1. Let us note that $\mathsf{P}(D_n \geq q)$, $0 \leq q \leq 1$, can directly be computed using (3) and (4), applying the determinantal formula for the rectangle probability in (3), due to Steck (1971), or the recurrence formula of Niederhausen (1981). However, such computations are slow, and may become unstable for sample sizes $n \geq 100$, as shown in Section 3.2, Example 3.5. We also note that $\mathsf{P}(D_n \geq q)$ is the p value corresponding to a fixed critical level $q \in [0,1]$. Thus, if $q = d_n$, where d_n is the value of the KS test statistic computed based on a sample $\{x_1, \ldots, x_n\}$, then the corresponding exact p value, $\mathsf{P}(D_n \geq d_n)$ can be obtained through (3) and (4).

Remark 2.2. We have described the procedure exact-KS-FFT for computing the complementary CDF of the two-sided KS statistic, D_n, defined in (1). It should be noted that by selecting the lower boundary $g(t) \equiv 0, \forall t$, and the upper boundary $h(t)$ as specified in (6) one can compute the complementary CDF for the one-sided KS statistic $D_n^+ = \sup_x(F_n(x) - F(x))$. By selecting the upper boundary $h(t) \equiv n, \forall t$, and the lower boundary $g(t)$ as specified in (6), one can compute the complementary CDF for the one-sided KS statistic $D_n^- = \sup_x(F(x) - F_n(x))$ (see, e.g., Gleser 1985). For the sake of consistency, in what follows, we illustrate the proposed FFT-based method for the two-sided version of the KS statistic.

As noted and also demonstrated in Section 3, the proposed FFT-based method for computing exact $\mathsf{P}(D_n \geq q)$ is highly numerically efficient and could be easily applied to sample sizes n up to hundreds of thousands (see also Moscovich and Nadler 2017). Nevertheless, it is still beneficial to know the asymptotic distribution of D_n as $n \to \infty$, since as demonstrated in Section 3, it can be efficiently applied to approximate $\mathsf{P}(D_n \geq q)$ for large and even moderate sample sizes and hypothesized distributions with small number of jumps. The asymptotic distribution of D_n will be considered in the next section.

2.2. The asymptotic distribution of D_n

Schmid (1958) has studied the asymptotic distribution of the form

$$\Phi(\lambda) = \lim_{n\to\infty} \mathsf{P}(D_n < \lambda n^{-\frac{1}{2}}) = \lim_{n\to\infty} \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| < \lambda n^{-\frac{1}{2}}), \tag{11}$$

where n denotes the sample size, and $F(x)$ is a CDF with countable number of jumps J, at $x = x_l$, $l = 1, 2, \ldots, J$ and increasing continuous segments between the jumps. Let $F(x_l-) = f_{2l-1}, F(x_l) = f_{2l}$, $l = 1, 2, \ldots, J$, with $f_0 = 0, f_{2J+1} \equiv 1$, and $f_{2l} < f_{2l+1}$, $l = 0, \ldots, J$. Under these assumptions on $F(x)$, Theorem 1 of Schmid (1958) states that

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_{J+1}=-\infty}^{\infty} (-1)^{j_1+\cdots+j_{J+1}} \times c \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda}$$
$$\exp\Big[-\frac{1}{2}\sum_{l=1}^{J} \frac{(z_{2l} - z_{2l-1})^2}{f_{2l} - f_{2l-1}} - \frac{1}{2}\sum_{l=0}^{J} \frac{(z_{2l+1} - (-1)^{j_{l+1}} z_{2l} - 2\lambda j_{l+1})^2}{f_{2l+1} - f_{2l}} \Big] dz_1 \cdots dz_{2J}, \tag{12}$$

where

$$z_0 = z_{2J+1} = 0, \text{and } c = (2\pi)^{-J} \prod_{j=1}^{2J+1} (f_j - f_{j-1})^{-1/2}.$$

In view of (12), when the sample size n is large, the limiting $\mathsf{P}(D_n \geq q)$ for mixed $F(x)$ can be calculated as

$$\lim_{n\to\infty} \mathsf{P}(D_n \geq q) = 1 - \Phi(\lambda), \tag{13}$$

where $\Phi(\lambda)$ is expressed as in (12), and $\lambda = qn^{\frac{1}{2}}$. However, Schmid's formula cannot be applied if the condition $f_{2l} < f_{2l+1}, l = 0, \ldots, J$ is not satisfied, since there will be division by 0 in the second denominator in (12). Therefore, (12) is not applicable if $F(x)$ has constant segments between (some of) the jumps, as is the case when $F(x)$ is purely discrete, or if $F(x)$ starts (ends) with a jump at 0 (at 1), as is the case for zero-inflated (mixed) distributions. Carnal (1962) has generalized Schmid (1958)'s formula to the case of arbitrary discontinuous $F(x)$ with finite number of jumps (cf., expression (5.1) therein). However, there is notational ambiguity (e.g., in the fourth summation in (5.1)) and because the embedded index structure is rather implicit, it is not straightforward to implement formula (5.1) numerically. Therefore, in what follows, we will derive an alternative formula for $\Phi(\lambda)$, for any discontinuous $F(x)$ with finite number of jumps (see Proposition 2.3). The latter formula may look cumbersome, but as we will see, it is notationally explicit and therefore easier to implement numerically. In addition, we believe that the clearer and more intuitive proof of Proposition 2.3 will facilitate better understanding of the structure underlying (15). However, one should note that formula (15) (respectively (20) and (22)) is only practically implementable for small/moderate number of jumps, J, in the null distribution, as otherwise the multidimensional integration becomes infeasible.

It is not difficult to see that any jump structure in $F(x)$ can be represented through only two different types of continuous segments of $F(x)$ followed by jumps. The first one is a segment of $F(x)$ increasing on $[x_{l-1}, x_l-]$, i.e., $f_{2l-2} < f_{2l-1}$, followed by a jump at x_l, and the second one is a constant segment of $F(x)$ on $[x_{l-1}, x_l-]$, i.e., $f_{2l-2} = f_{2l-1}$, followed by a jump at x_l. We will refer to these two types of segments as increasing-jump segment and flat-jump segment, respectively.

We will use the notation $\nu_1, \nu_2, \ldots$ to denote the sizes of groups of consecutive increasing-jump segments, i.e., ν_i denotes the number of consecutive jumps, preceded by an increasing segment, in the ith group. Similarly, by ω_k, $k = 1, 2, \ldots$, we denote the number of consecutive jumps preceded by a flat segment, in the kth group. Without loss of generality, we assume that there are m groups of increasing-jump and flat-jump segments, i.e., $\nu_1, \ldots, \nu_m$ and $\omega_1, \ldots, \omega_m$, and that these groups of jumps points, x_l, appear in the CDF in the following order:

$$\begin{aligned}\{&x_1, \ldots, x_{\nu_1}, x_{\nu_1+1}, \ldots, x_{\nu_1+\omega_1}, x_{\nu_1+\omega_1+1}, \ldots, x_{\nu_1+\omega_1+\nu_2}, x_{\nu_1+\omega_1+\nu_2+1}, \ldots,\\ &x_{\nu_1+\omega_1+\nu_2+\omega_2}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+1}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m},\\ &x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m+1}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m+\omega_m}\},\end{aligned} \tag{14}$$

where $\nu_1 + \omega_1 + \cdots + \nu_m + \omega_m = J$ is the total number of jumps in $F(x)$, and

$$\nu_1 \geq 0; \omega_1 \geq 0; \nu_1 + \omega_1 > 0; \nu_l > 0, 2 \leq l \leq m; \omega_l > 0, 2 \leq l \leq m-1; \omega_m \geq 0; \nu_m + \omega_m > 0.$$

It can be seen that (14) covers any possible order of the jumps of different type in $F(x)$ as illustrated on some examples below (see, e.g., Corollary 2.6 and Example 2.8). Under these

general assumptions on $F(x)$, in the following proposition we give a formula for $\Phi(\lambda)$ which generalizes (12).

Proposition 2.3. *Assuming that a CDF $F(x)$ has the structure of jumps as in (14) and that $f_{2J} = f_{2J+1} \equiv 1$, we have*

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_{v_m}=-\infty}^{\infty} \left((-1)^{j_1+\cdots+j_{v_m}}\right) c \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\{\psi\} dz_1 \cdots dz_{2v_m+w_m-1}, \quad (15)$$

where

$$\begin{aligned} c = \prod_{i=1}^{m} \Big(\prod_{l=1}^{\nu_i} & \left(f_{2(v_{i-1}+w_{i-1}+l)-1} - f_{2(v_{i-1}+w_{i-1}+l)-2}\right)^{-1/2} \\ & \left(f_{2(v_{i-1}+w_{i-1}+l)} - f_{2(v_{i-1}+w_{i-1}+l)-1}\right)^{-1/2} \Big) \\ & \times \Big(\prod_{l=1}^{\omega_i} \left(f_{2(v_i+w_{i-1}+l)} - f_{2(v_i+w_{i-1}+l)-1}\right)^{-1/2} \Big) (2\pi)^{-\frac{2v_m+w_m-1}{2}}, \end{aligned} \quad (16)$$

and

$$\begin{aligned} \psi = -\frac{1}{2} \sum_{i=1}^{m} \Big\{ \sum_{l=1}^{\nu_i} \Big[& \frac{\left(z_{2(v_{i-1}+l)+w_{i-1}} - z_{2(v_{i-1}+l)+w_{i-1}-1}\right)^2}{f_{2(v_{i-1}+w_{i-1}+l)} - f_{2(v_{i-1}+w_{i-1}+l)-1}} \\ & + \frac{\left(z_{2(v_{i-1}+l)+w_{i-1}-1} - (-1)^{j_{(v_{i-1}+l)}} z_{2(v_{i-1}+l)+w_{i-1}-2} - 2\lambda j_{(v_{i-1}+l)}\right)^2}{f_{2(v_{i-1}+w_{i-1}+l)-1} - f_{2(v_{i-1}+w_{i-1}+l)-2}} \Big] \\ & + \sum_{l=1}^{\omega_i} \Big[\frac{\left(z_{2v_i+w_{i-1}+l} - z_{2v_i+w_{i-1}+l-1}\right)^2}{f_{2(v_i+w_{i-1}+l)} - f_{2(v_i+w_{i-1}+l)-1}} \Big] \Big\}, \end{aligned} \quad (17)$$

with $\nu_0 = \omega_0 = 0; v_0 = w_0 = 0; v_i = \sum_{k=1}^{i} \nu_k; w_i = \sum_{k=1}^{i} \omega_k, v_m + w_m = J$, and $z_0 = z_{2v_m+w_m} = 0$.

Proof: The reasoning in the proof follows that of Schmid (1958) with some necessary adjustments to account for the fact that $f_{2l} \le f_{2l+1}$ as opposed to $f_{2l} < f_{2l+1}$, $l = 0, \ldots, J$. So, here we only give details related to those parts of the proof which are affected by the relaxed assumption on $F(x)$. Thus, following Schmid (1958), page 1014, denote by I the union of the closed intervals $[f_{2l}, f_{2l+1}]$, $l = 0, \ldots, J$ and let M_n be the set of integers j such that $j/n \in I$,

$$M_n = \{k_0 = 0, \ldots, k_1; k_2, k_2 + 1, \ldots, k_3; \ldots; k_{2J}, k_{2J} + 1, \ldots, k_{2J+1} = n\},$$

where k_i is such that $k_i/n \to f_i$, as $n \to \infty$. Note that if $f_{2l} = f_{2l+1}$ i.e., if we have a constant segment in the CDF $F(x)$, then $k_{2l} \equiv k_{2l+1}$ and both are included in the set M_n. Now, as demonstrated by Schmid (1958) (see expressions (20), (21) therein), the probability $\mathsf{P}_{0n} := \mathsf{P}(D_n < \lambda n^{-1/2})$ in (11), can be calculated as

$$\mathsf{P}_{0n} = \frac{n! e^n}{n^n} R_{0n},$$

where

$$R_{ik_{2l+1}} = \sum_{|j| < \lambda N^{1/2}} R_{jk_{2l}} \mathsf{P}[\mathcal{D}_{ik_{2l+1}} | \mathcal{D}_{ik_{2l}}], \quad l = 0, \ldots, J, \quad (18)$$

and

$$R_{ik_{2l}} = \sum_{|j|<\lambda N^{1/2}} R_{jk_{2l-1}} \frac{(k_{2l}-k_{2l-1})^{i-j+k_{2l}-k_{2l-1}}}{(i-j+k_{2l}-k_{2l-1})!e^{k_{2l}-k_{2l-1}}}, \quad l=0,\ldots,J, \tag{19}$$

and $R_{00}=1, R_{i0}=0$ for $i\neq 0$. Note that recursion (19) is related to the lth jump in $F(x)$, whereas recursion (18) is related to the continuous (increasing or flat) segment on $[x_l, x_{l+1}-]$ in $F(x)$. The events $\mathcal{D}_{ik}$ are specified in details in Schmid (1958) (see page 1016), but what is important here is to observe that when $k_{2l}=k_{2l+1}$ in M_n, $\mathsf{P}[\mathcal{D}_{ik_{2l+1}}|\mathcal{D}_{ik_{2l}}]=\mathbb{1}_{(i=j)}$. Thus, for a constant segment on $[x_l, x_{l+1}-]$ in $F(x)$, we have $R_{ik_{2l+1}}=R_{ik_{2l}}$ and so, recursion (18) is obsolete. Therefore, asymptotically, when $k_{2l}=k_{2l+1}$, we only need to consider the convergence of recursion (19) for a flat-jump segment in $F(x)$, whereas for increasing-jump segment, both recursions (18) and (19) generate terms in the resulting expression, in particular (16) and (17). Now, applying the asymptotic arguments outlined on page 1018 of Schmid (1958), one easily obtains formula (15). □

Let us note that Proposition 2.3 does not cover the case when $f_{2J}<f_{2J+1}\equiv 1$. This case is addressed in the following proposition, which follows by similar reasoning.

