

# EXACT NULL DISTRIBUTIONS OF THE OVL- $q$ GOODNESS-OF-FIT TESTS: A LINEAR-TIME ALGORITHM FOR $D_q$ , AND THE CELLS WITH $q \geq 3$

KOREA SUPERINTELLIGENCE LABS

**ABSTRACT.** Komaba, Johnno and Nakamoto attach to a pair of distribution functions  $F, G$  on  $\mathbb{R}$  and an integer  $q \geq 1$  the overlap discrepancy  $D_q(F, G) = 1 - \inf_{v_1 \leq \dots \leq v_q} \sum_{i=0}^q \min(F|_{v_i}^{v_{i+1}}, G|_{v_i}^{v_{i+1}})$ , and build from it a one-sample and a two-sample goodness-of-fit test (the OVL- $q$  tests;  $D_1$  is the Kolmogorov–Smirnov metric). Both papers close by recording that they cannot evaluate the statistic outside the case  $q = 2$  with equal sample sizes: “we are currently unable to rapidly perform the OVL-2 if  $m \neq n$  or the OVL- $q$  if  $q \geq 3$ ”, and “we have not presented a calculation method for the one-sample OVL- $q$  test when  $q > 2$ ”. Their own algorithm is a minimisation over all  $\binom{N+q}{q}$  chains of breakpoints. We give an  $O(qN)$  dynamic programme for  $D_q$ , obtained from those papers’ own description of  $D_q$  as half of a  $q$ -point zigzag total variation of  $F - G$ , and verify it against their definition word by word on  $\Gamma_{3,4}$  and  $\Gamma_{4,5}$ . With it we compute exact null distributions of the two-sample OVL- $q$  statistic for  $q \geq 3$ , which we have not been able to find anywhere in print: OVL-3 at sample sizes  $(4, 4)$ ,  $(5, 5)$ ,  $(6, 6)$ ,  $(4, 5)$ ,  $(5, 6)$  and OVL-4 at  $(6, 6)$ . The OVL-3 test is not a relabelling of OVL-2: on the 462 interleavings of a  $(5, 6)$  sample  $D_3 > D_2$  on exactly 350. We also record, as background rather than as a new result, that for  $q = 2$  the statistic is Kuiper’s  $V_n$ , so that the  $q = 2$  half of both stated obstacles is already answered by a literature the two papers do not cite. Every theorem and proposition below is checked by the Lean 4 kernel over Mathlib.

## 1. INTRODUCTION

**The statistic.** Write  $\mathcal{F}$  for the set of distribution functions on  $\mathbb{R}$  and  $F|_a^b = F(b) - F(a)$ , with the conventions  $F(-\infty) = 0$ ,  $F(+\infty) = 1$ . For  $F, G \in \mathcal{F}$ , an integer  $q \geq 1$  and a nondecreasing tuple  $v = (v_1 \leq \dots \leq v_q) \in \mathbb{R}^q$ , put  $v_0 = -\infty$ ,  $v_{q+1} = +\infty$  and

$$(1) \quad r_{F,G}(v) = \sum_{i=0}^q \min(F|_{v_i}^{v_{i+1}}, G|_{v_i}^{v_{i+1}}), \quad D_q(F, G) = 1 - \inf_{v_1 \leq \dots \leq v_q} r_{F,G}(v).$$

This is the definition of Komaba, Johnno and Nakamoto [5, 4]. The quantity  $r_{F,G}(v)$  is the mass the two distributions share when  $\mathbb{R}$  is cut into the  $q + 1$  intervals determined by  $v$ , so  $1 - D_q$  is an overlap coefficient computed at the best cut, and  $D_q$  is a discrepancy:  $D_1$  is the Kolmogorov–Smirnov metric, and  $D_q$  is nondecreasing in  $q$  because a chain of  $q$  points extends to a chain of  $q + 1$  points, repetitions being allowed, without changing the value of  $r_{F,G}$ .

From  $D_q$  the two papers build two tests. The *two-sample OVL- $q$  test* of [4] uses  $\rho_{q,m,n} = 1 - D_q(F_{0,m}, F_{1,n})$ , where  $F_{0,m}$  and  $F_{1,n}$  are the empirical distribution functions of independent samples of sizes  $m$  and  $n$ ; large  $D_q$  is evidence against the hypothesis that the two samples come from the same continuous law. The *one-sample OVL- $q$  test* of [5] uses  $D_q(F_n, F)$  for an empirical distribution function  $F_n$  of size  $n$  and a continuous reference  $F$ , with  $p$ -value  $p_{q,n}(x) = \mathbb{P}(x \leq D_q(F_n, F))$  and upper confidence limit  $u_{q,n}(\alpha) = \inf\{x : p_{q,n}(x) < \alpha\}$ .

**What those papers could not compute.** Both papers end on the same obstacle. The two-sample paper [4, Conclusion] writes, verbatim:

“One limitation is that we are currently unable to rapidly perform the OVL-2 if  $m \neq n$  or the OVL- $q$  if  $q \geq 3$ . To overcome this, we should explore the possibility of expanding fast and exact algorithms for the OVL- $q$ , or should investigate the

asymptotic distribution of  $\rho_{q,m,n}$  (as  $m, n \rightarrow \infty$ ) to approximate the OVL- $q$  in future works.”

The one-sample paper [5] reports the same gap on the other side. Its experiments say

“Beforehand, we computed  $p_{2,n}$ , as defined in (4), using the Monte Carlo method for  $n = 2^3, 2^4, \dots, 2^{12}$ , because exact  $p$ -values for the one-sample OVL-2 test could not be computed”,

where the cross-reference is to that paper’s display (4) defining  $p_{q,n}$ , quoted above, and its Conclusion lists as limitations “We have not presented a calculation method for the one-sample OVL- $q$  test when  $q > 2$ ” and “We have not proposed a method for calculating  $p$ -values when the sample size is small, particularly exact methods”.

The reason is visible in (1): as written,  $D_q$  is an infimum over a  $q$ -dimensional region. Reduced to the finitely many breakpoints of  $F$  and  $G$  it becomes a minimisation over all  $\binom{N+q}{q}$  nondecreasing index chains, where  $N$  is the number of breakpoints, and this — together with a generating function that handles  $q = 2$  with equal sample sizes [4, Thm. 3.9] — is the whole computational apparatus of the two papers.

**What is settled here.** Both papers also record, and prove, a structural fact about  $D_q$  which we restate as Theorem 3.1 below: writing  $H = F - G$ , the quantity  $D_q(F, G)$  is *half the largest total variation of  $H$  along a chain of  $q$  interior points, pinned at 0 at both ends* — a zigzag. This paper begins there and supplies what those papers do not have, namely an algorithm and values.

