A Family of Exact Goodness-of-Fit Tests for High-Dimensional Discrete Distributions

Feras A. Saad MIT
Cameron E. Freer MIT
Nathanael L. Ackerman Harvard University
Vikash K. Mansinghka MIT

Abstract

The objective of goodness-of-fit testing is to assess whether a dataset of observations is likely to have been drawn from a candidate probability distribution. This paper presents a rank-based family of goodness-of-fit tests that is specialized to discrete distributions on high-dimensional domains. The test is readily implemented using a simulation-based, linear-time procedure. The testing procedure can be customized by the practitioner using knowledge of the underlying data domain. Unlike most existing test statistics, the proposed test statistic is distribution-free and its exact (non-asymptotic) sampling distribution is known in closed form. We establish consistency of the test against all alternatives by showing that the test statistic is distributed as a discrete uniform if and only if the samples were drawn from the candidate distribution. We illustrate its efficacy for assessing the sample quality of approximate sampling algorithms over combinatorially large spaces with intractable probabilities, including random partitions in Dirichlet process mixture models and random lattices in Ising models.

1 Introduction

We address the problem of testing whether a dataset of observed samples was drawn from a candidate probability distribution. This problem, known as goodness-of-fit testing, is of fundamental interest and has applications in a variety of fields including Bayesian statistics [10; 31], high-energy physics [34], astronomy [22], genetic association studies [17], and psychometrics [3].

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Rank-based methods are a popular approach for assessing goodness-of-fit and have received great attention in the nonparametric statistics literature [15]. However, the majority of existing rank-based tests operate under the assumption of continuous distributions [16, VI.8] and analogous methods for discrete distributions that are theoretically rigorous, customizable using domain knowledge, and practical to implement in a variety of settings remain much less explored.

This paper presents a new connection between rank-based tests and discrete distributions on high-dimensional data structures. By algorithmically specifying an ordering on the data domain, the practitioner can quantitatively assess how typical the observed samples are with respect to resampled data from the candidate distribution. This ordering is leveraged by the test to effectively surface distributional differences.

More specifically, we propose to test whether observations $\{y_1, \dots, y_n\}$, taking values in a countable set $\mathcal{T}$, were drawn from a given discrete distribution $\mathbf{p}$ on the basis of the rank of each y_i with respect to m i.i.d. samples $\{x_1, \dots, x_m\}$ from $\mathbf{p}$. If y_i was drawn from $\mathbf{p}$ then we expect its rank to be uniformly distributed over $\{0, 1, \dots, m\}$. When the ranks show a deviation from uniformity, it is unlikely that the y_i were drawn from $\mathbf{p}$. A key step is to use continuous random variables to break any ties when computing the ranks. We call this statistic the Stochastic Rank Statistic (SRS), which has several desirable properties for goodness-of-fit testing:

1. The SRS is distribution-free: its sampling distribution under the null does not depend on $\mathbf{p}$. There is no need to construct ad-hoc tables or use Monte Carlo simulation to estimate rejection regions.

2. The exact (non-asymptotic) sampling distribution of the SRS is a discrete uniform. This exactness obviates the need to apply asymptotic approximations in small-sample and sparse regimes.

3. The test is consistent against all alternatives. We show that the SRS is distributed as a discrete uniform if and only if $\{y_1, \dots, y_n\} \sim^{\text{iid}} \mathbf{p}$.

4. The test gives the practitioner flexibility in deciding the set of properties on which the observations be checked to agree with samples from $\mathbf{p}$. This flexibility arises from the design of the ordering on the domain that is used to compute the ranks.

5. The test is readily implemented using a procedure that is linear-time in the number of observations. The test is simulation-based and does not require explicitly computing $\mathbf{p}(x)$, which is especially useful for distributions with intractable probabilities.

While the test is consistent for any ordering $(\mathcal{T}, \prec)$ over the domain that is used to compute the SRS, the power of the test depends heavily on the choice of $\prec$. We show how to construct orderings in a variety of domains by (i) defining procedures that traverse and compare discrete data structures; (ii) composing probe statistics that summarize key numerical characteristics; and (iii) using randomization to generate arbitrary orderings.

The remainder of the paper is organized as follows. Section 2 reviews the goodness-of-fit problem and discusses related work. Section 3 presents the proposed test and several theoretical properties. Section 4 gives conceptual examples for distributions over integers, binary strings, and partitions. Section 5 applies the method to (i) compare approximate Bayesian inference algorithms over mixture assignments in a Dirichlet process mixture model and (ii) assess the sample quality of random lattices from approximate samplers for the Ising model.

2 The Goodness-of-Fit Problem

Problem 2.1. Let $\mathbf{p}$ be a candidate discrete distribution over a finite or countably infinite domain $\mathcal{T}$. Given observations $\{y_1, \dots, y_n\}$ drawn i.i.d. from an unknown distribution $\mathbf{q}$ over $\mathcal{T}$, is there sufficient evidence to reject the hypothesis $\mathbf{p} = \mathbf{q}$?