Proposition 2.4. *Assuming that a CDF $F(x)$ has the structure of jumps as in (14) and that $f_{2J}<f_{2J+1}\equiv 1$, $v_m+w_m=J$, we have*

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty}\cdots\sum_{j_{v_m}=-\infty}^{\infty}\sum_{j_{v_m+1}=-\infty}^{\infty}\left((-1)^{j_1+\cdots+j_{v_m}+j_{v_m+1}}\right)c' \int_{-\lambda}^{\lambda}\cdots\int_{-\lambda}^{\lambda}\exp\{\psi'\}dz_1\cdots dz_{2v_m+w_m}, \tag{20}$$

where

$$c'=c(f_{2J+1}-f_{2J})^{-1/2}(2\pi)^{-1/2}, \quad \text{and} \quad \psi'=\psi+\frac{(-(-1)^{j_{v_m+1}}z_{2v_m+w_m}-2\lambda j_{v_m+1})^2}{f_{2J+1}-f_{2J}}, \tag{21}$$

with c and ψ in (21) defined in (16), (17), noting that $z_{2v_m+w_m}\neq 0$.

Remark 2.5. Let us note that (12) is a special case of (20) when $m=1, \omega_1\equiv w_1=0, \nu_1\equiv v_1=J$.

Corollary 2.6. *When $F(x)$ is purely discrete with J jumps, the limiting distribution $\Phi(\lambda)$ in (15) becomes*

$$\Phi(\lambda)=(2\pi)^{-\frac{J-1}{2}}\prod_{l=1}^{J}(f_{2l}-f_{2l-1})^{-\frac{1}{2}}\int_{-\lambda}^{\lambda}\cdots\int_{-\lambda}^{\lambda}\exp\Big[-\frac{1}{2}\Big(\sum_{l=1}^{J}\frac{(z_l-z_{l-1})^2}{f_{2l}-f_{2l-1}}\Big)\Big]dz_1\cdots dz_{J-1}, \tag{22}$$

where $z_0=z_J=0$.

Proof: Since the jump structure in this case includes only one group of flat-jump segments of size J, the first group of increasing-jump segments in (14) is empty, i.e., $m=1, \nu_1\equiv v_1=0$, $\omega_1\equiv w_1=J$, and by convention, $\prod_{l=1}^{\nu_1=0}(\cdot)=1, \sum_{l=1}^{\nu_1=0}(\cdot)=0$. Substituting these in (15), (16), and (17), we have

$$c=(2\pi)^{-\frac{J-1}{2}}\prod_{l=1}^{J}(f_{2l}-f_{2l-1})^{-\frac{1}{2}}, \quad \psi=-\frac{1}{2}\Big(\sum_{l=1}^{J}\frac{(z_l-z_{l-1})^2}{f_{2l}-f_{2l-1}}\Big),$$

and (15) becomes (22). □

Remark 2.7. It should be noted that (22) is the formula for the distribution of an $(J-1)$ dimensional Brownian bridge between $-\lambda$ and λ. The Brownian bridge interpretation has been used by Wood and Altavela (1978) to compute via Monte Carlo (MC) simulation the asymptotic distribution of D_n, without relating the interpretation to an explicit expression such as (22).

Next, we give an illustrative example on how to use the asymptotic distribution formula (15) given by Proposition 2.3, for mixed $F(x)$. Similarly, one can employ expressions (20) and (22) on appropriate specific examples.

Example 2.8. Consider a random variable X with CDF

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.2 + x & \text{if } 0 \le x < 0.2, \\ 0.5 & \text{if } 0.2 \le x < 0.8, \\ x - 0.1 & \text{if } 0.8 \le x < 1, \\ 1 & \text{if } x \ge 1. \end{cases} \tag{23}$$

Clearly, $F(x)$ is a CDF with four jumps, i.e., $J = 4$, at $x_1 = 0$, $x_2 = 0.2$, $x_3 = 0.8$, $x_4 = 1.0$, and $f_0 = f_1 = 0$, $f_2 = 0.2$, $f_3 = 0.4$, $f_4 = f_5 = 0.5$, $f_6 = 0.7$, $f_7 = 0.9$, $f_8 = f_9 = 1$. Since the jump structure of $F(x)$ in (23) is flat-jump, increasing-jump, flat-jump, increasing-jump segments, the first set of increasing-jump segments and the last set of flat-jump segments in (14) should be omitted. Therefore, $m = 3$, $\nu_1 = 0$, $\omega_1 = 1$, $\nu_2 = 1$, $\omega_2 = 1$, $\nu_3 = 1$, $\omega_3 = 0$, and $v_0 = 0$, $v_1 = 0$, $v_2 = 1$, $v_3 = 2$, $w_0 = 0$, $w_1 = 1$, $w_2 = 2$, $w_3 = 2$. Substituting these in (15), (16), and (17), we obtain

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \sum_{j_2=-\infty}^{\infty} c(-1)^{j_1+j_2} \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\{\psi\} dz_1 \cdots dz_5, \tag{24}$$

where

$$c = (2\pi)^{-\frac{5}{2}} (f_2 - f_1)^{-1/2} (f_3 - f_2)^{-1/2} (f_4 - f_3)^{-1/2} (f_6 - f_5)^{-1/2} (f_7 - f_6)^{-1/2} (f_8 - f_7)^{-1/2},$$

and

$$\psi = -\frac{1}{2}\Big(\frac{z_1^2}{f_2 - f_1} + \frac{(z_2 - (-1)^{j_1} z_1 - 2\lambda j_1)^2}{f_3 - f_2} + \frac{(z_3 - z_2)^2}{f_4 - f_3} + \frac{(z_4 - z_3)^2}{f_6 - f_5} + \frac{(z_5 - (-1)^{j_2} z_4 - 2\lambda j_2)^2}{f_7 - f_6} + \frac{z_5^2}{f_8 - f_7} \Big).$$

3. Software implementation and numerical analysis

In this section, we introduce the C++ and the R implementation of the proposed FFT-based method for computing $\mathsf{P}(D_n \ge q)$, described in Section 2.1 and study its numerical properties. In the sequel, we will refer to it as the exact-KS-FFT method. The method is implemented in the R package **KSgeneral** which can be downloaded from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=KSgeneral. In order to build

the **KSgeneral** package from source, a C++ compiler is required. The latter is contained in the Windows program **Rtools** (R Core Team 2020), or under MacOS in **Xcode**, downloadable from the App Store. The package **KSgeneral** uses **Rcpp** (Eddelbuettel and François 2011) in R, and utilizes the C++ code that efficiently computes $\mathsf{P}(D_n \geq q)$ using the exact-KS-FFT method (see the replication material to this paper available online). Since the latter requires computation of FFT, the **FFTW3** library developed by Frigo and Johnson (2005) needs to be installed from `http://www.fftw.org/`. It should be noted that both **Rtools** and **FFTW3** should be installed in the system `PATH`.

In this section, we also study the asymptotic formulae (15) and (22) given in Section 2.2, which have been implemented in Mathematica 10. For this purpose, in Sections 3.1 and 3.2, we compute the complementary CDF, $\mathsf{P}(D_n \geq q)$, for different values of n and q, and also compute related p values when $F(x)$ is mixed and discrete, respectively. Then, in Section 3.3 we consider $\mathsf{P}(D_n < q)$ and $\mathsf{P}(D_n \geq q)$ in the case of continuous $F(x)$. For the examples given in all three Sections 3.1, 3.2, 3.3 (and in the replication material), we give the lines of code that should be executed in C++ or R using **KSgeneral**. Furthermore, in the case when $F(x)$ is mixed (cf., Section 3.1), we compare the exact probabilities $\mathsf{P}(D_n \geq q)$, $q \in [0, 1]$, obtained using the exact-KS-FFT approach with those obtained using the asymptotic formula (15). In addition, when $F(x)$ is purely discrete (cf., Section 3.2), we also compare with the results of the Brownian bridge simulation-based algorithm of Wood and Altavela (1978). When $F(x)$ is continuous, in Section 3.3, Appendix B, Appendix C, and Appendix D, we compare the accuracy and speed of the exact-KS-FFT method to the results obtained from the R program of Carvalho (2015), and the C program due to Simard and L'Ecuyer (2011). The reported CPU times are obtained running the related C++ code on a machine with an 2.5GHz Intel Core i5 processor with 4GB RAM, running Mac OS X Yosemite.

3.1. Complementary CDF of D_n when $F(x)$ is mixed

In order to illustrate the performance of the exact-KS-FFT method of Section 2.1, we consider first the following example from excess-of-loss reinsurance.

Example 3.1. Consider an excess-of-loss reinsurance contract with a retention level M and a limiting level L, where $0 < M < L$ are positive constants. Under such a contract, given a loss amount random variable X with a continuous CDF $F_X(\cdot)$ on $[0, +\infty)$, the insurer and the reinsurer pay correspondingly the amounts Z and Y, where

$$Z = \begin{cases} X & \text{if } X \leq M, \\ M & \text{if } M < X \leq L, \\ M + X - L & \text{if } L < X, \end{cases} \quad \text{and} \quad Y = \begin{cases} 0 & \text{if } X \leq M, \\ X - M & \text{if } M < X \leq L, \\ L - M & \text{if } L < X. \end{cases}$$

Clearly, both Z and Y are mixed random variables with correspondingly, one and two jumps in their CDFs. For illustrative purposes, assume that the CDF of Y, $F_Y(y)$ is of the form

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ 1 - 0.5e^{-y} & \text{if } 0 \leq y < \log 2.5, \\ 1 & \text{if } y \geq \log 2.5, \end{cases} \tag{25}$$

where $M = \log 2$, $L = \log 5$, $F_X(x) = 1 - e^{-x}$. Assuming D_n in (1) is defined with respect to $F_Y(y)$, i.e., $F(x) \equiv F_Y(y)$ in (1), we have computed exact probabilities $\mathsf{P}(D_n \geq q)$, for

different values of n and q, applying the exact-KS-FFT method and also, the asymptotic formula (15). In order to apply (15), one should note that $F_Y(y)$ has two jumps, (i.e., $J = 2$) at $x_1 = 0, x_2 = \log 2.5$, and $f_0 = f_1 = 0, f_2 = 0.5, f_3 = 0.8, f_4 = f_5 = 1$. Since the jump structure of $F_Y(\cdot)$ in (25) is flat-jump, increasing-jump segments, the first set of increasing-jump segments and the last set of flat-jump segments in (14) should be omitted. Therefore, one should apply formula (15) with $m = 2, \nu_1 = 0, \omega_1 = 1, \nu_2 = 1, \omega_2 = 0$, and $v_0 = 0, v_1 = 0, v_2 = 1, w_0 = 0, w_1 = 1, w_2 = 1$.

The results for $\mathsf{P}(D_n \geq q)$ calculated using the proposed FFT-based method and the asymptotic formula (15), for different values of n, q, and respectively $\lambda = qn^{1/2}$, are shown in Table 1. For example, to obtain the probability $\mathsf{P}(D_n \geq q)$ using C++, for $n = 25$, $q = 0.60$ as shown in the column exact-KS-FFT of Table 1, according to step (i) of the procedure exact-KS-FFT, we first define the mixed CDF in (25) in the file `crossprob.cc` using the following code.

```
vector <double> MixDF (vector <double> obs){
    vector <double> observed = obs;
    set<double> s;
    for (int i = 0; i < obs.size(); ++i){
        s.insert(obs[i]);
    }
    obs.assign(s.begin(), s.end());
    vector <double> DF(obs.size());
    /* The distribution in the reinsurance example in (25) */
        for (int i = 0; i < obs.size(); ++i){
            if (obs[i] < 0.0){
                DF[i] = 0.0;
            }
            else if (obs[i] < log(2.5)){
                DF[i] = 1 - 0.5 * exp(-1.0 * obs[i]);
            }
            else
            {
                DF[i] = 1.0;
            }
        }
    return DF;
}
```

Also, since the mixed CDF in (25) has jumps at $y = 0$ and $y = \log 2.5$, we need to specify this by inputting `vector_input3 = {0.0, log(2.5)};` to the `int main()` function in the file `crossprob.cc`.

Next, we first run `make` in one of the command line tools (e.g., bash) to build the program for the exact-KS-FFT method developed in this paper, based on the code provided by Moscovich and Nadler (2017). Then, in the command line tool, we run the following line `./bin/crossprob ecdf 25 Boundary_Crossing_Time.txt`, where 25 is the input for the sample size. We will have the following screen prompts.

```
Please enter the distribution type: 1 for Continuous Distribution,
2 for Discontinuous Distributions:
```

We enter 2 since the CDF in (25) is not continuous.

```
2
```

Then, we can choose whether to calculate the KS complementary CDF, $\mathsf{P}(D_n \geq q)$, or the p value, $\mathsf{P}(D_n \geq d_n)$ corresponding to a value d_n computed based on a user provided data sample.

```
Please enter the objective: 1 for KS Complementary Distribution,
2 for P-Values:
```

Since we want to obtain the probability $\mathsf{P}(D_n \geq q)$, for $n = 25$, $q = 0.6$, we will enter 1.

```
1
```

Here, we enter the sample size n and the quantile q.

```
Please enter the sample size and quantile:

25
0.6

Probability: 0.0000000019082332
Time taken: 0.0000720000000000
```

Now, steps (ii), (iii), (iv) and (v) of the procedure exact-KS-FFT are performed. The result for $\mathsf{P}(D_n \geq q)$, for $n = 25$, $q = 0.60$, is 1.90823×10^{-9} as shown in the column exact-KS-FFT of Table 1. The corresponding computation time is also printed.

Remark 3.2. Note that the distribution of the KS test statistic D_n depends on the hypothesized distribution $F(x)$ when $F(x)$ is not continuous. Hence, to obtain $\mathsf{P}(D_n \geq q)$ for different mixed $F(x)$, the users should: 1) define the mixed CDF in the file `crossprob.cc` each time, and 2) in the file `crossprob.cc`, define the vector containing points where $F(x)$ has jumps, `vector_input3`.