- (A) *A linear-time algorithm* (§4). The largest zigzag of a fixed length along a sequence is computed by a dynamic programme with running maxima, at cost  $O(qN)$  rather than  $\binom{N+q}{q}$ . We check the implementation against the two papers’ own definition (1), transcribed literally, on every interleaving word of a (3, 4) sample for  $q = 1, 2, 3$  and of a (4, 5) sample for  $q = 2, 3$  (Theorem 4.1).
- (B) *Exact null distributions with  $q \geq 3$*  (§5). Under the null hypothesis the two-sample statistic is a function of the interleaving pattern alone, which is uniform on a finite set, so its null law is a finite count. We give it exactly for OVL-3 at sample sizes (4, 4), (5, 5), (6, 6), (4, 5), (5, 6) and for OVL-4 at (6, 6) (Theorem 5.1), together with the comparison  $D_3 \geq D_2$ , strict on exactly 350 of the 462 words at (5, 6), which shows that OVL-3 is a genuinely different test.

Two further things are settled, and both are background rather than new mathematics; we state them plainly because a reader of [4, 5] will want to know. First, for  $q = 2$  the zigzag collapses to a range,

$$(2) \quad D_2(F, G) = \sup(F - G) - \inf(F - G),$$

so that for an empirical  $F_n$  against a continuous  $F$  one has  $D_2(F_n, F) = D_n^+ + D_n^-$ , which is *Kuiper’s statistic*  $V_n$  [7]. Identity (2) is [5]’s own lemma, in exactly this generality (§3); what is missing there is only the name, and with the name comes sixty years of finite-sample theory [11, 1, 3]. In particular the  $m \neq n$  two-sample problem [4] calls open is solved and packaged: **Kuiper2sample** in the R package **KSgeneral** computes that exact  $p$ -value for arbitrary sample sizes [6, 3]. We therefore present our  $q = 2$  tables (§6) as reproductions, not discoveries, and our exact one-sample  $p$ -values (§8) as a labelled computation outside the formal development. Second, the asymptotic substitute used in [5] is quantitatively poor at the smallest sample size that paper works with: at  $n = 8$  it reports  $p = 0.106$  at the exact 5% point (Remark 8.2).

Everything stated below as a numbered theorem or proposition is checked by the Lean 4 kernel over Mathlib; §11 says exactly which parts rest on exhaustive machine computation and which are proved in the ordinary sense.

## 2. THE STATISTIC AND ITS NULL MODEL

**Reading  $D_q$  along a chain.** Let  $F, G \in \mathcal{F}$  be step functions whose jumps lie in a finite set  $x_1 < \dots < x_N$  — the case of two empirical distribution functions. Enumerate the breakpoint values as  $A_0, \dots, A_N$  and  $B_0, \dots, B_N$  with  $A_0 = B_0 = 0$  and  $A_N = B_N = 1$ ; then each admissible  $v$  in (1) may be replaced by a nondecreasing chain of indices  $0 \leq a_1 \leq \dots \leq a_q \leq N$  without changing the infimum, and, writing  $a_0 = 0$ ,  $a_{q+1} = N$ ,

$$(3) \quad r(a) = \sum_{i=0}^q \min(A_{a_{i+1}} - A_{a_i}, B_{a_{i+1}} - B_{a_i}), \quad D_q = 1 - \min_a r(a).$$

Formula (3) is the definition we compute against; it is transcribed literally into the formal development, where it is the function `dqNaive`. In the one-sample case, where  $G$  is continuous, the extrema of  $F_n - G$  occur at the sample points and at their left limits, so the same finite reading applies once the chain is taken through both; that is the chain used in Proposition 7.2.

**The two-sample null model.** Let  $\Gamma_{m,n}$  be the set of binary words with  $m$  zeros and  $n$  ones; a word  $\gamma \in \Gamma_{m,n}$  records the interleaving pattern of two samples of sizes  $m$  and  $n$ , a 0 marking an observation from the first sample. Under the null hypothesis that both samples come from the same continuous law,  $\gamma$  is uniform on  $\Gamma_{m,n}$  and  $\#\Gamma_{m,n} = \binom{m+n}{m}$ ; this is the null model of [4, Rem. 3.3]. Both  $F_{0,m}$  and  $F_{1,n}$  are determined by  $\gamma$ , so  $D_q(F_{0,m}, F_{1,n})$  is a function  $D_q(\gamma)$  of the word alone and its null distribution is a finite count.

Set  $L = \text{lcm}(m, n)$  and define the *scaled path* of  $\gamma = \gamma_1 \dots \gamma_{m+n}$  by

$$(4) \quad S_0 = 0, \quad S_j = S_{j-1} + \begin{cases} L/m, & \gamma_j = 0, \\ -L/n, & \gamma_j = 1, \end{cases}$$

so that  $S_j = L \cdot (F_{0,m} - F_{1,n})$  read after the  $j$ -th pooled observation, all  $S_j$  are integers, and  $S_0 = S_{m+n} = 0$ . Every quantity below is stated in this integer scale, and a stated value  $k$  of  $L D_2$  means  $D_2 = k/L$ .

## 3. THE ZIGZAG FORM

The three statements of this section are due to Komaba, Johno and Nakamoto and are proved in [5]; we state them because everything later rests on them, and because our formal development contains them in a slightly more general setting (an arbitrary linearly ordered field in place of  $\mathbb{R}$ ), which is a formalisation contribution and nothing more. Precise attribution, by the numbering of [5] at version 1: Theorem 3.1 is [5, Rem. 5.10], proved there by the same min identity; Lemma 3.2 is the sign analysis proving [5, Lem. 5.22] (referred there to [4, Lem. 7.4]); and Theorem 3.3 is [5, Lem. 5.23], “For any  $F, G \in \mathcal{F}$ ,  $D_2(F, G) = \bar{\delta}_{F,G} - \underline{\delta}_{F,G}$ ”, i.e. stated there for arbitrary  $F, G \in \mathcal{F}$ .

**Theorem 3.1** (Komaba–Johno–Nakamoto; the OVL objective is a total variation). *Let  $A, B : \mathbb{N} \rightarrow \alpha$  take values in a linearly ordered field and let  $n \in \mathbb{N}$ . Then, with  $H = A - B$ ,*

$$\sum_{i < n} \min(A_{i+1} - A_i, B_{i+1} - B_i) = \frac{(A_n - A_0) + (B_n - B_0)}{2} - \frac{1}{2} \sum_{i < n} |H_{i+1} - H_i|.$$

*In particular, for  $A_0 = B_0 = 0$  and  $A_n = B_n = 1$  the left-hand side is  $r$  of (3) and the identity reads  $r = 1 - \frac{1}{2} \sum_i |H_{i+1} - H_i|$ ; consequently  $D_q(F, G)$  is half the largest total variation of  $H = F - G$  along a chain of  $q$  interior points pinned at 0 at both ends.*

*Proof.* Pointwise  $\min(x, y) = \frac{1}{2}(x + y - |x - y|)$ , and  $(A_{i+1} - A_i) - (B_{i+1} - B_i) = H_{i+1} - H_i$ ; summing and telescoping the two increment sums gives the display. The consequence is (1) read through (3).  $\square$