In the parlance of statistical testing, we have the following null and alternative hypotheses:

$$\mathsf{H}_0 \coloneqq [\mathbf{p} = \mathbf{q}] \qquad \mathsf{H}_1 \coloneqq [\mathbf{p} \neq \mathbf{q}].$$

A statistical test $\phi_n \colon \mathcal{T}^n \to \{\mathsf{reject}, \mathsf{not\ reject}\}$ says, for each size n dataset, whether to reject or not reject the null hypothesis H_0. We define the significance level

$$\alpha \coloneqq \Pr\{\phi_n(Y_{1:n}) = \mathsf{reject} \mid \mathsf{H}_0\} \tag{1}$$

to be the probability of incorrectly declaring reject. For a given level α, the performance of the test ϕ_n is characterized by its power

$$\beta \coloneqq \Pr\{\phi_n(Y_{1:n}) = \mathsf{reject} \mid \mathsf{H}_1\}, \tag{2}$$

which is the probability of correctly declaring reject.

Classical goodness-of-fit tests for nominal (unordered) data include the multinomial test [14]; Pearson chi-square test [23]; likelihood-ratio test [33]; nominal Kolmogorov–Smirnov test [13; 24]; and power-divergence statistics [26]. For ordinal data, goodness-of-fit test statistics include the ordinal Watson, Cramér–von Mises, and Anderson–Darling [7] tests as well as the ordinal Kolmogorov–Smirnov [4; 8]. These approaches typically suffer from statistical issues in large domains. They assume that $\mathbf{p}(x)$ is easy to evaluate (which is rarely possible in modern machine-learning applications such as graphical models) and/or require that each discrete outcome $x \in \mathcal{T}$ has a non-negligible expectation $n\mathbf{p}(x)$ [20; 28] (which requires a large number of observations n even when $\mathbf{p}$ and $\mathbf{q}$ are noticeably far from one another). In addition, the rejection regions of these statistics are either distribution-dependent (which requires reestimating the region for each new candidate distribution $\mathbf{p}$) or asymptotically distribution-free (which is inexact for finite-sample data and imposes additional statistical assumptions on $\mathbf{p}$ and $\mathbf{q}$). The Mann–Whitney U [19], which is also a rank-based test that bears some similarity to the SRS, is only consistent under median shift, whereas the proposed method is consistent under general distributional inequality.

Recent work in the theoretical computer science literature has established computational and sample complexity bounds for testing approximate equality of discrete distributions [5]. These methods have been primarily studied from a theoretical perspective and have not been shown to yield practical goodness-of-fit tests in practice, nor have they attained widespread adoption in the applied statistics community. For instance, the test in [1] is based on a variant of Pearson chi-square. It requires enumerating over the domain $\mathcal{T}$ and representing $\mathbf{p}(x)$ explicitly. The test in [32] requires specifying and solving a complex linear program. While these algorithms may obtain asymptotically sample-optimal limits, they are designed to detect differences between $\mathbf{p}$ and $\mathbf{q}$ in a way that is robust to highly adversarial settings. These tests do not account for any structure in the domain $\mathcal{T}$ that can be leveraged by the practitioner to effectively surface distributional differences.

Permutation and bootstrap resampling of test statistics are another family of tests for goodness-of-fit [11]. Theoretically rigorous and consistent tests can be obtained using kernel methods, including the maximum mean discrepancy [12] and discrete Stein discrepancy [35]. Since the null distribution is unknown, rejection regions are estimated by bootstrap resampling, which may be inexact due to discreteness of the data. Instead of bootstrapping, the SRS can be used to obtain an exact, distribution-free test by defining an ordering using the kernel. This connection is left for future work.