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is mixed using the R package **KSgeneral**, one needs to input `mixed_ks_c_cdf(q, n, jump_points, Mixed_dist, ..., tol = 1e - 10)`, where `jump_points` is a numeric vector of the x coordinates of the jumps of $F(x)$, `Mixed_dist` specifies the mixed CDF $F(x)$, possibly followed by a list of parameters `...` specifying $F(x)$, and `tol` is the value of ϵ that is used to compute the values A_i and B_i, $i = 1, \ldots, n$, as detailed in equations (4) in Step 1 of Section 2.1. By default, `tol = 1e - 10`. Note that a value `NA` or `0` will lead to an error. For instance, if one wants to use the R package **KSgeneral** to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is the mixed CDF specified in Example 3.1 by equation (25), with $n = 25$, $q = 0.1$, one needs to run the following code in order to obtain the corresponding result, as shown in Table 1 for $n = 25, q = 0.1$.

```
R> Mixed_cdf_example <- function(x) {
+    result <- 0
+    if (x < 0) {
+      result <- 0
+    } else if (x == 0) {
+      result <- 0.5
+    } else if (x < log(2.5)) {
+      result <- 1 - 0.5 * exp(-x)
+    } else {
+      result <- 1
+    }
+    return(result)
+  }
R> mixed_ks_c_cdf(0.1, 25, c(0, log(2.5)), Mixed_cdf_example)

[1] 0.76768489
```

From Table 1, one can first see that the exact-KS-FFT method effectively computes $\mathsf{P}(D_n \geq q)$ for small, medium and large sample sizes n and various levels q, and gives exact probabilities in the range of 10^{-10} to 1. It should be noted though that the method could become numerically unstable (producing negative values) when calculating probabilities of 10^{-11} or smaller. Similar issue has been observed by Simard and L'Ecuyer (2011) in the case of continuous $F(x)$. The column "Rel. err. (%)" quantifies the relative error of the asymptotic value (for fixed λ) compared to the exact values in the column exact-KS-FFT (for various combinations of n and q resulting in the same λ). Furthermore, we see that when the sample size n is large, results using formula (15) approximate closely the exact $\mathsf{P}(D_n \geq q)$, except when $\mathsf{P}(D_n \geq q)$ is nearly zero (when $\lambda = 2, 3$ in Table 1). Also, asymptotic formula (15) gives better approximations to the exact values of $\mathsf{P}(D_n \geq q)$ as q decreases, or equivalently, as $\mathsf{P}(D_n \geq q)$ increases. Moreover, as λ decreases, values obtained from asymptotic formula (15) become better approximations to the exact $\mathsf{P}(D_n \geq q)$. Let us recall however that formula (15) (respectively (20) and (22)) is only practically implementable for small/moderate number of jumps, J, in the null distribution (which is the case with (25) illustrated in Table 1), as otherwise the multidimensional integration becomes infeasible.

In addition, as mentioned in Section 1, a null hypothesis that a sample comes from a discontinuous distribution will be accepted more often if one uses the continuous KS test, as opposed to using the discontinuous KS test. To illustrate this, assume that a random sample of size $n = 25$ follows $F(x) \equiv F_Y(y)$ in (25) under H_0, and that the KS test statistic for the sample is $d_n = 0.25$. Then, the exact p value of the test is $\mathsf{P}(D_n \geq 0.25|H_0) = 0.04496610$ and, with a significance level of 5%, one should reject H_0. On the other hand, a p value calculated using the complementary CDF of the distribution-free continuous KS test statistic D_n (i.e., when $F(x)$ in (1) is continuous) is $0.07360597 > 0.05$. Therefore, based on the latter p value, one will not reject H_0. Similar situations are illustrated in Table 2 for larger sample sizes and different values of the test statistic D_n, where one can see that the differences between the values in the last two columns are higher than 58% (our experience shows that these are typically in the range 50%–65%) and do not decrease with n. To the best of our knowledge, the KS test in software packages such as R, SPSS, Stata, MATLAB, Mathematica is based on the distribution-free continuous KS test statistic and the discontinuous (mixed and purely

λ	n	q	Exact-KS-FFT		Asympt. (15)	Rel. err. (%)
3	25	0.60	1.90823×10^{-9}	(0.000)	1.72031×10^{-8}	801.52
	100	0.30	9.49583×10^{-9}	(0.000)	(5155.54)	81.16
	400	0.15	1.41586×10^{-8}	(0.015)		21.50
	2500	0.06	1.62830×10^{-8}	(0.202)		5.65
	10000	0.03	1.67953×10^{-8}	(2.932)		2.43
	40000	0.015	1.69543×10^{-8}	(59.86)		1.47
	90000	0.01	1.70084×10^{-8}	(351.9)		1.14
	250000	0.006	1.74671×10^{-8}	(3524)		1.51
2	25	0.4	2.13209×10^{-4}	(0.000)	3.98459×10^{-4}	86.89
	100	0.2	3.27304×10^{-4}	(0.000)	(1.17)	21.74
	400	0.1	3.66979×10^{-4}	(0.015)		8.58
	2500	0.04	3.86968×10^{-4}	(0.195)		2.97
	10000	0.02	3.92912×10^{-4}	(2.707)		1.41
	40000	0.01	3.95740×10^{-4}	(57.14)		0.69
	90000	1/150	3.96661×10^{-4}	(341.3)		0.45
	250000	0.004	3.97390×10^{-4}	(3465)		0.27
1	25	0.2	0.151510006	(0.000)	0.174525238	15.19
	100	0.1	0.164499986	(0.000)	(0.73)	6.09
	400	0.05	0.169049900	(0.015)		3.24
	2500	0.02	0.172221536	(0.171)		1.34
	10000	0.01	0.173354312	(2.511)		0.68
	40000	0.005	0.173934996	(54.94)		0.34
	90000	1/300	0.174130680	(330.3)		0.23
	250000	0.002	0.174287993	(3423)		0.14
0.5	25	0.1	0.767684886	(0.000)	0.801038769	4.34
	100	0.05	0.782681427	(0.000)	(5.63)	2.35
	400	0.025	0.790339869	(0.015)		1.35
	2500	0.01	0.796406211	(0.156)		0.58
	10000	0.005	0.798664879	(2.441)		0.30
	40000	0.0025	0.799837547	(54.27)		0.15
	90000	1/600	0.800234794	(326.5)		0.10
	250000	0.001	0.800554870	(3410)		0.06
0.2	25	0.04	0.999798067	(0.000)	0.999961812	0.016
	100	0.02	0.999888190	(0.000)	(5.03)	0.007
	400	0.01	0.999925985	(0.015)		0.004
	2500	0.004	0.999948507	(0.156)		0.001
	10000	0.002	0.999955380	(2.364)		0.001
	40000	0.001	0.999958655	(53.62)		0.000
	90000	1/1500	0.999959721	(324.4)		0.000
	250000	0.0004	0.999960564	(3383)		0.000
0.15	25	0.03	0.999998692	(0.000)	0.999999978	0.000
	100	0.015	0.999999682	(0.000)	(0.51)	0.000
	400	0.0075	0.999999905	(0.015)		0.000
	2500	0.003	0.999999956	(0.156)		0.000
	10000	0.0015	0.999999969	(2.355)		0.000
	40000	0.00075	0.999999974	(53.45)		0.000
	90000	0.0005	0.999999975	(324.7)		0.000
	250000	0.0003	0.999999977	(3372)		0.000

Table 1: Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the exact-KS-FFT method and the asymptotic formula (15),when $\lambda = qn^{1/2} = 3$, 2, 1, 0.5, 0.2 and 0.15, respectively, when the underlying CDF $F(x)$ follows $F_Y(y)$ in (25). Numbers in () are run times in seconds.

n	$D_n = d_n$	Discontinuous KS p values	Continuous KS p values
25	0.25	0.04496610	0.07360597
100	0.13	0.03913182	0.06209234
400	0.065	0.04090172	0.06511744
2500	0.026	0.04200207	0.06690821
10000	0.013	0.04237475	0.06750119
40000	0.0065	0.04256212	0.06779695

Table 2: Discontinuous and continuous KS p values under null hypothesis $H_0 : F(x) \equiv F_Y(y)$, obtained via the exact-KS-FFT method.

discrete) version is not implemented due to the lack of efficient and robust method such as the exact-KS-FFT method we propose here.

Example 3.3. Another possible application of KS tests on mixed distributions appears in testing the goodness-of-fit in zero-inflated or/and one-inflated models. Many real data contain zeros and ones, i.e., have masses at zero and one, and therefore zero- and one-inflated distributions need to be applied. For example, Ospina and Ferrari (2010) have used the zero-and-one-inflated beta distribution to model the proportion of inhabitants living within a 200 kilometer wide coastal strip in 232 countries in the year 2000, denoted as Y. The data for years 1990, 2000 and 2010 are supplied by the Columbia University Centre for International Earth Science Information Network, see CIESIN (2012), and are available at `http://sedac.ciesin.columbia.edu/data/set/nagdc-population-landscape-climate-estimates-v3`.

The zero-and-one-inflated beta distribution considered by Ospina and Ferrari (2010) is of the following form

$$G_Y(y; \alpha, \gamma, \mu, \phi) = \alpha \text{Bernoulli}(y; \gamma) + (1 - \alpha)F(y; \mu, \phi), \quad 0 \le y \le 1,$$

where $\text{Bernoulli}(\cdot; \gamma)$ denotes the CDF of a Bernoulli random variable with parameter γ, $0 < \gamma < 1$, and $F(\cdot; \mu, \phi)$ denotes the CDF of a beta random variable with parameters μ, $0 < \mu < 1$, and $\phi > 0$. Hence, the zero-and-one-inflated distribution can be seen as a mixture of a (discrete) Bernoulli distribution and a (continuous) beta distribution, with weights α and $(1 - \alpha)$, respectively, $0 < \alpha < 1$.

According to Ospina and Ferrari (2010), the random variable Y has the following distribution

$$G_Y(y) = \begin{cases} 0 & \text{if } y < 0, \\ 0.1141 + 0.4795F_Y(y; \mu, \phi) & \text{if } 0 \le y < 1, \\ 1 & \text{if } y \ge 1, \end{cases}$$

where $F_Y(y; \mu, \phi)$ has a density function

$$f_Y(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1 - \mu)\phi)} y^{\mu\phi - 1}(1 - y)^{(1-\mu)\phi - 1},$$

with μ and ϕ estimated as $\mu = 0.6189, \phi = 0.6615$ based on the population data in 2000, and $\Gamma(\cdot)$ is the gamma function. Then, we can examine the goodness-of-fit of the distribution to the population data in 2010, denoted by $\tilde{Y}$, hypothesizing that $\tilde{Y}$ has the same distribution as $G_Y(y)$. Using (1) with $F(x) \equiv G_Y(y)$ and $F_n(x) \equiv G_n(y)$, where $G_n(y)$ is the EDF of $\tilde{Y}$

computed from the population data in 2010, we obtain the KS test statistic $d_n = 0.09047$. Using the exact-KS-FFT method, we compute a p value of $0.03403 < 0.05$. Alternatively, applying the asymptotic formula (15), we obtain a p value of 0.03641, which is reasonably accurate, given the sample size of 232. Therefore, the KS test indicates that the zero-and-one-inflated beta distribution estimated using population data in 2000 does not fit the population data in 2010 at a significance level of 5%, providing evidence for a change in the proportion of inhabitants in the decade.

In order to perform the one-sample two-sided KS test, when $F(x)$ is mixed, one needs to input the code `mixed_ks_test(x, jump_points, Mixed_dist, ..., tol = 1e-10)`, where `x` is a numeric vector of data sample values, and where other arguments are defined similarly as in the function `mixed_ks_c_cdf()`. For instance, if one wants to use the R package **KSgeneral** to calculate the p value for the KS test, when $F(x)$ follows a zero-and-one-inflated beta distribution as in Example 3.3, with a sample of size $n = 232$, one should run the following R code.

```
R> data("Population_Data", package = "KSgeneral")
R> mu <- 0.6189
R> phi <- 0.6615
R> a <- mu * phi
R> b <- (1 - mu) * phi
R> Mixed_cdf_example <- function(x) {
+    result <- 0
+    if (x < 0) {
+      result <- 0
+    } else if (x == 0) {
+      result <- 0.1141
+    } else if (x < 1) {
+      result <- 0.1141 + 0.4795 * pbeta(x, a, b)
+    } else {
+      result <- 1
+    }
+    return (result)
+  }
R> mixed_ks_test(Population_Data, c(0, 1), Mixed_cdf_example)

        One-sample Kolmogorov-Smirnov test

data:  Population_Data
D = 0.0904762, p-value = 0.034018
alternative hypothesis: two-sided
```

In the next section, assuming $F(x)$ is purely discrete, we apply the FFT-based methodology and the asymptotic formula (22) (cf., Corollary 2.6) to compute correspondingly, exact and approximate values of $\mathsf{P}(D_n \geq q)$.

3.2. Complementary CDF of D_n when $F(x)$ is purely discrete

There is an abundance of real-life applications in which purely discrete distributions are used to model count data, such as number of claims to an insurance company, number of jumps

in stock returns, number of trades on the stock exchange, number of manufacturing defects, number of diseased species and plants in biology and agricultural research, and many other count data applications. In all such cases, examining the goodness-of-fit of the model requires computing p values or $\mathsf{P}(D_n \geq q)$ for various of n and q. As an illustration, using the proposed FFT-based method, we will compute exact probabilities $\mathsf{P}(D_n \geq q)$ when the underlying $F(x)$ follows Binomial(r, π) distribution (see Example 3.4) and when it follows a discrete uniform distribution (see Example 3.5). In Example 3.4, we compare these exact probabilities with approximate ones obtained using the asymptotic distribution of D_n, given by (22), and using the asymptotic MC simulation-based method of Wood and Altavela (1978). In Example 3.5, we compare the exact results with those obtained using the R function `ks.test` of Arnold and Emerson (2011). The latter is a revised version of the same function from the base package **stats**.