**Lemma 3.2** (the pointwise  $q = 2$  collapse). *For all  $x, y$  in a linearly ordered field,*

$$|x| + |y - x| + |y| = 2 \left( \max\{0, x, y\} - \min\{0, x, y\} \right).$$

*Proof.* Eight cases on the signs of  $x$ ,  $y$  and  $y - x$ ; each is an identity between linear expressions. Both sides are twice the range of the three-element set  $\{0, x, y\}$ .  $\square$

**Theorem 3.3** (Komaba–Johno–Nakamoto;  $D_2$  is a range). *Let  $s_0, \dots, s_N$  satisfy  $s_0 = 0$ , and let  $M$  and  $m$  be an attained upper and an attained lower bound of  $\{s_0, \dots, s_N\}$ . Then*

$$\max\{|s_i| + |s_j - s_i| + |s_j| : 0 \leq i \leq j \leq N\} = 2(M - m).$$

*Consequently, if  $A, B$  are as in Theorem 3.1 with  $A_0 = B_0 = 0$ ,  $A_N = B_N = 1$ , and  $M, m$  are the attained maximum and minimum of  $H = A - B$ , then the minimum over  $0 \leq i \leq j \leq N$  of*

$$\min(A_i - A_0, B_i - B_0) + \min(A_j - A_i, B_j - B_i) + \min(A_N - A_j, B_N - B_j)$$

*equals  $1 - (M - m)$ , so that  $D_2 = M - m$ ; that is, (2) holds. For  $A = F_n$  empirical and  $B = F$  continuous this reads  $D_2(F_n, F) = D_n^+ + D_n^- = V_n$ , Kuiper’s statistic.*

*Proof.* The upper bound is Lemma 3.2 together with  $m \leq \min\{0, s_i, s_j\}$  and  $\max\{0, s_i, s_j\} \leq M$ , which use  $s_0 = 0$  and hence  $m \leq 0 \leq M$ ; the bound is attained at the pair of indices realising  $M$  and  $m$ , in whichever order they occur. The second statement follows by Theorem 3.1 applied to the three-term chain. In the formal development the two halves are recorded as an **IsGreatest** and an **IsLeast** statement, so that attainment is part of the claim.  $\square$

*Remark 3.4.* Theorem 3.3 identifies the one-sample OVL-2 statistic with Kuiper’s  $V_n$  [7], and this is why the limit law obtained in [5], namely  $\lim_n \mathbb{P}(\sqrt{n} D_2(F_n, F) \geq a) = 2 \sum_{i \geq 1} (4i^2 a^2 - 1) e^{-2i^2 a^2}$ , is Kuiper’s 1960 limit law. Solving it numerically at  $\alpha = 0.05$  gives  $a = 1.747260\dots$ , the classical Kuiper 5% point 1.747. The identification is of the *statistic*, not of the identity: the identity is [5]’s own.

#### 4. A LINEAR-TIME ALGORITHM FOR $D_q$

**The dynamic programme.** Fix a sequence  $s_0, \dots, s_N$  with  $s_0 = s_N = 0$  — in the two-sample case the scaled path (4), so that  $s_j = S_j$  and the answer will be  $2L D_q$ . For  $k \geq 1$  set

$$f_k(j) = \max\left\{\sum_{i < k} |s_{a_{i+1}} - s_{a_i}| : 0 = a_0 \leq a_1 \leq \dots \leq a_k = j\right\},$$

the largest total variation of a  $k$ -step zigzag from index 0 to index  $j$ . Then  $f_1(j) = |s_j|$  and, since  $|s_j - s_i| = \max(s_j - s_i, s_i - s_j)$ ,

$$(5) \quad f_{k+1}(j) = \max_{i \leq j} \left( f_k(i) + |s_j - s_i| \right) = \max \left( s_j + \max_{i \leq j} (f_k(i) - s_i), -s_j + \max_{i \leq j} (f_k(i) + s_i) \right).$$

The two inner maxima are running maxima, so one layer of (5) costs  $O(N)$  and

$$2 D_q \cdot L = f_{q+1}(N)$$

is computed in  $q$  layers, i.e. in  $O(qN)$  arithmetic operations, against the  $\binom{N+q}{q}$ -term minimisation (3) that [4, 5] use. The correctness of (5) is the elementary statement that an optimal  $(k+1)$ -step zigzag ending at  $j$  restricts to an optimal  $k$ -step zigzag ending at its last interior index; what converts it into a statement about  $D_q$  is Theorem 3.1, and only that. Neither step is part of the machine-checked development, whose content is the following exhaustive agreement of the implementation with the definition (3).

**Theorem 4.1** (the algorithm computes the sources’  $D_q$ ). *Let  $\text{dp}_q(\gamma)$  denote the output  $f_{q+1}(N)$  of the dynamic programme (5) on the scaled path of  $\gamma$ , and let  $\text{def}_q(\gamma)$  denote  $L \cdot D_q(\gamma)$  computed from the definition (3) by minimising  $r$  over every nondecreasing index chain. Then*

- (1)  $\text{dp}_q(\gamma) = 2 \text{def}_q(\gamma)$  for every  $\gamma \in \Gamma_{3,4}$  and every  $q \in \{1, 2, 3\}$ , and for every  $\gamma \in \Gamma_{4,5}$  and every  $q \in \{2, 3\}$ ;
- (2) the range  $\max_j S_j - \min_j S_j$  of the scaled path equals  $\text{def}_2(\gamma)$  for every  $\gamma \in \Gamma_{4,5}$  and every  $\gamma \in \Gamma_{4,6}$ ;
- (3)  $\#\Gamma_{m,n} = \binom{m+n}{m}$  for  $(m, n) \in \{(3, 4), (4, 5), (4, 6), (5, 6), (5, 5), (6, 6)\}$ , the six cells used in this paper.

*Proof, and what was computed.* Exhaustive enumeration. Part (1) evaluates both sides on all 35 words of  $\Gamma_{3,4}$  for three values of  $q$  and on all 126 words of  $\Gamma_{4,5}$  for two, i.e. 357 word-and- $q$  comparisons, each of which minimises  $r$  over all  $\binom{N+q}{q}$  chains (120 chains per word at  $N = 7, q = 3$ ; 220 at  $N = 9, q = 3$ ); part (2) adds 126 + 210 comparisons of the range form against the same definition. The claim is stated in the formal development as the assertion that the list of disagreeing words is empty, so that a single mismatch would be visible. Part (3) is the length of the enumerated list of words. The same comparison was run beforehand in an independent implementation outside the proof assistant, covering 1376 word-and- $q$  checks for  $q \leq 4$  over eight sample-size pairs including  $m \neq n$ , again with no mismatch.  $\square$

*Remark 4.2.* Theorem 4.1 is a verification on cells, not a proof of correctness for all inputs. What it excludes is a transcription error between (1) and the programme — the failure mode that matters here, since the object being computed is somebody else’s definition. A machine-checked proof of (5) for all sequences is a piece of work we have not done.