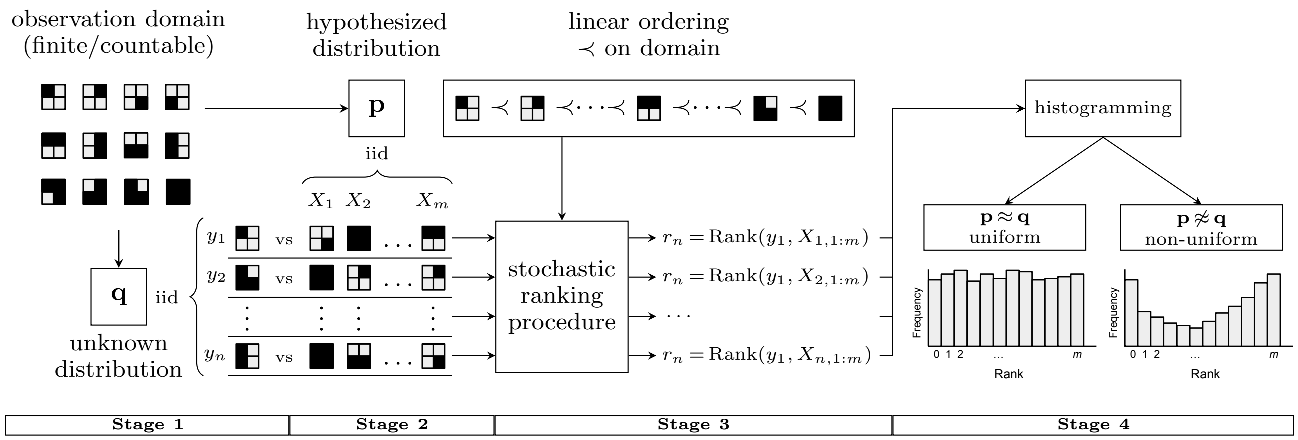


Figure 1: Overview of the proposed goodness-of-fit test for discrete distributions. **Stage 1**: Observations $\{y_1, \dots, y_n\}$ are assumed to be drawn i.i.d. from an unknown discrete distribution $\mathbf{q}$ over a finite or countable observation domain $\mathcal{T}$ (shown in the top-left corner). **Stage 2**: For each y_i, m samples $\{X_{i1}, \dots, X_{im}\}$ are simulated i.i.d. from the candidate distribution $\mathbf{p}$ over $\mathcal{T}$. **Stage 3**: Given a total order $\prec$ on $\mathcal{T}$ and the observed and simulated data, a stochastic ranking procedure returns the rank r_i of each y_i within $\{X_{i1}, \dots, X_{im}\}$, using uniform random numbers to ensure the ranks are unique. **Stage 4**: The histogram of the ranks $\{r_1, \dots, r_n\}$ is analyzed for uniformity over $\{0, 1, \dots, m\}$.

3 A Family of Exact and Distribution-Free GOF Tests

In this section we describe our proposed method for addressing the goodness-of-fit problem. The proposed procedure combines (i) the intuition from existing methods for ordinal data [7] that the deviation between the expected CDF and empirical CDF of the sample serves as a good signal for goodness-of-fit, with (ii) the flexibility of probe statistics in Monte Carlo-based resampling tests [11] to define, using an ordering $\prec$ on $\mathcal{T}$, characteristics of the distribution that are of interest to the experimenter. Figure 1 shows the step-by-step workflow of the proposed test and Algorithm 1 formally describes the testing procedure.

Algorithm 1 Exact GOF Test using SRS

Input: simulator for candidate dist. $\mathbf{p}$ over $\mathcal{T}$; i.i.d. samples $\{y_1, y_2, \dots, y_n\}$ from dist. $\mathbf{q}$; strict total order $\prec$ on T, of any order type; number $m \geq 1$ of datasets to resimulate; significance level α of hypothesis test;

Output: Decision to reject the null hypothesis $\mathsf{H}_0 : \mathbf{p} = \mathbf{q}$ versus alternative hypothesis $\mathsf{H}_1 : \mathbf{p} \neq \mathbf{q}$ at level α.

1: **for** $i = 1, 2, \dots, n$ **do**
2: $\quad X_1, X_2, \dots, X_m \sim^{\text{iid}} \mathbf{p}$
3: $\quad U_0, U_1, \dots, U_m \sim^{\text{iid}} \mathsf{Uniform}(0, 1)$
4: $\quad r_i \leftarrow \sum_{k=1}^{m} \mathbb{I}[X_k \prec y_i] + \mathbb{I}[X_k = y_i, U_k < U_0]$
5: Use a standard hypothesis test to compute p-value of $\{r_1, \dots, r_n\}$ under a discrete uniform on $\{0, \dots, m\}$.
6: **return** reject if $p \leq \alpha$, else not reject.

The proposed method addresses shortcomings of existing statistics in sparse regimes. It does not require the ability to compute $\mathbf{p}(x)$ and it is not based on comparing the expected frequency of each $x \in \mathcal{T}$ (which is often vanishingly small) with its observed frequency. Furthermore, the stochastic rank statistics r_i have an exact and distribution-free sampling distribution. The following theorem establishes that the r_i are uniformly distributed if and only if $\mathbf{p} = \mathbf{q}$. (Proofs are in the Appendix.)

Theorem 3.1. *Let $\mathcal{T}$ be a finite or countably infinite set, let $\prec$ be a strict total order on $\mathcal{T}$, let $\mathbf{p}$ and $\mathbf{q}$ be two probability distributions on $\mathcal{T}$, and let m be a positive integer. Consider the following random variables:*

$$X_0 \sim \mathbf{q} \tag{3}$$

$$X_1, X_2, \dots, X_m \sim^{\text{iid}} \mathbf{p} \tag{4}$$

$$U_0, U_1, U_2, \dots, U_m \sim^{\text{iid}} \mathsf{Uniform}(0, 1) \tag{5}$$

$$R = \textstyle\sum_{j=1}^{m} \mathbb{I}\left[X_j \prec X_0\right] + \mathbb{I}\left[X_j = X_0, U_j < U_0\right]. \tag{6}$$

Then $\mathbf{p} = \mathbf{q}$ if and only if for all $m \geq 1$, the rank R is distributed as a discrete uniform random variable on the set of integers $[m+1] := \{0, 1, \dots, m\}$.

Note that the r_i in line 4 of Algorithm 1 are n i.i.d. samples of the random variable R in Eq. (6), which is the rank of $X_0 \sim \mathbf{q}$ within a size m sample $X_{1:m} \sim^{\text{iid}} \mathbf{p}$. For Theorem 3.1, it is essential that ties are broken by pairing each X_i with a uniform random variable U_i, as opposed to, e.g., breaking each tie independently with probability 1/2, as demonstrated by the next example.