Wood and Altavela (1978)'s approach utilizes the connection between the asymptotic distribution of D_n and a multi-variate Brownian bridge (cf., Remark 2.7), and they directly simulate the latter, thus avoiding the necessity to derive and evaluate an explicit expression such as (22). Following the Wood and Altavela (1978)'s method, one should simulate from the $(J-1)$-variate normal random vector $(Z_1, Z_2, \ldots, Z_{J-1})$, where

$$\mathsf{E}(Z_i) = 0, \quad \mathsf{E}(Z_i, Z_k) = \min(f_{2i}, f_{2k}) - f_{2i} f_{2k}, \quad i, k = 1, \ldots, J-1, \tag{26}$$

and estimate the probability in $\Phi(\lambda)$ in (11) as

$$\frac{\sum_{i=1}^{N} \mathbb{1}_{\{(Z_1, Z_2, \ldots, Z_{J-1}) \in [-\lambda, \lambda]^{J-1}\}}}{N},$$

where N is the number of simulations, $\mathbb{1}_{\{\cdot\}}$ is an indicator function, and $[-\lambda, \lambda]^{J-1}$ is the $(J-1)$ dimensional hypercube. The authors further suggest a continuity correction for λ in (11), as $\lambda = qn^{1/2} - 0.5n^{-1/2}$. In the remainder of this section, we will refer to this method as W&A(a) method and to its version without the continuity correction, as W&A(b) method.

Example 3.4. Assume that $F(x)$ in (1) is Binomial(r, π) with $r = 3, 7, 15$ (i.e., with $J = r + 1$ number of jumps), and $\pi = 0.5$. In Tables 3, 4, and 5, for different values of n, q, and respectively $\lambda = qn^{1/2}$, we give the exact $\mathsf{P}(D_n \geq q)$ obtained with the exact-KS-FFT method, and compare with the asymptotic probabilities obtained using (22) (combined with (13)), and using the Wood and Altavela (1978) simulation-based approach. We have coded both the W&A(a) and W&A(b) versions in R as part of the **KSgeneral** R package and have simulated 1000000 realizations of the random vector $(Z_1, Z_2, \ldots, Z_{J-1})$. As before, the numbers in parentheses show the computation (run) times, in seconds. Let us note that the multidimensional numerical integration in (22) becomes unstable as the number of jumps, $J = r+1$, in $F(x)$ increases, and so we only use W&A(a) and W&A(b) to obtain approximate asymptotic probabilities in the case of $r = 15$ and $\pi = 0.5$ (see Table 5).

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is purely discrete using the R package **KSgeneral**, one needs to input `disc_ks_c_cdf(q, n, y, ..., exact = NULL, tol = 1e-08, sim.size = 1e+06, num.sim = 10)`, where `y` specifies the purely discrete CDF $F(x)$, possibly followed by a list of parameters `...` specifying $F(x)$, the input parameter `exact` is a logical variable specifying whether one wants to compute exact values for $\mathsf{P}(D_n \geq q)$ using the FFT-based method, `exact = TRUE` or wants to compute the approximate values for $\mathsf{P}(D_n \geq q)$ using the simulation-based algorithm of Wood and Altavela (1978), in which case `exact =`

FALSE. When exact = NULL and n <= 100000, the exact $\mathsf{P}(D_n \geq q)$ will be computed using the FFT-based method. The input parameter tol is the value of ϵ that is used to compute the values A_i and B_i, $i = 1, \ldots, n$, as detailed in equations (4) in Step 1 of Section 2.1. By default, tol = 1e-08. Note that a value of NA or 0 will lead to an error. The input parameter sim.size is the required number of simulated trajectories in order to produce one MC estimate (one MC run) of the asymptotic p value using the algorithm of Wood and Altavela (1978). By default, sim.size = 1e + 06. The input parameter num.sim is the number of MC runs, each producing one estimate (based on sim.size number of trajectories), which are then averaged in order to produce the final estimate for the asymptotic p value. This is done in order to reduce the variance of the final estimate. By default, num.sim = 10. For instance, if one wants to use the R package **KSgeneral** to compute the exact value for $\mathsf{P}(D_n \geq q)$, when $F(x)$ follows a Binomial$(3, 0.5)$ distribution as in Example 3.4, with $n = 400$, $q = 0.05$, one should run the following R code and obtain the corresponding result as shown in the column exact-KS-FFT of Table 3.

```
R> binom_3 <- stepfun(0:3, c(0, pbinom(0:3, 3, 0.5)))
R> disc_ks_c_cdf(0.05, 400, binom_3)

[1] 0.05611849
```

On the other hand, if one wants to use the simulation-based method of Wood and Altavela (1978) in order to approximate the asymptotic value for $\mathsf{P}(D_n \geq q)$, when $F(x)$ follows a Binomial$(3, 0.5)$ distribution, with $n = 400$, $q = 0.05$, one should use the W&A(a) method, by running the following R code and obtain the corresponding result as shown in the column W&A(a) of Table 3.

```
R> binom_3 <- stepfun(0:3, c(0, pbinom(0:3, 3, 0.5)))
R> disc_ks_c_cdf(0.05, 400, binom_3, exact = FALSE, tol = 1e-08,
+    sim.size = 1e+06, num.sim = 10)

[1] 0.0561998
```

Looking at Tables 3, 4, and 5, one can see that the exact-KS-FFT method effectively computes $\mathsf{P}(D_n \geq q)$ for small, medium and large sample sizes n and various levels q, and gives exact probabilities in the range 10^{-12} to 1. We also see that when the sample size n is large, results using formula (22) approximate closely the exact $\mathsf{P}(D_n \geq q)$, except when $\mathsf{P}(D_n \geq q)$ is nearly zero (when $\lambda = 2, 3$ in Tables 3 and 4). Similarly to the mixed $F(x)$ case, asymptotic formula (22) gives better approximations to the exact values of $\mathsf{P}(D_n \geq q)$ as q decreases, or equivalently, as $\mathsf{P}(D_n \geq q)$ increases. Moreover, as λ decreases, values obtained from asymptotic formula (22) become better approximations to the exact $\mathsf{P}(D_n \geq q)$. One can further observe that asymptotic formula (22) and W&A(b) method provide similar results. In particular, as the number of jumps in $F(x)$ increases, results obtained from these two methods almost coincide. In addition, when the number of jumps in $F(x)$ is small (in our case $J = 4$ or 8), we see that values obtained from W&A(a) method provide more accurate approximations to the exact probabilities. On the other hand, when the number of jumps in $F(x)$ is large (in our case $J = 16$), values obtained from W&A(b) method give closer approximations. In comparison with the exact-KS-FFT method, W&A(a) and W&A(b) deviate stronger from

the exact probabilities for moderate values of λ, e.g., $\lambda = 0.5, 1$, and this is more pronounced for small sample sizes, see $n \leq 400$.

With regards to computation time, looking at Tables 3, 4, and 5, for fixed sample size n and number of jumps J, as λ decreases, the computation time for the exact-KS-FFT method, W&A(a) method and W&A(b) method decreases. Furthermore, when the sample size n and q are fixed, as the number of jumps in $F(x)$, J, increases, the computation time for the exact-KS-FFT method decreases, whereas the computation time for W&A(a) and W&A(b) methods increases. And, as expected, when the sample size n increases, the exact-KS-FFT method becomes more time-consuming.

Example 3.5. Next, we consider another illustrative example where we compare the performance of the proposed exact-KS-FFT method with the R function `ks.test` from the package **dgof** (Arnold and Emerson 2011). Hypothesizing that the underlying $F(x)$ in (1) follows a discrete uniform distribution on $[1, 10]$, we have simulated random samples of size n, $25 \leq n \leq 100000$, from the discrete uniform distribution on $[1, 10]$ and have performed KS tests on the simulated samples. In Table 6, we compute p values corresponding to different values of the test statistic D_n for the simulated samples of size n.

In order to perform the one-sample two-sided KS test, when $F(x)$ is purely discrete, one needs to input the `disc_ks_test(x, y, ..., exact = NULL, tol = 1e-08, sim.size = 1e+06, num.sim = 10)`, where `x` is a numeric vector of data sample values, and where other arguments are defined similarly as in the function `disc_ks_c_cdf()`. For instance, in order to calculate the p value for the KS test, when $F(x)$ follows a discrete uniform distribution on $[1, 10]$ as in Example 3.5, with a sample size $n = 1000$, one should run the following R code.

```
R> set.seed(1234567)
R> x4 <- sample(1:10, 1000, replace = TRUE)
R> disc_ks_test(x4, ecdf(1:10), exact = TRUE)

        One-sample Kolmogorov-Smirnov test

data:  x4
D = 0.02, p-value = 0.54235
alternative hypothesis: two-sided
```

As can be seen from Table 6, the exact-KS-FFT method produces exact p values for all sample sizes $25 \leq n \leq 100000$, whereas the function `ks.test` becomes numerically unstable (NU) for $n \geq 100$, as noted also by Arnold and Emerson (2011). To avoid instability, for large n the function `ks.test` allows for estimating p values via simulation, which may be insufficiently accurate or prohibitively time consuming, depending on the choice of the number of simulations (cf., the column `ks.test`(simulation) in Table 6 where the number of simulations is 2000). In contrast to the `ks.test` function, using the exact-KS-FFT method, one obtains the exact p value 0.3343 for sample size $n = 100000$ in less than 2 minutes without any simulation. Moreover, note that the p values in the column `ks.test`(simulation) in Table 6 are based on the suggested default number of 2000 replicates (i.e., obtained by implementing the R code `dgof::ks.test(x, ecdf(1:10), simulated.p.value = TRUE, B = 2000)`). Thus, each estimated p value is likely to be different if we run another simulation and the relative

λ	n	q	Exact-KS-FFT		Asympt. (22)	Rel. err. (%)	W&A(a)		Rel. err. (%)	W&A(b)	Rel. err. (%)
3	25	0.60	1.15052×10^{-12}	(0.000)	1.973175×10^{-9}		0	(6.06)		0	
	400	0.15	2.04623×10^{-9}	(0.015)	(10.55)	-3.570	0	(6.14)		(6.29)	
	10000	0.03	2.07709×10^{-9}	(3.291)		-5.003	0	(6.10)			
	90000	0.01	1.90246×10^{-9}	(427.5)		3.717	0	(6.13)			
2	25	0.40	1.99454×10^{-5}	(0.000)	6.334532×10^{-5}		1.369×10^{-4}	(6.06)		5.91×10^{-5}	
	400	0.10	7.43068×10^{-5}	(0.015)	(10.82)	-14.752	7.22×10^{-5}	(6.13)	-2.835	(6.16)	-20.465
	10000	0.02	6.59391×10^{-5}	(3.010)		-3.934	6.17×10^{-5}	(6.14)	-6.429		-10.372
	90000	1/150	6.42285×10^{-5}	(414.1)		-1.375	5.98×10^{-5}	(6.16)	-6.895		-7.985
1	25	0.20	0.046850021	(0.000)	0.04943858		0.0817293	(5.92)		0.0494865	
	400	0.05	0.056118495	(0.015)	(5.30)	-11.903	0.0561998	(6.00)	0.145	(6.04)	-11.818
	10000	0.01	0.050721030	(2.776)		-2.528	0.0507686	(6.01)	0.094		-2.434
	90000	1/300	0.049863086	(400.2)		-0.851	0.0499037	(6.02)	0.081		-0.755
0.5	25	0.10	0.532599669	(0.000)	0.46014460		0.6300309	(5.15)		0.4604677	
	400	0.025	0.500282800	(0.015)	(10.65)	-8.023	0.5011142	(5.38)	0.166	(5.44)	-7.959
	10000	0.005	0.468139770	(2.574)		-1.708	0.4684648	(5.40)	0.069		-1.639
	90000	1/600	0.462807932	(392.2)		-0.575	0.4631272	(5.42)	0.069		-0.506
0.2	25	0.04	0.935407699	(0.000)	0.92701801		0.9896557	(4.13)		0.9270207	
	400	0.01	0.949180930	(0.015)	(10.84)	-2.335	0.9492119	(4.33)	0.003	(4.40)	-2.335
	10000	0.002	0.931797646	(2.527)		-0.513	0.9317807	(4.35)	-0.002		-0.513
	90000	1/1500	0.928630334	(389.4)		-0.174	0.9286420	(4.39)	0.001		-0.173
0.1	25	0.020	0.999999999	(0.000)	0.98963108		1	(3.88)		0.9896557	
	400	0.005	0.995546700	(0.015)	(10.84)	-0.594	0.9955451	(4.02)	0.000	(4.13)	-0.592
	10000	0.001	0.991072365	(2.480)		-0.145	0.9910796	(4.10)	0.000		-0.143
	90000	1/3000	0.990126719	(388.8)		-0.050	0.9901471	(4.11)	0.002		-0.048

Table 3: Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the exact-KS-FFT method, the asymptotic formula (22) and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(3, 0.5)$ distribution. Numbers in () are run times in seconds.