## 5. EXACT NULL DISTRIBUTIONS FOR $q \geq 3$

We can now answer, for small sample sizes, the question both papers leave open. Recall from §2 that under the null hypothesis the two-sample statistic is a function of the interleaving word, uniform on  $\Gamma_{m,n}$ .

**Theorem 5.1** (exact null distributions of OVL-3 and OVL-4). *Under the null hypothesis the two-sample OVL-3 statistic  $D_3$  has the following exact distributions, an entry  $t:c$  meaning that  $D_3 = t$  on exactly  $c$  of the  $\binom{m+n}{m}$  interleaving words:*

| $(m, n)$ | exact null distribution of $D_3$ |                    |                    |                    |                                                                                                    |
|----------|----------------------------------|--------------------|--------------------|--------------------|----------------------------------------------------------------------------------------------------|
| (4, 4)   | $\frac{1}{2}:8,$                 | $\frac{3}{4}:36,$  | $1:26$             |                    |                                                                                                    |
| (5, 5)   | $\frac{2}{5}:10,$                | $\frac{3}{5}:82,$  | $\frac{4}{5}:118,$ | $1:42$             |                                                                                                    |
| (6, 6)   | $\frac{1}{3}:12,$                | $\frac{1}{2}:162,$ | $\frac{2}{3}:412,$ | $\frac{5}{6}:276,$ | $1:62$                                                                                             |
| (4, 5)   | $\frac{2}{5}:1,$                 | $\frac{11}{20}:3,$ | $\frac{3}{5}:23,$  | $\frac{3}{4}:12,$  | $\frac{4}{5}:54, \quad 1:33$                                                                       |
| (5, 6)   | $\frac{1}{3}:1,$                 | $\frac{7}{15}:6,$  | $\frac{1}{2}:34,$  | $\frac{3}{5}:8,$   | $\frac{19}{30}:50, \quad \frac{2}{3}:130, \quad \frac{4}{5}:50, \quad \frac{5}{6}:132, \quad 1:51$ |

and the two-sample OVL-4 statistic at (6, 6) has the exact distribution

$$\frac{1}{3}:2, \quad \frac{1}{2}:60, \quad \frac{2}{3}:304, \quad \frac{5}{6}:396, \quad 1:162.$$

Moreover  $D_3 \geq D_2$  on every one of the 462 words of  $\Gamma_{5,6}$ , with strict inequality on exactly 350 of them.

*Proof, and what was computed.* Exhaustive enumeration of  $\Gamma_{m,n}$ , evaluating  $D_q$  on each word by the dynamic programme (5), whose agreement with the definition (3) is Theorem 4.1; the six cells enumerate 70, 252, 924, 126, 462 and 924 words respectively. The distributions are stated in the formal development as lists of (value, multiplicity) pairs in the integer scale  $2L D_q$ , and the multiplicities sum to  $\binom{m+n}{m}$  in each row. The final assertion is a pair of counts over  $\Gamma_{5,6}$ , carried in the formal development by a single declaration, which the ledger files with the controls of §7: the number of words with  $D_3 < D_2$  is 0, and the number with  $D_2 < D_3$  is 350.  $\square$

*Remark 5.2.* The inequality  $D_2 \leq D_3$  is elementary (§1); the content of the last clause of Theorem 5.1 is the strictness count. It says that OVL-3 is not a relabelling of OVL-2: on more than three quarters of the sample space the two statistics disagree, so the two tests order the sample space differently and have different null distributions — compare the (5, 6) rows of Theorem 5.1 and of Proposition 6.2 below. We are not aware of any previously published value of  $D_q$  for  $q \geq 3$ ; see §10.

## 6. THE CELLS WITH $q = 2$

This section records the  $q = 2$  half of the picture. None of it is new mathematics: by Theorem 3.3 the two-sample OVL-2 statistic is the two-sample Kuiper statistic  $\varsigma_{m,n} = \sup_t[F_m(t) - G_n(t)] - \inf_t[F_m(t) - G_n(t)]$ , whose exact distribution is computed for arbitrary sample sizes by the algorithm of Dimitrova, Jia and Kaishev [3] as shipped in **KSgeneral** [6]. We give the values because they are the ones [4] asks for, in that paper’s own vocabulary, and because the equal-sample case is our control on everything else.

**Proposition 6.1** (reproduction of the sources’ equal-sample law). *For  $m = n$  the exact null distribution of  $D_2$  is*

$$n = 4 : \quad \frac{1}{4}:2, \frac{1}{2}:28, \frac{3}{4}:32, 1:8; \quad n = 6 : \quad \frac{1}{6}:2, \frac{1}{3}:124, \frac{1}{2}:390, \frac{2}{3}:300, \frac{5}{6}:96, 1:12.$$

*Proof, and what was checked.* Exhaustive enumeration of  $\Gamma_{n,n}$  (70 and 924 words) with  $D_2$  evaluated as the range of the scaled path, which is Theorem 3.3. These are exactly the counts predicted by [4]: that paper proves  $\#\{\gamma \in \Gamma_{n,n} : \rho_2(\gamma) \geq 1 - k/n\} = [x^n](Q'_{k+1}/Q_k - Q'_{k+2}/Q_{k+1})$  with  $Q_0 = Q_1 = 1$  and  $Q_{i+2} = Q_{i+1} - xQ_i$ , giving the cumulative counts 0, 2, 30, 62, 70 at  $n = 4$  and 0, 2, 126, 516, 816, 912, 924 at  $n = 6$ , which are the partial sums above. Before any of the values of §5 were computed we implemented that generating function independently in exact rational power series and compared it with the enumeration for  $n = 1, \dots, 10$  and every  $k$ : 65 cells, 65 agreements. The same numbers appear in the On-Line Encyclopedia of Integer Sequences [10] as sequence **A274878** (a statistic on orbital systems, contributed by P. Luschny), whose rows 8 and 12 are  $[0, 2, 28, 32, 8]$  and  $[0, 2, 124, 390, 300, 96, 12]$  — that triangle is indexed by the number of sectors, here  $m + n$ ; and the cumulative counts  $\#\{\gamma \in \Gamma_{n,n} : D_2 \leq 2/n\} = 1, 2, 6, 14, 30, 62, 126, 254, 510, 1022$  for  $n = 0, \dots, 9$  are **A095121**.  $\square$

**Proposition 6.2** (exact null distributions and  $p$ -values at  $m \neq n$ ). *The exact null distribution of the two-sample OVL-2 statistic  $D_2$  at unequal sample sizes is*

| $(m, n)$ | exact null distribution of $D_2$                                                                                                                                            |
|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| (3, 4)   | $\frac{1}{2}:7, \frac{2}{3}:7, \frac{3}{4}:14, 1:7$                                                                                                                         |
| (3, 5)   | $\frac{7}{15}:8, \frac{3}{5}:8, \frac{2}{3}:16, \frac{4}{5}:16, 1:8$                                                                                                        |
| (4, 5)   | $\frac{2}{5}:9, \frac{1}{2}:9, \frac{11}{20}:18, \frac{3}{5}:36, \frac{3}{4}:18, \frac{4}{5}:27, 1:9$                                                                       |
| (4, 6)   | $\frac{1}{3}:5, \frac{5}{12}:10, \frac{1}{2}:45, \frac{7}{12}:40, \frac{2}{3}:40, \frac{3}{4}:30, \frac{5}{6}:30, 1:10$                                                     |
| (5, 6)   | $\frac{1}{3}:11, \frac{2}{5}:11, \frac{13}{30}:22, \frac{7}{15}:44, \frac{1}{2}:88, \frac{3}{5}:44, \frac{19}{30}:66, \frac{2}{3}:88, \frac{4}{5}:33, \frac{5}{6}:44, 1:11$ |