Example 3.2. Let $\mathcal{T}$ contain a single element. Then all the X_i (for $0 \leq i \leq m$) are equal almost surely. Break each tie between X_0 and X_j by flipping a fair coin. Then R is binomially distributed with m trials and weight 1/2, not uniformly distributed over $[m+1]$.

We now establish theoretical properties of R which form the basis of the goodness-of-fit test in Algorithm 1.

First note that in the case where all the X_i are almost surely distinct, the forward direction of Theorem 3.1, which establishes that if $\mathbf{p} = \mathbf{q}$ then the rank R is uniform for all $m \geq 1$, is easy to show and is known in the statistical literature [2]. However no existing results make the connection between rank statistics and discrete random variables over countable domains with ties broken stochastically. Nor do they establish that $\mathbf{p} = \mathbf{q}$ is a *necessary* condition for uniformity of R (across all m beyond some integer) and can therefore be used as the basis of a consistent goodness-of-fit test. We now state an immediate consequence of Theorem 3.1.

Corollary 3.3. *If $\mathbf{p} \neq \mathbf{q}$, then there is some $M \geq 1$ such that R is not uniformly distributed on $[M + 1]$.*

The next theorem significantly strengthens Corollary 3.3 by showing that if $\mathbf{p} \neq \mathbf{q}$, the rank statistic is non-uniform for *all but finitely many m.*

Theorem 3.4. *Let $\mathbf{p} \neq \mathbf{q}$ and M be defined as in Corollary 3.3. Then for all $m \geq M$, the rank R is not uniformly distributed on $[m + 1]$.*

In fact, unless $\mathbf{p}$ and $\mathbf{q}$ satisfy an adversarial symmetry relationship under the selected ordering $\prec$, the rank is non-uniform for *all* $m \geq 1$.

Corollary 3.5. *Let $\lhd$ denote the lexicographic order on $\mathcal{T} \times [0, 1]$ induced by $(\mathcal{T}, \prec)$ and $([0, 1], <)$. Suppose $\Pr\{(X, U_1) \lhd (Y, U_0)\} \neq 1/2$ for $Y \sim \mathbf{q}$, $X \sim \mathbf{p}$, and $U_0, U_1 \overset{\text{iid}}{\sim} \mathsf{Uniform}(0, 1)$. Then for all $m \geq 1$, the rank R is not uniformly distributed on $[m + 1]$.*

The next theorem establishes the existence of an ordering on $\mathcal{T}$ satisfying the hypothesis of Corollary 3.5.

Theorem 3.6. *If $\mathbf{p} \neq \mathbf{q}$, then there is an ordering $\prec^*$ whose associated rank statistic R is non-uniform for $m = 1$ (and hence by Theorem 3.4 for all $m \geq 1$).*

Intuitively, $\prec^*$ sets elements $x \in \mathcal{T}$ which have a high probability under $\mathbf{q}$ to be "small" in the linear order, and elements $x \in \mathcal{T}$ which have a high probability under $\mathbf{p}$ to be "large" in the linear order. More precisely, $\prec^*$ maximizes the sup-norm distance between the induced cumulative distribution functions $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$ of $\mathbf{p}$ and $\mathbf{q}$, respectively (Figure 3). Under a slight variant of this ordering, for finite $\mathcal{T}$, the next theorem establishes the sample complexity required to obtain exponentially high power in terms of the statistical distance $L_\infty(\mathbf{p}, \mathbf{q}) = \sup_{x \in \mathcal{T}} |\mathbf{p}(x) - \mathbf{q}(x)|$ between $\mathbf{p}$ and $\mathbf{q}$.

Theorem 3.7. *Given significance level $\alpha = 2\Phi(-c)$ for $c > 0$, there is an ordering for which the proposed test with $m = 1$ achieves power $\beta \geq 1 - \Phi(-c)$ using*

$$n \approx 4c^2 / L_\infty(\mathbf{p}, \mathbf{q})^4 \tag{7}$$

samples from $\mathbf{q}$, where Φ is the cumulative distribution function of a standard normal.

This key result is independent of the domain size and establishes a lower bound for any $\prec$ because it is based on the optimal ordering $\prec^*$. The next theorem derives the exact sampling distribution for any pair of distributions $(\mathbf{p}, \mathbf{q})$, which is useful for simulation studies (e.g., Figure 3) that characterize the power of the SRS.