λ	n	q	Exact-KS-FFT		Asympt. (22)	Rel. err. (%)	W&A(a)		Rel. err. (%)	W&A(b)	Rel. err. (%)
3	25	0.60	2.74895×10^{-10}	(0.000)	6.9080877×10^{-4}		0	(11.43)		0	
	400	0.15	2.06160×10^{-9}	(0.015)	(32.43)		0	(10.92)		(10.93)	
	10000	0.03	2.08096×10^{-9}	(2.074)			0	(10.99)			
	90000	0.01	1.91653×10^{-9}	(259.6)			0	(10.94)			
2	25	0.40	4.20725×10^{-5}	(0.000)	$-5.7657653 \times 10^{-5}$		1.543×10^{-4}	(10.95)		6.54×10^{-5}	
	400	0.10	7.91684×10^{-5}	(0.015)	(34.01)		8.10×10^{-5}	(10.93)	2.314	(10.98)	−17.391
	10000	0.02	6.93244×10^{-5}	(1.840)			6.80×10^{-5}	(10.94)	−1.910		−5.661
	90000	1/150	6.75595×10^{-5}	(244.9)			6.64×10^{-5}	(10.98)	−1.716		−3.196
1	25	0.20	0.068266018	(0.000)	0.07016012		0.1154450	(10.58)		0.0700717	
	400	0.05	0.074899103	(0.015)	(35.39)	−6.327	0.0796742	(10.65)	6.375	(10.75)	−6.445
	10000	0.01	0.070933439	(1.606)		−1.090	0.0719082	(10.71)	1.374		−1.215
	90000	1/300	0.070290581	(233.2)		−0.186	0.0706828	(10.74)	0.558		−0.311
0.5	25	0.10	0.619487745	(0.000)	0.56366243		0.7375863	(8.31)		0.5638245	
	400	0.025	0.583754412	(0.015)	(45.96)	−3.442	0.6068312	(8.76)	3.953	(8.96)	−3.414
	10000	0.005	0.567662656	(1.481)		−0.705	0.5723828	(8.93)	0.832		−0.676
	90000	1/600	0.564996352	(221.4)		−0.236	0.5666604	(8.95)	0.295		−0.207
0.2	25	0.04	0.976334785	(0.000)	0.97713513		0.9993882	(6.01)		0.9771208	
	400	0.01	0.983298737	(0.015)	(89.49)	−0.627	0.9875645	(6.61)	0.434	(6.80)	−0.628
	10000	0.002	0.978475846	(1.404)		−0.137	0.9795275	(6.73)	0.107		−0.138
	90000	1/1500	0.977587940	(216.5)		−0.046	0.9779527	(6.78)	0.037		−0.048
0.1	25	0.020	0.999999999	(0.000)	0.99938850		1	(4.97)		0.9993882	
	400	0.005	0.999745396	(0.015)	(11.75)	−0.036	0.9998985	(5.72)	0.015	(6.00)	−0.036
	10000	0.001	0.999472182	(1.388)		−0.008	0.9995561	(5.90)	0.008		−0.008
	90000	1/3000	0.999417006	(214.6)		−0.003	0.9994476	(5.97)	0.003		−0.003

Table 4: Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the exact-KS-FFT method, the asymptotic formula (22) and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(7, 0.5)$ distribution. Numbers in () are run times in seconds.

λ	n	q	Exact-KS-FFT		W&A(a)		Rel. err. (%)	W&A(b)	Rel. err. (%)
3	25	0.60	4.08521×10^{-10}	(0.000)	0	(21.75)		0	
	400	0.15	2.32765×10^{-9}	(0.015)	0	(21.78)		(21.87)	
	10000	0.03	2.21509×10^{-9}	(1.622)	0	(21.79)			
	90000	0.01	2.10453×10^{-9}	(186.3)	0	(21.86)			
2	25	0.40	6.62012×10^{-5}	(0.000)	1.993×10^{-4}	(21.60)		8.02×10^{-5}	
	400	0.10	9.95661×10^{-5}	(0.015)	1.011×10^{-4}	(21.63)	1.541	(21.80)	−19.45
	10000	0.02	9.05026×10^{-5}	(1.387)	8.36×10^{-5}	(21.70)	−7.627		−11.38
	90000	1/150	8.88601×10^{-5}	(173.5)	8.16×10^{-5}	(21.76)	−8.170		−9.75
1	25	0.20	0.089163050	(0.000)	0.1451183	(20.65)		0.0895510	
	400	0.05	0.093364526	(0.015)	0.1014416	(21.02)	8.651	(21.11)	−4.085
	10000	0.01	0.090270911	(1.138)	0.0918437	(21.13)	1.742		−0.798
	90000	1/300	0.089721687	(161.2)	0.0903139	(21.10)	0.660		−0.190
0.5	25	0.10	0.715781619	(0.000)	0.8181717	(15.09)		0.6503478	
	400	0.025	0.659784355	(0.015)	0.6937999	(16.38)	5.156	(16.57)	−1.430
	10000	0.005	0.652226764	(1.045)	0.6590198	(16.49)	1.042		−0.288
	90000	1/600	0.650966899	(155.2)	0.6532165	(16.54)	0.346		−0.095
0.2	25	0.04	0.992964654	(0.000)	0.9999664	(11.24)		0.9936400	
	400	0.01	0.994406641	(0.015)	0.9973480	(12.10)	0.296	(12.35)	−0.077
	10000	0.002	0.993769635	(0.967)	0.9945804	(12.24)	0.082		−0.013
	90000	1/1500	0.993672471	(151.6)	0.9939733	(12.34)	0.030		−0.003
0.1	25	0.020	0.999999999	(0.000)	1	(7.52)		0.9999664	
	400	0.005	0.999974260	(0.015)	0.9999980	(10.80)	0.002	(11.22)	−0.001
	10000	0.001	0.999966686	(0.951)	0.9999789	(11.04)	0.001		0.000
	90000	1/3000	0.999965549	(150.1)	0.9999703	(11.14)	0.000		0.000

Table 5: Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the exact-KS-FFT method and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(15, 0.5)$ distribution. Numbers in () are run times in seconds.

n	$D_n = d_n$	Exact-KS-FFT		`ks.test`		`ks.test`(simulation)		W&A(a)	
25	0.2	0.1523	(0.0000)	0.1523	(0.007)	0.1475	(0.79)	0.1910	(12.63)
30	0.2	0.1133	(0.0000)	0.1133	(0.007)	0.1075	(0.84)	0.1194	(12.73)
50	0.22	0.007164	(0.0000)	0.007167	(0.014)	0.0045	(1.10)	0.0078125	(13.36)
100	0.2	0.00021	(0.0000)	NU		0.0000	(4.10)	0.0002210	(13.80)
1000	0.02	0.5424	(0.0150)	NU		0.5450	(8.35)	0.5429	(11.08)
5000	0.0094	0.4779	(0.2340)	NU		0.4870	(68.37)	0.4781	(10.92)
10000	0.0065	0.4975	(0.8890)	NU		0.4980	(123.98)	0.4977	(11.08)
100000	0.00241	0.3343	(118.85)	NU		–	–	0.3344	(11.80)

Table 6: p values obtained via the exact-KS-FFT method, the R function `ks.test`, and W&A(a) method, when the underlying CDF $F(x)$ follows a discrete uniform distribution on $[1, 10]$. Numbers in () are run times in seconds.

$F(x)$	n, q	Exact-KS-FFT	`ks.test`(simulation)	Rel. err.
Binomial$(3, 0.5)$	10000, 0.02	0.0000659	0	100%
			0	100%
			0	100%
Binomial$(3, 0.5)$	400, 0.05	0.05612	0.0595	6.0%
			0.0520	7.3%
			0.0600	6.9%
Binomial$(7, 0.5)$	10000, 0.01	0.07093	0.0755	6.4%
			0.0795	12.1%
			0.0690	2.7%
Binomial$(7, 0.5)$	400, 0.05	0.07490	0.0755	0.8%
			0.0670	10.5%
			0.0820	9.5%

Table 7: Differences between the exact and simulated values of $\mathsf{P}(D_n \geq q)$ obtained via the exact-KS-FFT method and the R function `ks.test`, respectively, for certain $n > 100$ and q, when the underlying $F(x)$ follows Binomial$(3, 0.5)$ or Binomial$(7, 0.5)$ distribution.

error will also vary substantially, as we demonstrate in Table 7. To reduce the variation of the simulated p values, one may wish to increase the number of simulations but that will increase even more the computation time and make it prohibitive even for $n > 1000$. In addition, mainly due to the way it has been implemented, for $n > 1000$ the number of simulations cannot be significantly increased, e.g., go beyond 4000 replicates.

As can be seen from Table 7 (which extends Tables 3 and 4) and as also supported by many additional calculations we have run, even for $n \leq 10000$ the accuracy of the R function `dgof::ks.test` may vary substantially for p values in the (rather important) range $(0, 0.1)$.

For small, moderate to large sample sizes (e.g., $25 \leq n \leq 10000$), looking at the column W&A(a) of Table 6, one can see that the alternative MC simulation-based W&A(a) method produces less accurate results and can be significantly slower than the exact-KS-FFT method. W&A(a) performs better in terms of the trade-off between accuracy and speed for very large sample sizes, e.g., $n = 100000$.

To conclude, the proposed method outperforms the R function `ks.test` from the package **dgof** in all of the tested cases. When the number of jumps in the underlying $F(x)$ is small, the asymptotic p value obtained from (22) may not be a good estimate unless sample sizes are very large (e.g., ≥ 40000). Whereas when the number of jumps in $F(x)$ is large, one may use the asymptotic p values to approximate the exact ones for large samples. In the next section, we turn our attention to the case of KS tests with continuous null distributions, which has been widely studied in the literature and for which very efficient numerical procedures have been recently developed.

3.3. (Complementary) CDF of D_n when $F(x)$ is continuous

Our purpose in this section is to illustrate the numerical performance of the proposed FFT-based approach of Section 2.1 and compare it with the state-of-the-art routines of Simard and L'Ecuyer (2011) and Carvalho (2015) developed especially for the case when the underlying CDF, $F(x)$, is strictly continuous. These authors have summarized and enhanced further the

most accurate and efficient methods for computing the distribution of D_n for $F(x)$ continuous, developed earlier in a series of papers e.g., by Durbin (1968), Durbin (1973), Pomeranz (1974), Ruben and Gambino (1982), Marsaglia *et al.* (2003) and Brown and Harvey (2008). For comparison and further details on the implementations of these methods in various statistical software packages, we refer to Simard and L'Ecuyer (2011) and Brown and Harvey (2007). In their recent paper, Simard and L'Ecuyer (2011) have combined into one state-of-the-art program different exact methods to compute the distribution of D_n for different combinations of n and q, based on the relative efficiency and accuracy of the methods. Moreover, for certain combinations of n and q, where the implementations of the exact methods break down (due to cancellation errors, loss of precision and/or prohibitive running time), e.g., for very large n or when the CDF of D_n is close to one, Simard and L'Ecuyer (2011) incorporate in their program various asymptotic formulae for the limiting distribution of D_n. We refer the reader to Section 4 in Simard and L'Ecuyer (2011) for further details. More recently, Carvalho (2015), by avoiding the direct calculation of powers of matrices as required by the approach of Durbin (1973), developed the R package **kolmim** with function `pkolmim` that produces results with similar accuracy as those obtained by the routine of Marsaglia *et al.* (2003), but much faster. However, the related R function becomes too slow when $n > 10000$ as the running time is proportional to n^3 on average. We will show this in Appendix D.

Let us reemphasize that the proposed FFT-based method developed in Section 2.1 is general and thus, applicable also for the case when $F(x)$ is continuous. Hypothesizing on a continuous distribution $F(x)$ leads to certain simplifications. In particular, (3) of Step 1 simplifies to

$$\mathsf{P}(D_n \geq q) = 1 - \mathsf{P}\left(\frac{i}{n} - q \leq U_{(i)} \leq \frac{i-1}{n} + q,\ 1 \leq i \leq n\right), \tag{27}$$

which confirms that the distribution of D_n no longer depends on $F(x)$. Also, (5) of Step 2 simplifies to (27) since the boundaries in (6) become $g(t) = nt - nq$ and $h(t) = nt + nq$, $q \geq 0$ as shown by Durbin (1968). This special case of the proposed FFT-based method has been considered by Moscovich and Nadler (2017) in the general context of computing the probability of non-crossing an upper and a lower boundaries by a Poisson process.

Similarly to Simard and L'Ecuyer (2011) (see Sections 4 and 5 therein), we consider three regions of n, (i) $n \leq 140$, (ii) $140 < n \leq 10^5$, and (iii) $n > 10^5$, forming various sub-regions with respect to q, as specified in Appendix B and Appendix C. Within these sub-regions Simard and L'Ecuyer (2011) use different methods to compute the distribution of D_n. We have performed a thorough numerical comparison across these regions with details given in Appendix B and Appendix C, and can report that, with only a few exceptions, the exact-KS-FFT method returns values that are of at least the same precision as those obtained from the R or C program.

4. Conclusions

We have provided a fast and accurate method to compute $\mathsf{P}(D_n \geq q)$ when $F(x)$ is arbitrary, discontinuous (i.e., mixed or purely discrete) or continuous. The approach we take is to express $\mathsf{P}(D_n \geq q)$ as an appropriate rectangle probability for uniform order statistics and to compute the latter probability using the FFT method. We demonstrate that the proposed exact-KS-FFT method is numerically efficient and robust when hypothesizing on either discontinuous or continuous $F(x)$. In particular, when $F(x)$ is purely discrete the proposed

method outperforms in terms of speed and accuracy the R function of Arnold and Emerson (2011), especially for large sample sizes. Furthermore, in the case of continuous $F(x)$ the exact-KS-FFT method represents a viable alternative to the state-of-the-art methods of Simard and L'Ecuyer (2011) and Carvalho (2015) as it returns values that are of at least the same precision. In the case when $F(x)$ is mixed, to the best of our knowledge no alternative methods have been proposed in the literature to compute the exact distribution of D_n.

In this paper, we have also derived a useful extension of Schmid (1958)'s asymptotic formula, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. As demonstrated numerically, the extended asymptotic formula provides reasonably close approximations to the exact values of $\mathsf{P}(D_n \geq q)$ and can successfully be used for small to moderate number of jumps in $F(x)$ and large sample sizes.

As part of a separate ongoing research, we have also demonstrated that the FFT-based method can be successfully applied to compute the complementary CDF of the weighted version of the KS test statistic

$$K_n = \sup_x \sqrt{n}|F_n(x) - F(x)|\sqrt{\psi[F(x)]},$$

where $\psi(t) \geq 0, \forall t \in [0,1]$ is a weight function, first considered by Anderson and Darling (1952). The result of this additional research will appear elsewhere. Finally, as noted in Remark A.3, the complementary CDFs $\mathsf{P}(D_n \geq q)$ and $\mathsf{P}(D_n > q)$ are non-increasing functions with jumps at some values of q. Characterizing in detail the distribution of D_n, in particular the points of discontinuity, in relation to $F(x)$ is also a subject of ongoing research.

Acknowledgments

The authors would like to thank Amit Moscovich for his valuable help on running and interpreting the code from Moscovich and Nadler (2017). The third author gratefully acknowledges the financial support received by Cass Business School, City, University of London through a PhD scholarship.

References

Anderson TW, Darling DA (1952). "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes." *The Annals of Mathematical Statistics*, **23**(2), 193–212. doi:10.1214/aoms/1177729437.

Arnold TA, Emerson JW (2011). "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions." *The R Journal*, **3**(2), 34–39. doi:10.32614/rj-2011-016.