*In particular the exact upper-tail  $p$ -values are  $\mathbb{P}(D_2 \geq 1) = \frac{1}{21}$ ,  $\mathbb{P}(D_2 \geq \frac{5}{6}) = \frac{4}{21}$ ,  $\mathbb{P}(D_2 \geq \frac{3}{4}) = \frac{1}{3}$  at (4, 6), and  $\mathbb{P}(D_2 \geq 1) = \frac{1}{42}$ ,  $\mathbb{P}(D_2 \geq \frac{5}{6}) = \frac{5}{42}$ ,  $\mathbb{P}(D_2 \geq \frac{4}{5}) = \frac{4}{21}$  at (5, 6).*

*Proof, and what was computed.* Exhaustive enumeration of  $\Gamma_{m,n}$  (35, 56, 126, 210 and 462 words), with  $D_2$  evaluated as the range of the scaled path (4); the  $p$ -values are the corresponding counts divided by  $\binom{m+n}{m}$ , as exact rationals.  $\square$

*Remark 6.3.*  $1/21$  and  $1/42$  are the smallest attainable significance levels of the two-sample OVL-2 test at (4, 6) and (5, 6), which is the number a practitioner needs before running the test at all. Note also that the support of  $D_2$  is not an interval in the natural scale: at (4, 6) the value  $11/12$  is attained by none of the 210 words (Proposition 7.1).

*Remark 6.4.* We have not found these particular tables printed anywhere, and searches of the On-Line Encyclopedia [10] for the multiplicity sequences

$$5, 10, 45, 40, 40, 30, 30, 10 \quad \text{and} \quad 11, 11, 22, 44, 88, 44, 66, 88, 33, 44, 11$$

return nothing. That is a statement about the printed record only: the algorithm of [3], described in the documentation of [6] as “accurate and valid for arbitrary (possibly large) sample sizes” with worst-case running time  $O((mn)^2)$ , computes exactly this distribution. That documentation also places the algorithm in its lineage; we quote it rather than paraphrase, since we did not read [3] itself (the publisher’s full text was not accessible to us): `Kuiper2sample` “calculates the exact  $p$ -value of the Kuiper test using an algorithm from Dimitrova, Jia, Kaishev (2024), which is based on extending the algorithm provided by Nikiforov (1994) and generalizing the method due to Maag and Stephens (1968) and Hirakawa (1973)” [6, `Kuiper2sample`]. The three works named there are U. R. Maag and M. A. Stephens, *The  $V_{NM}$  two-sample test*, Ann. Math. Statist. **39** (1968), no. 3, 923–935 (doi:10.1214/aoms/1177698324); K. Hirakawa, *The two-sample Kuiper test*, TRU Mathematics **9** (1973), 99–118 (Zbl 0295.62047); and A. M. Nikiforov, *Algorithm AS 288: exact Smirnov two-sample tests for arbitrary distributions*, Appl. Statist. **43** (1994), no. 1, 265–270 (doi:10.2307/2986126). Those bibliographic data are from Crossref and zbMATH, consulted 3 September 2026; none of the three papers was read here.

## 7. CONTROLS

Exhaustive counts prove nothing if the enumeration is silently empty or the scaling is off by one. The following two propositions are the controls; they are not results.

**Proposition 7.1** (negative controls on the two-sample tables). *On  $\Gamma_{4,5}$ , where  $L = 20$ : the claim that every word has  $LD_2 \leq 16$  is false, exactly 9 words having  $LD_2 > 16$ ; and the claim that every word has  $LD_2 \geq 11$  is false, exactly 18 words having  $LD_2 < 11$ . On  $\Gamma_{4,6}$ , where  $L = 12$ : the value  $LD_2 = 11$  is attained by 0 words, while  $LD_2 = 4$  is attained by 5 and  $LD_2 = 12$  by 10. Finally, on  $\Gamma_{5,6}$  the number of words with  $D_3 < D_2$  is 0 — the first half of the comparison quoted in Theorem 5.1, which the formal development carries as one declaration and the ledger files here.*

*Proof, and what was computed.* Filter lengths over the full enumerations of  $\Gamma_{4,5}$  (126 words),  $\Gamma_{4,6}$  (210) and  $\Gamma_{5,6}$  (462). The first two counts refute a too-small and a too-large claim about the same table, so neither an empty enumeration nor a misplaced quantifier could pass; the third shows both that the extreme values of Proposition 6.2 are attained and that the support has a hole.  $\square$

**Proposition 7.2** (the one-sample statistic, at instances). *Let  $x_1 \leq x_2 \leq x_3$  run over all 455 nondecreasing triples with entries in  $\{0, \frac{1}{12}, \dots, 1\}$ , let  $F_3$  be the corresponding empirical distribution function and  $U$  the uniform distribution function on  $[0, 1]$ ; write  $\Delta^+ = \max_i (\frac{i}{3} - x_i)^+$  and  $\Delta^- = \max_i (x_i - \frac{i-1}{3})^+$  for the two one-sided Kolmogorov–Smirnov statistics of the sample. Then  $D_2(F_3, U)$ , computed as  $\sup(F_3 - U) - \inf(F_3 - U)$  over the chain of breakpoints and their left limits, equals  $\Delta^+ + \Delta^-$  on every one of the 455 samples; it differs from the Kolmogorov–Smirnov statistic  $\max(\Delta^+, \Delta^-)$  on 385 of them; and  $D_2(F_3, U) \geq \frac{1}{3}$  on all of them.*

*Proof, and what was computed.* Exhaustive evaluation on the 455 samples, with all values carried as exact rationals. The chain includes the left limit of  $F_3$  at each order statistic, so that the infimum in (1) over half-open intervals is computed exactly; this is the point of the proposition, together with the second clause, which shows that the identity of Theorem 3.3 is not the trivial statement that  $D_2$  is the Kolmogorov–Smirnov statistic.  $\square$

## 8. COMPUTATIONS OUTSIDE THE FORMAL DEVELOPMENT

The two computations of this section are reported as measurements. Neither is a theorem here: the first is a volume computation that the formal development does not carry, the second is an observation about a bibliography.