Theorem 3.8. *The distribution of R is given by*

$$\Pr\{R = r\} = \sum_{x \in \mathcal{T}} H(x, m, r)\, \mathbf{q}(x) \tag{8}$$

for $0 \leq r \leq m$, where $H(x, m, r) :=$

$$\begin{cases} \displaystyle\sum_{e=0}^{m} \Bigg\{ \left[\sum_{j=0}^{e} \binom{m-e}{r-j} \left[\frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{r-j} \right. \\ \qquad \left. \left[1 - \frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{(m-e)-(r-j)} \left(\frac{1}{e+1}\right) \right] \\ \qquad \displaystyle\binom{m}{e} [\mathbf{p}(x)]^m \, [1-\mathbf{p}(x)]^{e-m} \Bigg\} & \text{if } 0 < \mathbf{p}(x) < 1, \\ \displaystyle\binom{r}{m} [\tilde{\mathbf{p}}(x)]^r \, [1 - \tilde{\mathbf{p}}(x)]^{m-r} & \text{if } \mathbf{p}(x) = 0, \\ \dfrac{1}{m+1} & \text{if } \mathbf{p}(x) = 1, \end{cases}$$

and $\tilde{\mathbf{p}}(x) := \sum_{x' \prec x} \mathbf{p}(x)$ is the CDF of $\mathbf{p}$.

4 Examples

We now apply the proposed test to a countable domain and two high-dimensional finite domains, illustrating a power comparison and how distributional differences can be detected when the number of observations is much smaller than the domain size. We use Pearson chi-square to assess uniformity of the SRS for Algorithm 1 (see [29] for alternative ways to test for a uniform null).

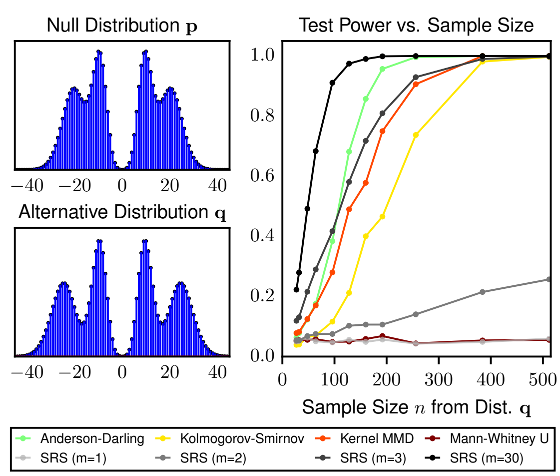
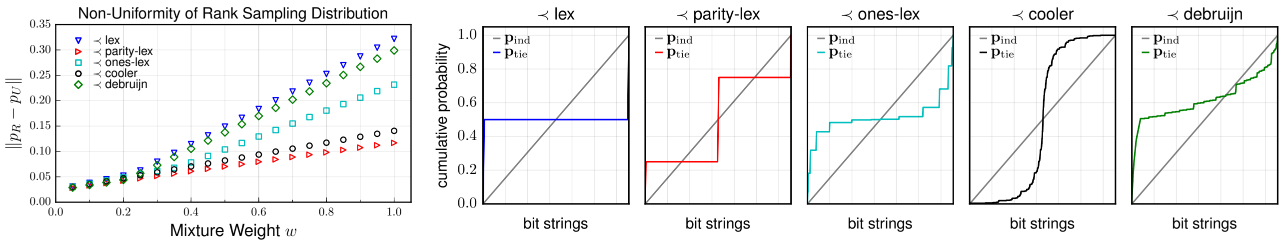
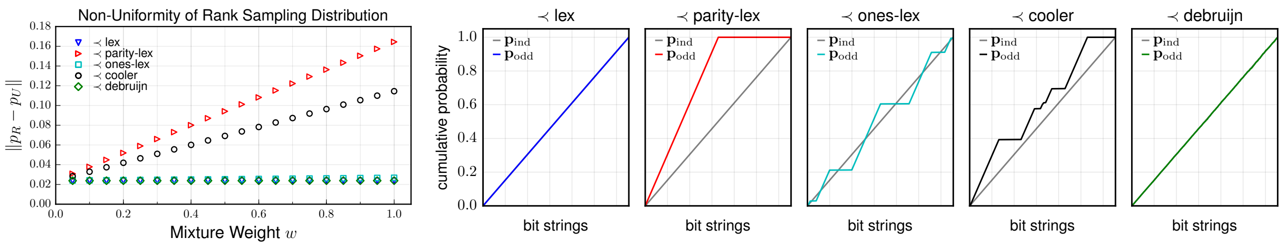


Figure 2: The left panel shows a pair $(\mathbf{p}, \mathbf{q})$ of reflected, bimodal Poisson distributions with slight location shift. The right plot compares the power of testing $\mathbf{p} = \mathbf{q}$ using the SRS (for various choices of m) to several baseline methods.



(a) $\mathbf{p} \coloneqq \mathbf{p}_{\text{ind}} \qquad \mathbf{q} \coloneqq w\mathbf{p}_{\text{tie}} + (1-w)\mathbf{p}_{\text{ind}}$



(b) $\mathbf{p} \coloneqq \mathbf{p}_{\text{ind}} \qquad \mathbf{q} \coloneqq w\mathbf{p}_{\text{odd}} + (1-w)\mathbf{p}_{\text{ind}}$

Figure 3: In each of the two panels (a) and (b), the left plot shows the sup-norm distance between the sampling distribution of the rank statistic and the discrete uniform (using Eq. 8 in Theorem 3.8), for a uniform null $\mathbf{p} \coloneqq \mathbf{p}_{\text{ind}}$ on $\{0,1\}^{16}$ against alternative distributions of the form $\mathbf{q} \coloneqq w\mathbf{p}_{\text{alt}} + (1-w)\mathbf{p}_{\text{ind}}$, for increasing mixture weight $0 \le w \le 1$ and six different orderings on the binary strings. The right plot compares the cumulative distribution function of the null distribution (diagonal line in gray) with the cumulative distribution functions of the alternative distribution (when $w = 1$) as obtained by sorting the binary strings according to each ordering. Orderings which induce a greater distance between the cumulative distribution functions of the null and alternative distributions result in more power to detect the alternative.