Brown JR, Harvey ME (2007). "Rational Arithmetic Mathematica Functions to Evaluate the One-Sided One Sample K-S Cumulative Sampling Distribution." *Journal of Statistical Software*, **19**(6), 1–32. doi:10.18637/jss.v019.i06.

Brown JR, Harvey ME (2008). "Rational Arithmetic Mathematica Functions to Evaluate the Two-Sided One Sample K-S Cumulative Sampling Distribution." *Journal of Statistical Software*, **26**(2), 1–40. doi:10.18637/jss.v026.i02.

Calabrese R, Zenga M (2010). "Bank Loan Recovery Rates: Measuring and Nonparametric Density Estimation." *Journal of Banking and Finance*, **34**(5), 903–911. doi:10.1016/j.jbankfin.2009.10.001.

Carnal H (1962). "Sur Les Théorèmes De Kolmogorov et Smirnov Dans Le Cas D'une Distribution Discontinue." *Commentarii Mathematici Helvetici*, **37**(1), 19–35. doi:10.1007/bf02566958.

Carvalho L (2015). "An Improved Evaluation of Kolmogorov's Distribution." *Journal of Statistical Software*, **65**(3), 1–7. doi:10.18637/jss.v065.c03.

CIESIN (2012). *National Aggregates of Geospatial Data Collection: Population, Landscape, and Climate Estimates, Version 3 (PLACE III)*. Palisades. doi:10.7927/H4F769GP.

Conover WJ (1972). "A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions." *Journal of the American Statistical Association*, **67**(339), 591–596. doi:10.1080/01621459.1972.10481254.

Dimitrova D, Ignatov Z, Kaishev V (2017). "On the First Crossing of Two Boundaries by an Order Statistics Risk Process." *Risks*, **5**(3), 43–56. doi:10.3390/risks5030043.

Dimitrova DS, Kaishev VK, Tan S (2020). ***KSgeneral**: Computing p-Values of the K-S Test for (Dis)Continuous Null Distribution*. R package version 1.0.0, URL https://CRAN.R-project.org/package=KSgeneral.

Doob JL (1949). "Heuristic Approach to the Kolmogorov-Smirnov Theorems." *The Annals of Mathematical Statistics*, **20**(3), 393–403. doi:10.1214/aoms/1177729991.

Durbin J (1968). "The Probability That the Sample Distribution Function Lies between Two Parallel Straight Lines." *The Annals of Mathematical Statistics*, **39**(2), 398–411. doi:10.1214/aoms/1177698404.

Durbin J (1973). *Distribution Theory for Tests Based on the Sample Distribution Theory*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM. doi:10.1137/1.9781611970586.

Eddelbuettel D, François R (2011). "**Rcpp**: Seamless R and C++ Integration." *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08.

Epanechnikov VA (1968). "The Significance Level and Power of the Two-Sided Kolmogorov Test in the Case of Small Sample Sizes." *Theory of Probability and Its Applications*, **13**(4), 686–690. doi:10.1137/1113085.

Feller W (1948). "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions." *The Annals of Mathematical Statistics*, **19**(2), 177–189. doi:10.1214/aoms/1177730243.

Frigo M, Johnson SG (2005). "The Design and Implementation of FFTW3." *Proceedings of the IEEE*, **93**(2), 216–231. doi:10.1109/jproc.2004.840301.

Gleser LJ (1985). "Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions." *Journal of the American Statistical Association*, **80**(392), 954–958. doi:10.1080/01621459.1985.10478210.

IBM Corp (2017). *IBM SPSS Statistics for Windows, Version 25.0*. IBM Corp., Armonk.

Khmaladze E, Shinjikashvili E (2001). "Calculation of Noncrossing Probabilities for Poisson Processes and Its Corollaries." *Advances in Applied Probability*, **33**(3), 702–716. doi:10.1239/aap/1005091361.

Kolmogorov AN (1933). "Sulla Determinazione Empirica Di Una Legge Di Distribuzione." *Giornale dell'Istituto Italiano degli Attuari*, **4**, 83–91. doi:10.1484/j.gif.5.101956.

Marsaglia G, Tsang WW, Wang J (2003). "Evaluating Kolmogorov's Distribution." *Journal of Statistical Software*, **8**(18), 1–4. doi:10.18637/jss.v008.i18.

Massey FJ (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*, **46**(253), 68–78. doi:10.1080/01621459.1951.10500769.

Miller LH (1956). "Table of Percentage Points of Kolmogorov Statistics." *Journal of the American Statistical Association*, **51**(273), 111–121. doi:10.1080/01621459.1956.10501314.

Moscovich A, Nadler B (2017). "Fast Calculation of Boundary Crossing Probabilities for Poisson Processes." *Statistics & Probability Letters*, **123**, 177–182. doi:10.1016/j.spl.2016.11.027.

Niederhausen H (1981). "Sheffer Polynomials for Computing Exact Kolmogorov-Smirnov and Rényi Type Distributions." *The Annals of Statistics*, **9**(5), 923–944. doi:10.1214/aos/1176345574.

Noé M (1972). "The Calculation of Distributions of Two-Sided Kolmogorov-Smirnov Type Statistics." *The Annals of Mathematical Statistics*, **43**(1), 58–64. doi:10.1214/aoms/1177692700.

Noether GE (1963). "Note on the Kolmogorov Statistic in the Discrete Case." *Metrika*, **7**(1), 115–116. doi:10.1007/bf02613966.

Ospina R, Ferrari SLP (2010). "Inflated Beta Distributions." *Statistical Papers*, **51**(1), 111–126. doi:10.1007/s00362-008-0125-4.

Pelz W, Good IJ (1976). "Approximating the Lower Tail-Areas of the Kolmogorov-Smirnov One-Sample Statistic." *Journal of the Royal Statistical Society B*, **38**(2), 152–156. doi:10.1111/j.2517-6161.1976.tb01579.x.

Pettitt AN, Stephens MA (1977). "The Kolmogorov-Smirnov Goodness-of-Fit Statistic with Discrete and Grouped Data." *Technometrics*, **19**(2), 205–210. doi:10.2307/1268631.

Pomeranz J (1974). "Algorithm 487: Exact Cumulative Distribution of the Kolmorgorov-Smirnov Statistic for Small Samples." *Communications of the ACM*, **17**(12), 703–704. doi:10.1145/361604.361628.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ruben H, Gambino J (1982). "The Exact Distribution of Kolmogorov's Statistic D_n for $n \leq 10$." *The Annals of the Institute of Statistical Mathematics*, **34**(1), 167–173. doi:10.1007/bf02481019.

Schmid P (1958). "On the Kolmogorov and Smirnov Limit Theorems for Discontinuous Distribution Functions." *The Annals of Mathematical Statistics*, **29**(4), 1011–1027. doi:10.1214/aoms/1177706438.

Shorack GR, Wellner JA (2009). *Empirical Processes with Applications to Statistics*. SIAM, Philadelphia. doi:10.1137/1.9780898719017.

Simard R, L'Ecuyer P (2011). "Computing the Two-Sided Kolmogorov-Smirnov Distribution." *Journal of Statistical Software*, **39**(11), 1–18. doi:10.18637/jss.v039.i11.

Slakter MJ (1965). "A Comparison of the Pearson Chi-Square and Kolmogorov Goodness-of-Fit Tests with Respect to Validity." *Journal of the American Statistical Association*, **60**(311), 854–858. doi:10.2307/2283251.

Smirnov N (1939). "Sur Les écarts De La Courbe De Distribution Empirique." *Matematicheskii Sbornik*, **48**(1), 3–26. doi:10.1787/888932596384.

Smirnov N (1948). "Table for Estimating the Goodness of Fit of Empirical Distributions." *The Annals of Mathematical Statistics*, **19**(2), 279–281. doi:10.1214/aoms/1177730256.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC., College Station.

Steck GP (1971). "Rectangle Probabilities for Uniform Order Statistics and the Probability That the Empirical Distribution Function Lies between Two Distribution Functions." *The Annals of Mathematical Statistics*, **42**(1), 1–11. doi:10.1214/aoms/1177693490.

The MathWorks Inc (2018). *MATLAB – The Language of Technical Computing, Version R2018a*. The MathWorks Inc., Natick.

Walsh JE (1963). "Bounded Probability Properties of Kolmogorov-Smirnov and Similar Statistics for Discrete Data." *The Annals of the Institute of Statistical Mathematics*, **15**(1), 153–158. doi:10.1007/bf02865912.

Wolfram Research, Inc (2018). *Mathematica, Version 11.3*. Wolfram Research, Inc., Champaign.

Wood CL, Altavela MM (1978). "Large-Sample Results for Kolmogorov–Smirnov Statistics for Discrete Distributions." *Biometrika*, **65**(1), 235–239. doi:10.1093/biomet/65.1.235.

A. Expressing complementary CDFs of D_n

In this appendix, we express $\mathsf{P}(D_n > q)$ and $\mathsf{P}(D_n \geq q)$ in terms of a rectangle probability with respect to the uniform order statistics.

Lemma A.1. *The following holds true*

$$\mathsf{P}(D_n > q) = 1 - \mathsf{P}(\tilde{A}_i \leq U_{(i)} \leq \tilde{B}_i, 1 \leq i \leq n),$$

where $\tilde{A}_i = F\Big((F^{-1}(\frac{i}{n} - q))-\Big)$ *and* $\tilde{B}_i = F\Big(F^{-1}((\frac{i-1}{n} + q)+)\Big)$ *and* $F^{-1}(y+) = \lim_{\epsilon \downarrow 0} F^{-1}(y+\epsilon)$.

Proof: We have

$$\begin{aligned}
\mathsf{P}(D_n > q) &= \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| > q) \\
&= 1 - \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \leq q) \\
&= 1 - \mathsf{P}(|F_n(x) - F(x)| \leq q, \text{ for all } x) \\
&= 1 - \mathsf{P}(-q \leq F_n(x) - F(x) \leq q, \text{ for all } x) \\
&= 1 - \mathsf{P}(F(x) - q \leq F_n(x) \leq F(x) + q, \text{ for all } x) \\
&= 1 - \mathsf{P}\Big(F(X_{(i)}-) - q \leq F_n(X_{(i-1)}) \text{ and } F_n(X_{(i)}) \leq F(X_{(i)}) + q, \text{ for } 1 \leq i \leq n\Big) \\
&= 1 - \mathsf{P}\Big(F(X_{(i)}-) \leq \frac{i-1}{n} + q \text{ and } \frac{i}{n} - q \leq F(X_{(i)}), \text{ for } 1 \leq i \leq n\Big) \\
&= \mathsf{P}\Big(F(X_{(i)}-) > \frac{i-1}{n} + q \text{ or } \frac{i}{n} - q > F(X_{(i)}), \text{ for some } 1 \leq i \leq n\Big) \\
&= \mathsf{P}\left(F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) + \Big) < X_{(i)} \text{ or } F^{-1}\Big(\frac{i}{n} - q\Big) > X_{(i)} \text{ for some } 1 \leq i \leq n\right),
\end{aligned}$$

where in the last equality we have applied that $u < F(x-)$ if and only if $F^{-1}(u+) < x$ and that $x < F^{-1}(u)$ if and only if $F(x) < u$ (see, e.g., Lemma 1 (iii) and (v) of Gleser (1985)). Therefore, we now have

$$\begin{aligned}
\mathsf{P}(D_n > q) &= 1 - \mathsf{P}\left(F^{-1}\Big(\frac{i}{n} - q\Big) \leq X_{(i)} \leq F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) + \Big) \text{ for } 1 \leq i \leq n\right) \\
&= 1 - \mathsf{P}\left(F\Big(\Big(F^{-1}\Big(\frac{i}{n} - q\Big)\Big) - \Big) \leq U_{(i)} \leq F\Big(F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) + \Big)\Big) \text{ for } 1 \leq i \leq n\right),
\end{aligned} \tag{28}$$

where in the last equality we have applied Lemma 1 of Dimitrova, Ignatov, and Kaishev (2017). The statement now follows noting that one can rewrite the last equality in terms of $\tilde{A}_i$ and $\tilde{B}_i$. □

Remark A.2. The fact that the non-crossing probability

$$\begin{aligned}
&\mathsf{P}(F(x) - q \leq F_n(x) \leq F(x) + q, \text{ for all } x) \\
&\qquad = \mathsf{P}\left(F^{-1}\Big(\frac{i}{n} - q\Big) \leq X_{(i)} \leq F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) + \Big) \text{ for } 1 \leq i \leq n\right)
\end{aligned}$$

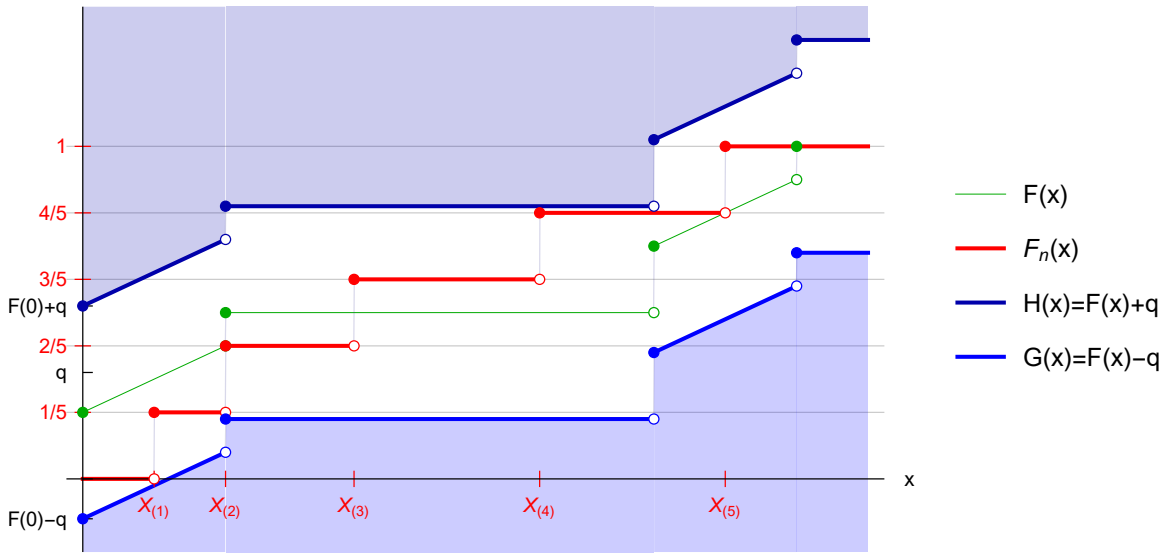

Figure 3: Illustration of the equivalence of $\mathsf{P}(F(x) - q \le F_n(x) \le F(x) + q, \text{ for all } x)$ to $\mathsf{P}\Big(F^{-1}(\frac{i}{n} - q) \le X_{(i)} \le F^{-1}((\frac{i-1}{n} + q) +) \text{ for } 1 \le i \le n\Big)$ (cf., Remark A.2), for $F(x)$ defined as in (23) with $n = 5$.

shown in the proof of Lemma A.1 is illustrated in Figure 3 with $F(x)$ (the green piecewise linear function) defined in (23) (cf., Example 2.8), for $n = 5$.