*Remark 8.1* (exact one-sample  $p$ -values). By Theorem 3.3 the one-sample OVL-2 statistic is Kuiper's  $V_n$ , whose null law does not depend on the continuous reference  $F$  (as [5] itself proves for  $D_q(F_n, F)$ ). We computed  $p_{2,n}(v) = \mathbb{P}(v \leq D_2(F_n, F))$  as an exact rational for  $n = 6, 8, 10, 12, 14, 16$ . At  $n = 8$  and  $n = 16$ , the two smallest sample sizes in the experiments of [5]:

| $v$ | $p_{2,8}(v)$ | decimal    | $v$  | $p_{2,16}(v)$                       | decimal      |
|-----|--------------|------------|------|-------------------------------------|--------------|
| 2/8 | 16069/16384  | 0.98077393 | 2/16 | 35183733575957/35184372088832       | 0.9999818524 |
| 3/8 | 77837/131072 | 0.59384918 | 3/16 | 4379073653923621/4503599627370496   | 0.9723496794 |
| 4/8 | 2489/16384   | 0.15191650 | 4/16 | 408710943621437/562949953421312     | 0.7260164800 |
| 5/8 | 4299/262144  | 0.01639938 | 5/16 | 12825043049693093/36028797018963968 | 0.3559664521 |
| 6/8 | 81/131072    | 0.00061798 | 6/16 | 266003045907425/2251799813685248    | 0.1181290825 |
| 7/8 | 1/262144     | 0.00000381 | 7/16 | 977296666358231/36028797018963968   | 0.0271254315 |
|     |              |            | 8/16 | 4830793155673/1125899906842624      | 0.0042906062 |

A 14-step bisection on the exact distribution function encloses the exact 5% upper confidence limits:

$$u_{2,8}(0.05) \in \left[ \frac{74519}{131072}, \frac{37263}{65536} \right] = [0.568535, 0.568588],$$

$$u_{2,16}(0.05) \in \left[ \frac{108349}{262144}, \frac{27091}{65536} \right] = [0.413319, 0.413376].$$

*The method.* Take  $F$  uniform on  $[0, 1]$  and let  $U_{(1)} \leq \dots \leq U_{(n)}$  be the order statistics, of joint density  $n!$  on the ordered simplex, so that  $D_n^+ = \max_i (\frac{i}{n} - U_{(i)})^+$  and  $D_n^- = \max_i (U_{(i)} - \frac{i-1}{n})^+$ . The event  $\{V_n \leq v\}$  is not a union of polytopes indexed by a combinatorial pattern but a *projection*: it is the set of samples for which there exists  $t \in [0, v]$  with  $D_n^- = t$  and  $D_n^+ \leq v - t$ . We slice at  $t = D_n^-$ . Note first that  $D_n^- > 0$  almost surely, since  $D_n^- = 0$  forces  $U_{(1)} \leq 0$ ; likewise  $D_n^+ > 0$  almost surely, since  $D_n^+ = 0$  forces  $U_{(n)} \geq 1$ . So the slice at  $t = 0$  contributes nothing and may be dropped. For  $t \in (0, v]$  the index  $k$  at which  $D_n^-$  is attained is almost surely unique, and fixing it fixes  $U_{(k)} = c := \frac{k-1}{n} + t$ . Given  $k$  and  $t$ , the constraints  $D_n^- \leq t$  and  $D_n^+ \leq v - t$ , together with  $U_{(i)} \in [0, 1]$ , read

$$\alpha_i(t) \leq U_{(i)} \leq \beta_i(t), \quad \alpha_i(t) = \left( \frac{i}{n} + t - v \right)^+, \quad \beta_i(t) = \min \left( 1, \frac{i-1}{n} + t \right),$$

and they cut the remaining coordinates into two blocks that interact only through  $U_{(k)} = c$ . For  $i < k$  the ordering constraint  $U_{(i)} \leq c$  is automatic, because  $\beta_{k-1} \leq \frac{k-2}{n} + t < c$ ; for  $i > k$  it must be imposed, by replacing  $\alpha_i$  with  $\max(\alpha_i, c)$ . Writing  $A_k(t)$  for the volume of the first band  $\{x_1 \leq \dots \leq x_{k-1}, \alpha_i \leq x_i \leq \beta_i\}$  and  $B_k(t)$  for the volume of the second,  $\{x_{k+1} \leq \dots \leq x_n, \max(\alpha_i, c) \leq x_i \leq \beta_i\}$ , and integrating the coordinate  $U_{(k)} = c$  out as  $t$  (unit Jacobian) gives

$$\mathbb{P}(V_n \leq v) = \sum_{k=1}^n \int_0^v n! A_k(t) B_k(t) dt \quad \left( \frac{1}{n} \leq v \leq 1 \right):$$

$n$  one-dimensional integrals in place of an  $n$ -dimensional volume. (Our implementation carries a boundary term for the slice  $t = 0$ ; by the paragraph above it is identically zero, and it evaluates to zero because its first band requires  $0 \leq x_1 \leq 0$ .) Each integrand is piecewise polynomial in  $t$ ; on each piece the implementation fits a polynomial of degree at most  $n - 1$  through  $n$  exact rational values, *verifies* the fit at two further points and bisects the piece if the verification fails, then integrates the fitted polynomial exactly. All arithmetic is over  $\mathbb{Q}$ .

*Status of these numbers.* They are correct but not new: exact closed forms for the finite-sample distribution of  $V_n$  are classical, [11] being the standard reference and [1] a later treatment of the same  $p$ -value. Neither of those two was read here (both are paywalled or offline for us), so what follows is a check against the formula as it is transmitted, not against a primary text. Three checks were made.

*First, the classical upper-tail formula.* We take it in the form implemented by the function `kuiper_false_positive_probability` of `astropy` [2], whose documentation records that “Eq 7 of Paltani 2004 appears to incorrectly quote the original formula (Stephens 1965). This function implements the original formula, as it produces a result closer to Monte Carlo simulations.” It reads

$$\mathbb{P}(V_n \geq v) = \sum_{t=0}^{\lfloor n(1-v) \rfloor} \binom{n}{t} \left(1 - v - \frac{t}{n}\right)^{n-t-1} T_t, \quad y = v + \frac{t}{n},$$

$$T_t = y^{t-3} \left[ y^3 n - y^2 t \left(3 - \frac{2}{n}\right) + \frac{y t(t-1)}{n} \left(3 - \frac{2}{n}\right) - \frac{t(t-1)(t-2)}{n^2} \right],$$

used when  $v > \frac{1}{2}$  ( $n$  even) or  $v > \frac{n-1}{2n}$  ( $n$  odd), with two further closed forms for  $v < 2/n$  and  $v < 3/n$ . Evaluated in exact rational arithmetic this reproduces our values *exactly*, as rationals, at every point where one of its branches applies: at  $n = 8$  for  $v = \frac{1}{4}, \frac{5}{8}, \frac{3}{4}, \frac{7}{8}$ , at  $n = 16$  for  $v = \frac{1}{8}$ , and at  $(n, v) = (3, \frac{7}{10}), (3, \frac{4}{5}), (4, \frac{2}{5}), (5, \frac{7}{10}), (5, \frac{4}{5}), (6, \frac{1}{4}), (6, \frac{2}{5}), (8, \frac{1}{5}), (8, \frac{3}{10})$ . The remaining entries of the two tables above lie in the range  $3/n \leq v \leq \frac{1}{2}$ , which no branch of the formula covers. This also settles a discrepancy we had recorded: the transcription of the same formula in the secondary source [12] did *not* reproduce our  $n = 8$  values (at  $v = 5/8$ : ours 0.01639938, that transcription 0.02410984). It differs from the form above in two places — its middle term carries an extra factor  $1/n$ , and the term  $y t(t-1)(3-2/n)/n$  is absent — and those two differences account for the whole gap. We make no claim about which rendering matches the printed text of [11], which we have not seen.