4.1 Bimodal, Symmetric Poisson

We first investigate the performance of the SRS for testing a pair of symmetric, multi-modal distributions over the integers with location shift. In particular, for $x \in \mathbb{Z}$, define distribution $\mathbf{f}(x; \lambda_1, \lambda_2) \coloneqq \frac{1}{2}\left(\frac{1}{2}\text{Poisson}(|x|; \lambda_1) + \frac{1}{2}\text{Poisson}(|x|; \lambda_2)\right)$. Note $\mathbf{f}$ is a mixture of Poisson distributions with rates λ_1 and λ_2, reflected symmetrically about $x = 0$. We set $\mathbf{p}(x) \coloneqq \mathbf{f}(x; 10, 20)$ and $\mathbf{q} \coloneqq \mathbf{f}(x; 10, 25)$ so that $\mathbf{q}$ is location-shifted in two of the four modes (Figure 2, left panel).

The right plot of Figure 2 compares the power for various sample sizes n from $\mathbf{q}$ according to the SRS ($m = 1, 2, 3, 30$, shown in increasing shades of gray) and several baselines (shown in color). The baselines (AD, MMD, KS, and Mann–Whitney U) are used to assess goodness-of-fit by performing a two-sample test on n samples from $\mathbf{q}$ with samples drawn i.i.d. from $\mathbf{p}$. The power (at level $\alpha = 0.05$) is estimated as the fraction of correct answers over 1024 independent trials. The Mann–Whitney U, which is also based on rank statistics with a correction for ties, has no power for all n as it can only detect median shift, as does the SRS with $m = 1$ (see Corollary 3.5). The SRS becomes non-uniform for $m = 2$ although this choice results in low power. The SRS with $m = 3$ has comparable power to the AD and MMD tests. The SRS with $m = 30$ is the most powerful, although it requires more computational effort and samples from $\mathbf{p}$ (Algorithm 1 scales as $O(mn)$).

4.2 Binary strings

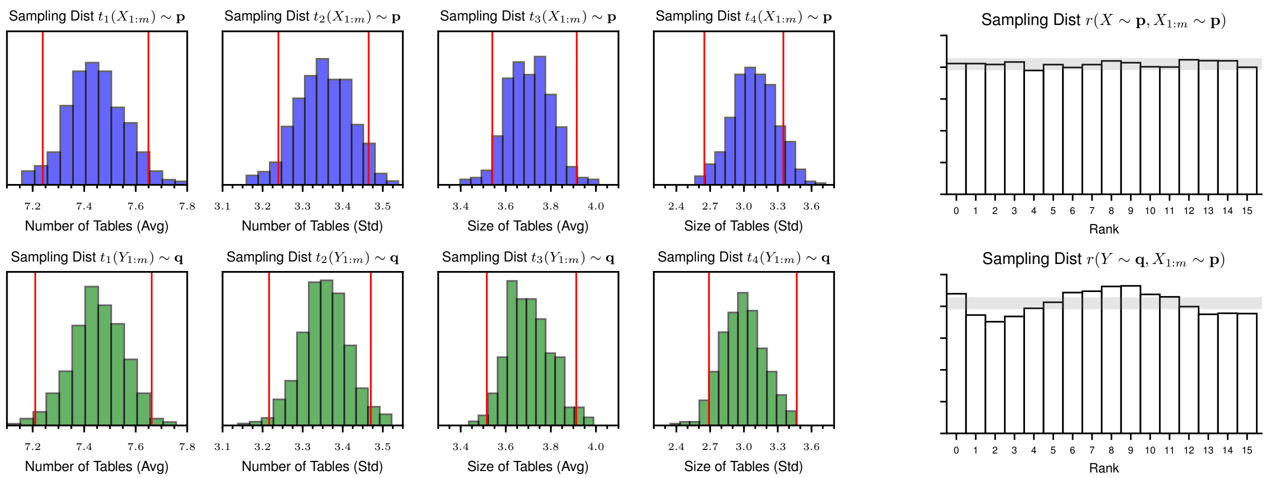
Let $\mathcal{T} \coloneqq \{0,1\}^k$ be the set of all length k binary strings. Define the following distributions to be uniform over all strings $x = (x_1, \dots, x_k) \in \{0,1\}^k$ which satisfy the given predicates:

$$\begin{aligned}
\mathbf{p}_{\text{ind}} &: \text{ uniform on all strings,} \\
\mathbf{p}_{\text{odd}} &: \textstyle\sum_{i=1}^{k} x_i \equiv 1 \pmod 2, \\
\mathbf{p}_{\text{tie}} &: x_1 = x_2 = \cdots = x_{k/2}.
\end{aligned}$$