Remark A.3. The statement of Lemma A.1 holds true also for $\mathsf{P}(D_n \ge q)$, as stated in (3), with A_i and B_i defined as in (4). The proof is similar but more involved than that of Lemma A.1 and is therefore omitted. It should also be noted that the complementary CDFs $\mathsf{P}(D_n \ge q)$ and $\mathsf{P}(D_n > q)$ are non-increasing functions with jumps at some values of q. In fact, these two functions coincide, except at the jumps where $\mathsf{P}(D_n \ge q)$ is left-continuous and $\mathsf{P}(D_n > q)$ is right-continuous. This is a consequence of the fact that the pairs A_i, B_i and $\tilde{A}_i$, $\tilde{B}_i$ coincide except at their points of discontinuity, where A_i, B_i are correspondingly right- and left- continuous, whereas $\tilde{A}_i$, $\tilde{B}_i$ are correspondingly left- and right-continuous.

Remark A.4. Let us note that the result of Lemma A.1 coincides with Theorem 1 of Gleser (1985).

B. Computing the CDF of D_n for continuous $F(x)$

In this appendix, we compute the values of the CDF $\mathsf{P}(D_n \le q)$ for different n and q using the exact-KS-FFT method and compare the results to those obtained with the C program due to Simard and L'Ecuyer (2011) and R function `pkolmim` from the package **kolmim** by Carvalho (2015), which is claimed to be highly efficient and precise. Hence, we calculate an absolute error as the absolute difference between our results and the R outputs, from which we can infer the number of decimal digits of precision of our results.

In order to compute $\mathsf{P}(D_n \le q)$, when $F(x)$ is continuous using the R package **KSgeneral**, one needs to input `cont_ks_cdf(q, n)`. For example, in order to compute the value for

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	2.320196159531E-08	2.320196159531E-08	2.320196159531E-08	1.1250E-22
40	6.749093037884E-17	6.749093037884E-17	6.749093037884E-17	1.0847E-30
60	1.702549809333E-25	1.702549809333E-25	1.702549809333E-25	2.2500E-39
80	4.050687717856E-34	4.050687717855E-34	4.050687717855E-34	3.0705E-47
100	9.332621544394E-43	9.332621544394E-43	9.332621544394E-43	3.7597E-56
120	2.106901932614E-51	2.106901932614E-51	2.106901932614E-51	2.4154E-64
140	4.690131222299E-60	4.690131222299E-60	4.690131222299E-60	9.1198E-74

Table 8: Values of $\mathsf{P}(D_n \leq q)$ for $q = 1/n$.

$\mathsf{P}(D_n \leq q)$, when $F(x)$ is continuous, for $n = 40$, $nq^2 = 0.76$, one should run the following R code and obtain the corresponding result as shown in Table 10 for $n = 40$ in the column exact-KS-FFT.

```
R> cont_ks_cdf(sqrt(0.76/40), 40)

[1] 0.6032371
```

Simard and L'Ecuyer (2011) consider the following regions: 1) $n \leq 140$ and $q \leq 1/n$; 2) $n \leq 140$ and $q \geq 1 - 1/n$; 3) $n \leq 140$ and $1/n < nq^2 < 0.754693$; 4) $n \leq 140$ and $0.754693 \leq nq^2 < 4$; 5) $n \leq 140$ and $4 \leq nq^2 < 18$; 6) $n \leq 140$ and $nq^2 \geq 18$; 7) $140 < n \leq 10^5$ and $nq^{3/2} < 1.4$; 8) $140 < n \leq 10^5$ and $nq^{3/2} \geq 1.4$; and 9) $n > 10^5$ where they use different methods to compute the distribution of D_n (cf., Simard and L'Ecuyer 2011, Section 4).

Following the segmentation of regions, we have computed the distribution of D_n with the proposed FFT-based method and can report that for regions 1), 2), 3), 4), 7), our approach gives results that are of at least the same precision as those obtained from the R or C program. In regions 5) and 6), when $n \leq 140$ and $nq^2 > 12$, our approach may be unsuitable due to numerical instabilities which may occur.

More specifically, when 1) $n \leq 140$ and $q \leq 1/n$, or when 2) $n \leq 140$ and $q \geq 1 - 1/n$, Simard and L'Ecuyer (2011) use the Ruben and Gambino (1982) formula to calculate the distribution of D_n, returning results with at least 13 decimal digits of precision. As can be seen from Table 8, in these regions our method gives results that are of similar accuracy as those from the R function `pkolmim` or the C program of Simard and L'Ecuyer (2011).

When 3) $n \leq 140$ and $1/n < nq^2 < 0.754693$, Simard and L'Ecuyer (2011) use the Durbin matrix algorithm to calculate the distribution of D_n, returning results with at least 13 decimal digits of precision. As can be seen from Table 9, in this region our method gives results of at least the same accuracy.

When 4) $n \leq 140$ and $0.754693 \leq nq^2 < 4$, Simard and L'Ecuyer (2011) use the Pomeranz (1974) method to calculate the distribution of D_n, returning results with at least 13 decimal digits of precision. In this region, again our method gives results of at least the same accuracy as shown in Tables 10 and 11 for $nq^2 = 0.76$ and $nq^2 = 3.9$, respectively.

When 5) $n \leq 140$ and $4 \leq nq^2 < 18$, Simard and L'Ecuyer (2011) first use the Miller (1956) approximation to estimate $\mathsf{P}(D_n \geq q)$, and then calculate the distribution of D_n by $\mathsf{P}(D_n \leq q) = 1 - \mathsf{P}(D_n \geq q)$. The authors claim that the approximated values of $\mathsf{P}(D_n \leq q)$ have 14 decimal digits of precision. As illustrated in Tables 12 and 13 for $nq^2 = 4.1$ and $nq^2 = 12$,

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.6089841201379	0.6089841201379	0.6089841201379	3.2196E-15
40	0.5951497241008	0.5951497241008	0.5951497241008	1.1102E-15
60	0.5888010590107	0.5888010590107	0.5888010590107	5.6621E-15
80	0.5849488429478	0.5849488429478	0.5849488429478	3.5971E-14
100	0.5822897960079	0.5822897960080	0.5822897960080	2.2649E-14
120	0.5803108927579	0.5803108927579	0.5803108927579	6.5614E-14
140	0.5787632928760	0.5787632928760	0.5787632928760	4.9960E-14

Table 9: Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 0.75$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.6169412955835	0.6169412955835	0.6169412955835	8.8818E-16
40	0.6032370735674	0.6032370735674	0.6032370735674	7.2164E-15
60	0.5969494784897	0.5969494784898	0.5969494784897	7.5495E-15
80	0.5931349807275	0.5931349807274	0.5931349807274	5.0737E-14
100	0.5905022875562	0.5905022875562	0.5905022875562	3.2641E-14
120	0.5885431553286	0.5885431553286	0.5885431553285	6.1395E-14
140	0.5870111081551	0.5870111081552	0.5870111081551	2.4425E-15

Table 10: Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 0.76$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.9995468293485	0.9995468293485	0.9995468293485	4.3299E-15
40	0.9994205337332	0.9994205337332	0.9994205337332	1.4655E-14
60	0.9993680770022	0.9993680770022	0.9993680770022	1.4544E-14
80	0.9993382289964	0.9993382289964	0.9993382289964	7.4607E-14
100	0.9993185558110	0.9993185558111	0.9993185558110	4.4076E-14
120	0.9993044245858	0.9993044245858	0.9993044245857	1.0636E-13
140	0.9992936831012	0.9992936831012	0.9992936831012	9.9920E-15

Table 11: Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 3.9$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.9997098154630	0.9997098154630	0.9997098154630	3.9968E-15
40	0.9996202540524	0.9996202540523	0.9996202540523	1.3767E-14
60	0.9995829210883	0.9995829210883	0.9995829210883	1.0325E-14
80	0.9995616853087	0.9995616853087	0.9995616853087	7.3608E-14
100	0.9995477016848	0.9995477016848	0.9995477016848	4.5519E-14
120	0.9995376676397	0.9995376676396	0.9995376676396	1.0769E-13
140	0.9995300481355	0.9995300481355	0.9995300481355	1.3989E-14

Table 12: Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 4.1$.

our method gives results of at least the same accuracy when $n \leq 140$ and $4 \leq nq^2 \leq 12$. For $n \leq 140$ and $12 < nq^2 < 18$, since our implementation uses floating numbers in C++, numerical instabilities may occur.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.99999999999963	0.99999999999962	0.99999999999962	7.3275E-15
40	0.99999999999135	0.99999999999134	0.99999999999134	1.4877E-14
60	0.99999999998168	0.99999999998167	0.99999999998167	1.1102E-14
80	0.99999999997414	0.99999999997407	0.99999999997407	7.3053E-14
100	0.99999999996822	0.99999999996827	0.99999999996827	4.5297E-14
120	0.99999999996387	0.99999999996376	0.99999999996376	1.1113E-13
140	0.99999999996019	0.99999999996017	0.99999999996017	2.0872E-14

Table 13: Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 12$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
140	6.378698330644E-02	6.378698330644E-02	6.378698330644E-02	7.4940E-16
500	7.365490405432E-03	7.365490405433E-03	7.365490405433E-03	4.4409E-16
1000	1.383862966202E-03	1.383862966202E-03	1.383862966202E-03	1.4051E-16
5000	3.811342214262E-06	3.811342214277E-06	3.811342214276E-06	1.4021E-17
10000	8.999089573398E-08	8.999089573405E-08	8.999089573401E-08	2.3810E-20
50000	1.979970859427E-13	1.979970859405E-13	1.979970859397E-13	3.0054E-24
100000	5.388085736386E-17	5.388085736389E-17	5.388085736345E-17	4.0889E-28

Table 14: Values of $\mathsf{P}(D_n \leq q)$ for $nq^{3/2} = 1.3$.

When 6) $n \leq 140$ and $nq^2 \geq 18$, $\mathsf{P}(D_n \geq q) < 5 \times 10^{-16}$. Equivalently, $\mathsf{P}(D_n \leq q) = 1 - \mathsf{P}(D_n \geq q) > 1 - 5 \times 10^{-16}$. Hence, returning $\mathsf{P}(D_n \leq q) = 1$ will give results with 15 decimal digits of precision.

When 7) $140 < n \leq 10^5$ and $nq^{3/2} < 1.4$, Simard and L'Ecuyer (2011) use the Durbin matrix algorithm to obtain the exact distribution of D_n, returning probabilities with at least 13 decimal digits of precision. As illustrated in Table 14, our method returns values of at least the same accuracy.

In region 8), when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 \leq 18$, Simard and L'Ecuyer (2011) apply the Pelz and Good (1976) approximation that gives five decimal digits of precision for values of $\mathsf{P}(D_n \leq q)$. In contrast, when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 \leq 10$, our approach gives results with at least 11 decimal digits of precision even though it is using floating numbers in calculation. The results when $nq^{3/2} = 1.4$ and when $nq^2 = 10$ are shown in Tables 15 and 16, respectively. However, in region 8), when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 > 10$, our approach may be unsuitable due to numerical instabilities. In particular, it will return results with at least 11 decimal digits of precision, but the resulting values of $\mathsf{P}(D_n \leq q)$ may not be decreasing in n, due to the errors in calculations with floating numbers. When $140 < n \leq 10^5$ and $nq^2 \geq 18$, returning $\mathsf{P}(D_n \leq q) = 1$ will give results with 15 decimal digits of precision.

Finally, in region 9), Simard and L'Ecuyer (2011) apply the Pelz and Good (1976) approximation to obtain values of $\mathsf{P}(D_n \leq q)$ when $nq^2 < 18$, and set $\mathsf{P}(D_n \leq q) = 1$ when $nq^2 \geq 18$. As illustrated in Table 17 for $n = 100001$, our approach tends to be more accurate when $\mathsf{P}(D_n \leq q)$ is very small. However, Pelz and Good (1976) approximation may provide higher accuracy when $\mathsf{P}(D_n \leq q)$ tends to one.