*Second, an independent recomputation.* The decomposition above was re-implemented from scratch, slicing on  $t = D_n^+$  rather than  $D_n^-$ , with a different recursion for the band volumes and a different fitting and verification rule; it reproduces all thirteen exact rationals of the two tables and all four endpoints of the two confidence-limit enclosures.

*Third, Monte Carlo.* A  $2 \cdot 10^7$ -sample run in a separate program with its own generator, sharing no code with the exact route, agrees with all six  $n = 8$  values to within 1.7 standard errors, and exact-versus-Monte-Carlo agreement was also confirmed at  $n = 2, \dots, 5$ . The band-volume primitive is checked against  $1/j!$  for the unconstrained simplex,  $j = 1, \dots, 7$ .

*An observation.* The computed values satisfy  $p_{2,n}(\frac{n-1}{n}) = n^{-(n-2)}$  at  $n = 6, 8, 10, 12$ . This is what the upper-tail formula displayed above would give: at  $v = (n-1)/n$  the sum runs over  $t = 0, 1$ , the  $t = 1$  summand carries the factor  $(1 - v - \frac{1}{n})^{n-2} = 0$ , and the  $t = 0$  summand is  $n(1/n)^{n-1}$ . We record the identity as an observation about the computed table, not as a theorem of this paper: it rests on a formula we have checked only numerically, against values it was used to check.

*Remark 8.2* (a bibliographic observation). It is worth saying plainly where the obstacle of [5] lies, because it is not where that paper places it. The paper proves (2) itself, and derives from it the limit law of Remark 3.4 by quoting a standard fact about the range of a Brownian bridge; yet it reports that exact  $p$ -values could not be computed. The missing step is a name: the word “Kuiper” does not occur in [5], and neither Kuiper [7] nor Stephens [11] appears among the eight entries of its bibliography; the same holds for [4]. (Checked by case-insensitive search for “kuiper” and “stephens” over the extracted text of the arXiv v1 PDF of [5], over the L<sup>A</sup>T<sub>E</sub>X source and the .bib of the arXiv v2 e-print of [4], and over the extracted text of the published version of [4]: zero occurrences in each.) With the name, the one-sample exact distribution is classical [11] and the two-sample case at  $m \neq n$  — the other stated limitation — is available as software [6, 3].

The cost of the gap is measurable. Using the asymptotic critical value  $1.747260/\sqrt{n}$  in place of the exact one costs +8.65% at  $n = 8$  and +5.68% at  $n = 16$  relative to the exact values of Remark 8.1, and at the exact 5% point for  $n = 8$  the asymptotic formula reports  $p = 0.106$  — anti-conservative by more than a factor of two at the smallest sample size used in the experiments of [5]. This is a quantitative form of that paper’s own caveat that its asymptotic  $p$ -values are not applicable when the sample size is small. None of this diminishes the OVL- $q$  construction, whose interest is precisely the regime  $q \geq 3$  that has no classical counterpart.

## 9. QUESTIONS LEFT OPEN

**Question 9.1.** Theorem 5.1 enumerates  $\Gamma_{m,n}$  word by word, at cost  $\binom{m+n}{m} \cdot O(qN)$ ;  $(10, 10)$  is 184756 words and takes seconds, but  $(20, 20)$  is about  $1.4 \cdot 10^{11}$  words and is out of reach. Is there a dynamic programme in the *zigzag state* — carrying the layer values of (5) as the word is built letter by letter — which produces the whole null distribution without enumerating words? For  $q = 2$  such a lattice dynamic programme exists and makes the equal- and unequal-sample tables free.

**Question 9.2.** For  $q = 2$  and  $m = n$ , [4] has a closed form: a generating function in the Fibonacci-like polynomials  $Q_i$  (Proposition 6.1). Is there an analogue for  $q = 3$ ? The tables of Theorem 5.1 are the data to guess it from.

**Question 9.3.**  $D_q(\gamma)$  is nondecreasing in  $q$  and equals 1 as soon as the chain is allowed to take every index, since the total variation of the path (4) over all of  $\{0, \dots, m+n\}$  is  $2L$ . What is the distribution, over  $\Gamma_{m,n}$ , of the least  $q$  with  $D_q(\gamma) = 1$ ?

## 10. WHAT WAS SEARCHED

The novelty claims made here are claims about a printed record, so we state what was searched. Over a local mirror of the arXiv full texts (about 86 GB), a literal search for the strings `OVL-2` and `OVL-$q$` returns exactly two papers, [4] and [5], and neither contains a single value of  $D_q$  for  $q \geq 3$ ; a search for a total-variation or zigzag algorithm for this family returned nothing. The forward-citation record is thin: OpenAlex reports a citation count of 0 for the arXiv record of each paper, and Semantic Scholar reports 0 for [5] and 6 for [4] — one of them [5] itself and the other five applications (ROC analysis, diagnostic accuracy, imaging screens, banking), none touching exact  $p$ -values, Kuiper’s statistic, or  $q \geq 3$ . The On-Line Encyclopedia of Integer Sequences [10] returns nothing for the OVL-3 multiplicity sequences 12, 162, 412, 276, 62 at  $(6, 6)$  and 10, 82, 118, 42 at  $(5, 5)$ , while their  $q = 2$  analogues are found at once (Proposition 6.1). All of these searches were run on 3 September 2026, against [5] at version 1 and [4] at version 2.

Read the negative as “not found”, not as “new”. A maximum-total-variation dynamic programme is a standard algorithmic shape; what we claim for §4 is that it has not been applied to this family, not that the technique is new. And the  $q = 2$  material of §6 and Remark 8.1 is explicitly *not* claimed as new, for the reason given there.

## 11. VERIFICATION

Theorems 3.1, 3.3, 4.1, 5.1, Lemma 3.2 and Propositions 6.1, 6.2, 7.1, 7.2 are checked, in the form stated, by the kernel of the Lean 4 proof assistant [8] over the Mathlib library [9], in two modules. The first is kernel-clean: six declarations — the identities of Theorem 3.1 and Lemma 3.2, the two halves of Theorem 3.3 (recorded separately as a greatest-element and a least-element statement), and two auxiliary identities, the pointwise min formula and the three-term expansion used in the proof of Theorem 3.3 — proved over an arbitrary linearly ordered field, using no evaluation and depending only on `propext`, `Classical.choice` and `Quot.sound`. The second carries the finite tables: 23 theorems, each closed by `native_decide`, so that each additionally depends on one evaluation axiom for the compiled evaluation of the corresponding closed Boolean term. There is no `sorry` anywhere in the development. The statements of that module are in the integer scale of §2 — a distribution is a list of (value, multiplicity) pairs for  $LD_2$  or  $2LD_q$ , a  $p$ -value an exact rational, a count a list length — and the translation into the fractions printed above is the scaling (4) and nothing else. What each exhaustive computation ranges over is stated with the corresponding proof above.