Each of these distributions assigns marginal probability 1/2 to each bit x_i (for $1 \le i \le k$), so all deviations from the uniform distribution $\mathbf{p}_{\text{ind}}$ are captured by higher-order relationships. The five orderings used for comparing binary strings are

$$\begin{aligned}
\prec_{\text{lex}} &: \text{Lexicographic (dictionary) ordering,} \\
\prec_{\text{par}} &: \text{Parity of ones, ties broken using } \prec_{\text{lex}}, \\
\prec_{\text{one}} &: \text{Number of ones, ties broken using } \prec_{\text{lex}}, \\
\prec_{\text{coo}} &: \text{Cooler ordering (randomly generated) [30],} \\
\prec_{\text{dbj}} &: \text{De Bruijn sequence ordering.}
\end{aligned}$$

We set the null distribution $\mathbf{p} \coloneqq \mathbf{p}_{\text{ind}}$ and construct alternative distributions $\mathbf{q} \coloneqq w\mathbf{p}_{\text{c}} + (1-w)\mathbf{p}_{\text{ind}}$ as mixtures of $\mathbf{p}_{\text{ind}}$ with the other two distributions, where $w \in [0,1]$ and $c \in \{\text{odd}, \text{not}\}$. We take bit strings of length $k = 16$ with $n = 256$ observations so that $|\mathcal{T}| = 65{,}536$ and 0.4% of the domain size is observed.



(a) Sampling distribution of four different probe statistics $\{t_1, t_2, t_3, t_4\}$ of a dataset of partitions, as sampled from $\mathbf{p}$ (Eq. (9); blue) and from $\mathbf{q}$ (Eq. (10); green) estimated by Monte Carlo simulation. Vertical red lines indicate 2.5% and 97.5% quantiles. Even though $\mathbf{p} \neq \mathbf{q}$, the distributions of these statistics are aligned in such a way that a statistic $t_j(Y_{1:m}) \sim \mathbf{q}$ is unlikely to appear as an extreme value in the sampling distribution of the corresponding statistic $t_j(X_{1:m}) \sim \mathbf{p}$, which leads to under-powered resampling-based tests.

(b) Monte Carlo simulation of the rank statistic illustrates its significant uniform distribution under the null hypothesis (top) and significant non-uniform distribution under the alternative hypothesis (bottom).

Figure 4: Comparison of the sampling distribution of (a) various bootstrapped probe statistics [11] with (b) the stochastic rank statistic, for goodness-of-fit testing the Chinese restaurant processes on $N = 20$ customers. Discussion in main text.

Figure 3 shows how the non-uniformity of the SRS (computed using Theorem 3.8) varies for each of the two alternatives and five orderings ($m = 6$). Each ordering induces a different CDF over $\{0,1\}^k$ for the alternative distribution, shown in the right panel for $w = 1$. Orderings with a greater maximum vertical distance between the null and alternative CDF attain greater rank non-uniformity. No single ordering is more powerful than all others in both test cases. However, in each case, some ordering detects the difference even at low weights w, despite the sparse observation set.

The alternative $\mathbf{q} = \mathbf{p}_{\text{odd}}$ in Figure 3b is especially challenging: in a sample, all substrings (not necessarily contiguous) of a given length $j < k$ are equally likely. Even though the SRS is non-uniform for all orderings, the powers vary significantly. For example, comparing strings using $\prec_{\text{lex}}$ does not effectively distinguish between $\mathbf{p}_{\text{ind}}$ and $\mathbf{p}_{\text{odd}}$, as strings with an odd number of ones are lexicographically evenly interspersed within the set of all strings. The parity ordering (which is optimal for this alternative) and the randomly generated cooler ordering have increasing power as w increases.

4.3 Partition testing

We next apply the SRS to test distributions on the space of partitions of the set $\{1, 2, \ldots, N\}$. Let Π_N denote the set of all such partitions. We define a distribution on Π_N using the two-parameter Chinese Restaurant Process (CRP) [6, Section 5.1]. Letting $(x|y)_N \coloneqq (x)(x+y)\cdots(x+(N-1)y)$, the probability of a partition $\pi \coloneqq \{\pi_1, \ldots, \pi_k\} \in \Pi_N$ with k tables (blocks) is given by

$$\mathsf{CRP}(\pi; a, b) \coloneqq \begin{cases} \frac{(b|a)_k}{(b|1)_N} \prod_{i=1}^{k} (1-a)_{c_k - 1} & (\text{if } a > 0) \\ \frac{b^k}{(b|1)_N} \prod_{i=1}^{k} (c_k - 1)! & (\text{if } a = 0), \end{cases}$$

where c_i is the number of customers (integers) at table π_i ($1 \le i \le k$). Simulating a CRP proceeds by sequentially assigning customers to tables [6, Def. 7]. Even though we can compute the probability of any partition, the cardinality of Π_N grows exponentially in N (e.g., $|\Pi_{20}| \approx 5.17 \times 10^{13}$). The expected frequency of any partition is essentially zero for sample size $n \ll |\Pi_N|$, so Pearson chi-square or likelihood-ratio tests on the raw data are inappropriate. Algorithm 2 defines a total order on the partition domain Π_N.