To conclude, apart from the regions where $n \leq 140$ and $12 < nq^2 < 18$; or $140 < n \leq 10^5$,

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
140	9.026232947500E-02	9.026232947500E-02	9.026232947500E-02	7.4940E-16
500	1.302425400211E-02	1.302425400211E-02	1.302425400211E-02	6.9215E-16
1000	2.894937251699E-03	2.894937251698E-03	2.894937251698E-03	4.2284E-16
5000	1.423550831459E-05	1.423550831465E-05	1.423550831464E-05	5.6458E-17
10000	4.833454107672E-07	4.833454107673E-07	4.833454107671E-07	1.5014E-19
50000	3.714790944124E-12	3.714790944070E-12	3.714790944055E-12	6.8725E-23
100000	2.212360525520E-15	2.212360525494E-15	2.212360525477E-15	4.3390E-26

Table 15: Values of $\mathsf{P}(D_n \le q)$ for $nq^{3/2} = 1.4$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.9999999974397	0.9999999974396	0.9999999974396	3.8969E-14
500	0.9999999965419	0.9999999965420	0.9999999965420	9.8810E-14
1000	0.9999999962965	0.9999999962983	0.9999999962963	1.5710E-13
5000	0.9999999960263	0.9999999960309	0.9999999960307	4.4144E-12
10000	0.9999999959799	0.9999999959799	0.9999999959798	8.3489E-14
50000	0.9999999959304	0.9999999959196	0.9999999959197	1.0697E-11
100000	0.9999999959211	0.9999999959067	0.9999999959068	1.4232E-11

Table 16: Values of $\mathsf{P}(D_n \le q)$ for $nq^2 = 10$.

q	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
$\frac{1}{10\sqrt{n}}$	2.350089150932E-52	2.269812366890E-52	2.350089151281E-52	3.4925E-62
$\frac{1}{8\sqrt{n}}$	1.969026572909E-33	1.962478061021E-33	1.969026573193E-33	2.8391E-43
$\frac{1}{6\sqrt{n}}$	1.018454527587E-18	1.018350562852E-18	1.018454527742E-18	1.5588E-28
$\frac{1}{4\sqrt{n}}$	2.907074248661E-08	2.907073793359E-08	2.907074249157E-08	4.9669E-18
$\frac{1}{2\sqrt{n}}$	3.639199759646E-02	3.639199759703E-02	3.639199760172E-02	5.2639E-12
$\frac{1}{\sqrt{n}}$	7.305646846155E-01	7.305646847185E-01	7.305646847159E-01	1.0047E-10
$\frac{2}{\sqrt{n}}$	9.993319331628E-01	9.993319333086E-01	9.993319333086E-01	1.4579E-10

Table 17: Values of $\mathsf{P}(D_n \le q)$ for $n = 100001$.

$nq^{3/2} \ge 1.4$, and $10 < nq^2 < 18$; or $nq^2 \ge 18$, the exact-KS-FFT method returns values of $\mathsf{P}(D_n \le q)$ that are at least as accurate as those obtained by Simard and L'Ecuyer (2011). This is shown in Figure 4. Moreover, for $n > 10^5$, the proposed method may be accurate when $\mathsf{P}(D_n \le q)$ is very small.

C. Computing the complementary CDF for continuous $F(x)$

It is well known that

$$D_n = \sup_x |F_n(x) - F(x)| \longrightarrow 0 \quad a.s.,$$

as $n \to \infty$. Hence, when n is very large, $\mathsf{P}(D_n \le q)$ is close to one. Also, it can be seen that $D_n \in [0, 1]$, so $\mathsf{P}(D_n \le q)$ is close to one when q is close to one. In these cases, cancellation

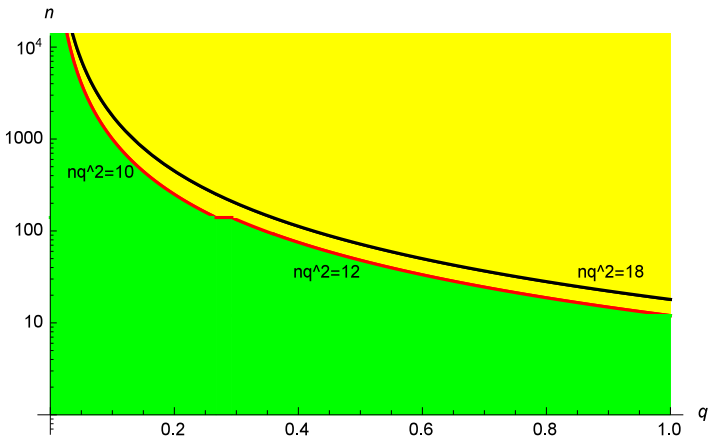

Figure 4: Approximate regions where the exact-KS-FFT method returns $\mathsf{P}(D_n \leq q)$ efficiently and accurately.

errors may occur when trying to numerically compute the p value

$$\mathsf{P}(D_n \geq q) = 1 - \mathsf{P}(D_n \leq q). \tag{29}$$

Similarly to previous section, we compute the values of $\mathsf{P}(D_n \geq q)$ for different n and q using the exact-KS-FFT method and compare the results to those obtained with the R program of Carvalho (2015), and the C program due to Simard and L'Ecuyer (2011).

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is continuous using the R package **KSgeneral**, one needs to input `cont_ks_c_cdf(q, n)`. For instance, in order to compute the value for $\mathsf{P}(D_n \geq q)$, for $n = 141$, $nq^2 = 2.1$, one should run the following R code and obtain the corresponding result as shown in Table 19 for $n = 141$ in the column exact-KS-FFT.

```
R> cont_ks_c_cdf(sqrt(2.1/141), 141)

[1] 0.02743689
```

Simard and L'Ecuyer (2011) consider the following regions: 1) $n \leq 140$ and $nq^2 < 4$; 2) $n \leq 140$ and $nq^2 \geq 4$; 3) $n > 140$ and $nq^2 < 2.2$; and 4) $n > 140$ and $nq^2 \geq 2.2$ where they use different methods to compute the complementary CDF of D_n (cf., Simard and L'Ecuyer 2011, Section 5).

Following the segmentation of regions, we have computed the complementary CDF of D_n with the proposed FFT-based method. Consequently, we can report that for region 1), our approach gives results that are of at least the same accuracy as those obtained from the R or C program. In region 2), $\mathsf{P}(D_n \leq q)$ is close to one and our method may be unsuitable due to cancellation errors which may occur when calculating the complementary CDF via (29). A comparison for $nq^2 = 4$ is shown in Table 18.

In region 3), when $140 < n \leq 10^5$ and $nq^2 < 2.2$, Simard and L'Ecuyer (2011) use the Pelz and Good (1976) approximation and apply (29) to calculate the complementary CDF, returning results with at least five decimal digits of precision. Our approach also applies (29), but returns results with at least nine decimal digits of precision. A comparison for $nq^2 = 2.1$ is given in Table 19.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	3.627396978E-04	3.627396978E-04	3.627396978E-04	3.8858E-15
40	4.691487961E-04	4.691487961E-04	4.691487961E-04	1.3434E-14
60	5.134182982E-04	5.134182982E-04	5.134182982E-04	1.0325E-14
80	5.386021476E-04	5.386021476E-04	5.386021476E-04	7.1498E-14
100	5.551927328E-04	5.551927328E-04	5.551927328E-04	4.3299E-14
120	5.671032850E-04	5.671032851E-04	5.671032851E-04	1.0847E-13
140	5.761521040E-04	5.761521040E-04	5.761521040E-04	7.5495E-15

Table 18: Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 4$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.02743688914	0.02743688914	0.02743688914	3.5305E-14
500	0.02866250073	0.02866250073	0.02866250073	4.2966E-14
1000	0.02905830828	0.02905831976	0.02905830828	8.5154E-14
10000	0.02969964418	0.02969964429	0.02969964418	6.5747E-13
100000	0.02989926163	0.02989926162	0.02989926162	1.1226E-11

Table 19: Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 2.1$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.02239633302	0.02239633302	0.02239633302	4.4409E-14
500	0.02343606481	0.02343606481	0.02343606481	4.9516E-14
1000	0.02377033994	0.02377035110	0.02377033994	1.9440E-13
10000	0.02431016269	0.02431016281	0.02431016270	1.0915E-12
100000	0.02447768610	0.02447768610	0.02447768610	7.2897E-13

Table 20: Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 2.2$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	1.2484863E-06	1.2484863E-06	1.2484863E-06	3.8969E-14
500	1.4796907E-06	1.4796906E-06	1.4796907E-06	8.7708E-15
1000	1.5434600E-06	1.5434599E-06	1.5434600E-06	7.7716E-14
10000	1.6309285E-06	1.6309265E-06	1.6309266E-06	1.8935E-12
100000	1.6535082E-06	1.6534983E-06	1.6534982E-06	1.0010E-11

Table 21: Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 7$.

In region 4), when $140 < n \leq 10^5$ and $nq^2 \geq 2.2$, Simard and L'Ecuyer (2011) use the Miller (1956) approximation and obtain complementary CDF with at least six decimal digits of precision. In this region, the proposed FFT-based method may give more accurate results when $140 < n \leq 10^5$ and $2.2 \leq nq^2 \leq 7$. For example, for $nq^2 = 2.2$ and $nq^2 = 7$, Tables 20 and 21 show that the exact-KS-FFT method returns complementary CDF with at least 10 decimal digits of precision. When $140 < n \leq 10^5$ and $nq^2 > 7$, our method may be unsuitable due to cancellation errors as previously discussed.

Finally, when $n > 10^5$ and $nq^2 < 370$, Simard and L'Ecuyer (2011) use the Miller (1956) approximation and obtain complementary CDF with a few correct decimal digits. These

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
100001	4.939303426E-03	4.939303336E-03	4.939303263E-03	1.6332E-10
200000	4.944654924E-03	4.944654660E-03	4.944654584E-03	3.3974E-10
300000	4.947020041E-03	4.947020013E-03	4.947019947E-03	9.3853E-11

Table 22: Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 3$.

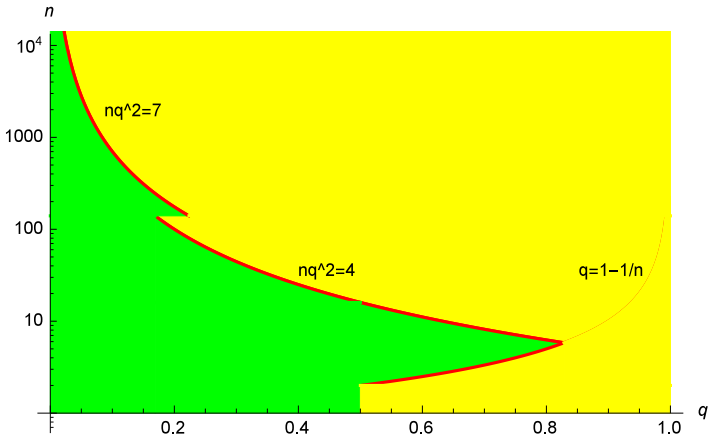

Figure 5: Approximate regions where the exact-KS-FFT method returns $\mathsf{P}(D_n \geq q)$ efficiently and accurately.

authors have shown that complementary CDF can be set to be zero when $nq^2 \geq 370$. Recall that in Table 17, we have shown that the exact-KS-FFT method tends to be more accurate when $\mathsf{P}(D_n \leq q)$ is very small, or when q is small. In this case, we can apply (29) to calculate the complementary CDF, without incurring large cancellation errors. More specifically, when $n > 10^5$ and $nq^2 \leq 3$, the exact-KS-FFT method returns complementary CDF with at least seven decimal digits of precision as demonstrated in Table 22. The accuracy of course deteriorates when $n > 10^5$ and $3 < nq^2 < 370$.

To summarize, apart from the regions where $n \leq 140$ and $nq^2 \geq 4$; or $n \leq 140$ and $q \geq 1-1/n$; or $140 < n \leq 10^5$ and $nq^2 > 7$, the exact-KS-FFT method returns values of the probability $\mathsf{P}(D_n \geq q)$ that are at least as accurate as those obtained by Simard and L'Ecuyer (2011). This is shown in Figure 5. Moreover, when $n > 10^5$ and $nq^2 \leq 3$, the proposed approach may be more accurate than Simard and L'Ecuyer (2011) method.

D. Speed comparison

Tables 23, 24 and 25 report the CPU times to compute $\mathsf{P}(D_n \geq q)$ 100 times, for selected values of n and $\lambda = qn^{1/2}$. As expected, Simard and L'Ecuyer (2011) C program which combines the most efficient methods for computing the distribution of D_n for $F(x)$ continuous, is the fastest among the three procedures. However, the exact-KS-FFT method proves to be a viable alternative especially given its generality and applicability to the case of discontinuous $F(x)$.

$n\backslash\lambda$	0.25	0.5	1	2	3	4
10	0.00003	0.00016	0.00031	0.00004	0.00001	0.00000
100	0.00030	0.00125	0.00463	0.00035	0.00029	0.00024
140	0.00029	0.00123	0.00851	0.01715	0.00032	0.00026
141	0.00024	0.00099	0.00602	0.00040	0.00032	0.00027
1000	0.00336	0.02888	0.00002	0.00269	0.00319	0.00228
10000	0.16970	0.00001	0.00002	0.02691	0.02385	0.01999
100000	8.35199	0.00002	0.00004	0.28366	0.23891	0.20484

Table 23: CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the Simard and L'Ecuyer (2011) C program.

$n\backslash\lambda$	0.25	0.5	1	2	3	4
10	0.065	0.080	0.074	0.058	0.065	0.006
100	0.212	0.183	0.199	0.203	0.211	0.000
140	0.307	0.292	0.313	0.302	0.409	0.001
141	0.347	0.335	0.354	0.368	0.402	0.001
1000	2.483	2.337	2.650	3.006	3.704	0.001
10000	82.654	88.667	96.662	114.566	122.466	0.001
100000	8294.874	8316.082	8646.217	8934.526	9551.063	0.001

Table 24: CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the exact-KS-FFT method.

$n\backslash\lambda$	0.25	0.5	1	2	3	4
10	0.002	0.002	0.001	0.001	0.002	0.001
100	0.002	0.002	0.004	0.009	0.015	0.027
140	0.002	0.002	0.004	0.013	0.029	0.053
141	0.001	0.002	0.004	0.014	0.028	0.053
1000	0.013	0.043	0.166	0.740	2.472	5.872
10000	1.053	4.075	31.292	152.414	316.529	523.194
100000	135.772	965.544	3462.484	11951.990	24563.235	41220.483

Table 25: CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the Carvalho (2015) R program.

Affiliation:

Dimitrina S. Dimitrova, Vladimir K. Kaishev, Senren Tan
Faculty of Actuarial Science and Insurance
Cass Business School
City, University of London
106 Bunhill Row, EC1Y 8TZ London, United Kingdom
E-mail: D.Dimitrova@city.ac.uk, v.kaishev@city.ac.uk, Senren.Tan@cass.city.ac.uk
URL: http://www.cass.city.ac.uk/experts/D.Dimitrova,
http://www.cass.city.ac.uk/experts/V.Kaishev

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2020, Volume 95, Issue 10 *Submitted:* 2016-10-20
doi:10.18637/jss.v095.i10 *Accepted:* 2018-07-13