The computations of Remark 8.1 are outside the formal development, and are labelled as such:  $\mathbb{P}(V_n \leq v)$  is an  $n$ -dimensional volume, and binding it to a machine-checked statement needs measure theory this development does not carry. A route exists — Steck’s determinant for ordered-uniform

rectangle probabilities, combined with the decomposition of Remark 8.1 — and we have not taken it. Remark 8.2 is likewise not a theorem.

For orientation, the measured costs in a stand-alone implementation of the same algorithms: the OVL-3 table at  $(6, 6)$  (924 words) takes 0.02 s and the OVL-4 table at  $(6, 6)$  0.03 s; one exact evaluation of  $\mathbb{P}(V_n \leq v)$  takes 0.9 s at  $n = 8$ , 12 s at  $n = 16$  and 73 s at  $n = 20$ , growing roughly like  $n^5$ , so that one exact critical value costs about 13 min per evaluation and 3 CPU-hours in total at  $n = 32$  and roughly 100 CPU-hours at  $n = 64$ ; we computed neither. Larger cells for  $q \geq 3$  wait on Question 9.1 rather than on more compute. The repository is private; repository reference to be added at submission.

## REFERENCES

- [1] H. Arsham, *Kuiper's P-value as a measuring tool and decision procedure for the goodness-of-fit test*, Journal of Applied Statistics **15** (1988), no. 2, 131–135; doi:10.1080/02664768800000020. Author, title, journal, volume, issue, pages and DOI are as recorded by Crossref (consulted 3 September 2026). Not read here; cited only as a finite-sample treatment of Kuiper's  $p$ -value.
- [2] Astropy project, function `kuiper_false_positive_probability` in `astropy/stats/funcs.py`, read from the project's source repository on 3 September 2026 (file at commit `d011a6c`). Its documentation says the function “uses the set of four formulas described in Paltani 2004” — S. Paltani, *Searching for periods in X-ray observations using Kuiper's test. Application to the ROSAT PSPC archive*, Astronomy and Astrophysics (2004) — and adds the note quoted in Remark 8.1, which says that the implementation follows [11] rather than Paltani's rendering of it. We use the function only as a transmission of [11], which we did not read.
- [3] D. S. Dimitrova, Y. Jia and V. K. Kaishev, *Efficient exact calculation of p-values of the two-sample Kolmogorov–Smirnov and Kuiper tests*, Journal of Statistical Computation and Simulation, published online 27 August 2026, 19 pp.; doi:10.1080/00949655.2026.2721410. Authors, title, journal, date and DOI are as recorded by Crossref (consulted 3 September 2026), which assigns the article no volume or issue; the full text was not accessible to us and was not read.
- [4] A. Komaba, H. Johno and K. Nakamoto, *A novel statistical approach for two-sample testing based on the overlap coefficient*, Journal of Mathematical Sciences, the University of Tokyo **30** (2023), no. 2, 205–240; MR4622637; preprint arXiv:2206.03166, whose version 2 is the accepted manuscript and is the one we read. Its §1 defines  $D_q$  and  $\rho_{q,m,n}$ ; the null model on  $\Gamma_{m,n}$  (Remark 3.3), the equal-sample generating function (Theorem 3.9) and the Conclusion quoted in §1 are as located there. Volume, issue, pages and year, and the numbering of Remark 3.3 and Theorem 3.9, were checked against the journal's own published PDF (consulted 3 September 2026); the journal registers no DOI.
- [5] A. Komaba, H. Johno and K. Nakamoto, *Extension of the one-sample Kolmogorov–Smirnov test*, arXiv:2408.10612 (2024), v1, which is the version we read; MSC2020 62G10, 62G20, 62-04. Definition (1), the identities of §3, the limit law of Remark 3.4 and the sentences quoted in §1 are from that version, and all numbered references to it above are that version's numbering. As of 3 September 2026 the arXiv record carries no journal reference and no publisher DOI, the submission history shows only v1, and a Crossref search by title and by authors finds no published version.
- [6] D. S. Dimitrova, Y. Jia, V. K. Kaishev and S. Tan, *KSgeneral: computing P-values of the one-sample K-S test and the two-sample K-S and Kuiper tests for (dis)continuous null distribution*, package for R version 2.1.0, published on the Comprehensive R Archive Network 23 August 2026; doi:10.32614/CRAN.package.KSgeneral. The function `Kuiper2sample` computes the exact  $p$ -value of the two-sample Kuiper test by the algorithm of [3]; the sentences quoted in §6 and Remark 6.4 are from the reference manual of that version, consulted on 3 September 2026.
- [7] N. H. Kuiper, *Tests concerning random points on a circle*, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A **63** (1960), 38–47. Not read here. Title, author, series, volume, year and pages are as recorded by zbMATH (Zbl 0096.12504) and by Crossref (doi:10.1016/s1385-7258(60)50006-0), consulted 3 September 2026; the statistic  $V_n = D_n^+ + D_n^-$  and its limit law are quoted as stated in [12].
- [8] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: A. Platzer and G. Sutcliffe (eds.), Automated Deduction — CADE 28, Lecture Notes in Computer Science **12699**, Springer, Cham, 2021, pp. 625–635; doi:10.1007/978-3-030-79876-5\_37.
- [9] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, pp. 367–381; doi:10.1145/3372885.3373824.
- [10] OEIS Foundation Inc., *The On-Line Encyclopedia of Integer Sequences*, <https://oeis.org>; sequences **A274878** (P. Luschny) and **A095121**, and the negative searches reported in §6 and §10, consulted on 3 September 2026.
- [11] M. A. Stephens, *The goodness-of-fit statistic  $V_N$ : distribution and significance points*, Biometrika **52** (1965), no. 3–4, 309–321; doi:10.1093/biomet/52.3-4.309 (JSTOR 2333685). Author, title, journal, volume, issue,

pages and year are as recorded by Crossref and zbMATH (Zbl 0192.26106), consulted 3 September 2026; the article itself is behind a paywall and was not read. The upper-tail formula attributed to it in [12] and in [2] is discussed in Remark 8.1; the two attributions differ, and only the second reproduces our values.

- [12] H.-Y. Zhang, Z.-Q. Feng, H. Liu, R.-J. Lin and Y. Zhou, *High order expansion method for Kuiper's  $V_n$  statistic in goodness-of-fit test*, arXiv:2310.19576. Read here for its statements of Kuiper's limit law and of the formulae of [11].