Algorithm 2 Total order $\prec$ on the set of partitions Π_N

Input: $\begin{cases} \text{Partition } \pi \coloneqq \{\pi_1, \pi_2, \ldots, \pi_k\} \in \Pi_N \text{ with } k \text{ blocks.} \\ \text{Partition } \nu \coloneqq \{\nu_1, \nu_2, \ldots, \nu_l\} \in \Pi_N \text{ with } l \text{ blocks.} \end{cases}$

Output: LT if $\pi \prec \nu$; GT if $\pi \succ \nu$; EQ if $\pi = \nu$.

```
if k < l then return LT                                  ▷ ν has more blocks
if k > l then return GT                                  ▷ π has more blocks
π̃ ← blocks of π sorted by value of least element in the block
ν̃ ← blocks of ν sorted by value of least element in the block
for b = 1, 2, ..., l do
    if |π̃_b| < |ν̃_b| then return LT                     ▷ ν̃_b has more elements
    if |π̃_b| > |ν̃_b| then return GT                     ▷ π̃_b has more elements
    π'_b ← values in π̃_b sorted in ascending order
    ν'_b ← values in ν̃_b sorted in ascending order
    for i = 1, 2, ..., |π'_b| do
        if π'_{b,i} < ν'_{b,i} then return LT            ▷ π'_b has smallest element
        if π'_{b,i} > ν'_{b,i} then return GT            ▷ ν'_b has smallest element
return EQ
```

We consider the following pair of distributions:

$$\mathbf{p} := \mathsf{CRP}(0.26, 0.76)/2 + \mathsf{CRP}(0.19, 5.1)/2 \quad (9)$$
$$\mathbf{q} := \mathsf{CRP}(0.52, 0.52). \quad (10)$$

These distributions are designed to ensure that partitions from $\mathbf{p}$ and $\mathbf{q}$ have similar distributions on the number and sizes of tables. Figure 4a shows a comparison of using Monte Carlo simulation of various bootstrapped probe statistics for assessing goodness-of-fit versus using the SRS with the ordering in Algorithm 2.

In Figure 4a, each probe statistic takes a size m dataset $X_{1:m}$ (where each X_i is a partition) and produces a numerical summary such as the average of the number of tables in each sample. A resampling test [11] that uses these probe statistics will report (with high probability) that an observed statistic $t(Y_{1:m}) \sim \mathbf{q}$ drawn from the alternative distribution is a non-extreme value in the null distribution $t(X_{1:m}) \sim \mathbf{p}$ (as indicated by alignment of their quantiles, shown in red) and will therefore have insufficient evidence to reject $\mathbf{p} = \mathbf{q}$.

On the other hand, Figure 4b shows that when ranked using the ordering obtained from Algorithm 2 (which is based on a multivariate combination of the univariate probe statistics in Figure 4a specified procedurally), a partition $Y \sim \mathbf{q}$ is more likely to lie in the center of a dataset $X_{1:m} \sim^{\text{iid}} \mathbf{p}$, as illustrated by the non-uniform rank distribution under the alternative hypothesis (the gray band shows 99% variation for a uniform histogram). By comparing the top and bottom panels of Figure 4b, the SRS shows that partitions from $\mathbf{q}$ have a poor fit with respect to partitions from $\mathbf{p}$, despite their agreement on multiple univariate summary statistics shown in Figure 4a.

5 Applications

We next apply the proposed test to assess the sample quality of random data structures obtained from approximate sampling algorithms over combinatorially large domains with intractable probabilities.

5.1 Dirichlet process mixture models

The recent paper [31] describes simulation-based calibration (SBC), a procedure for validating samples from algorithms that can generate posterior samples for a hierarchical Bayesian model. More specifically, for a prior $\pi(z)$ over the parameters z and likelihood function $\pi(x|z)$ over data x, integrating the posterior over the joint distribution returns the prior distribution:

$$\pi(z) = \int [\pi(z|x')\pi(x'|z')\mathrm{d}x']\, \pi(z')\mathrm{d}z'. \quad (11)$$

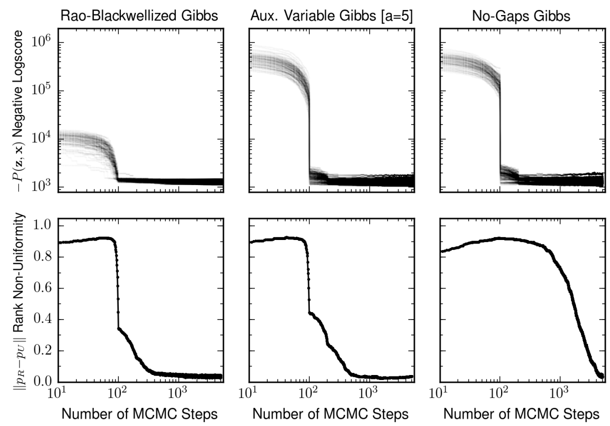
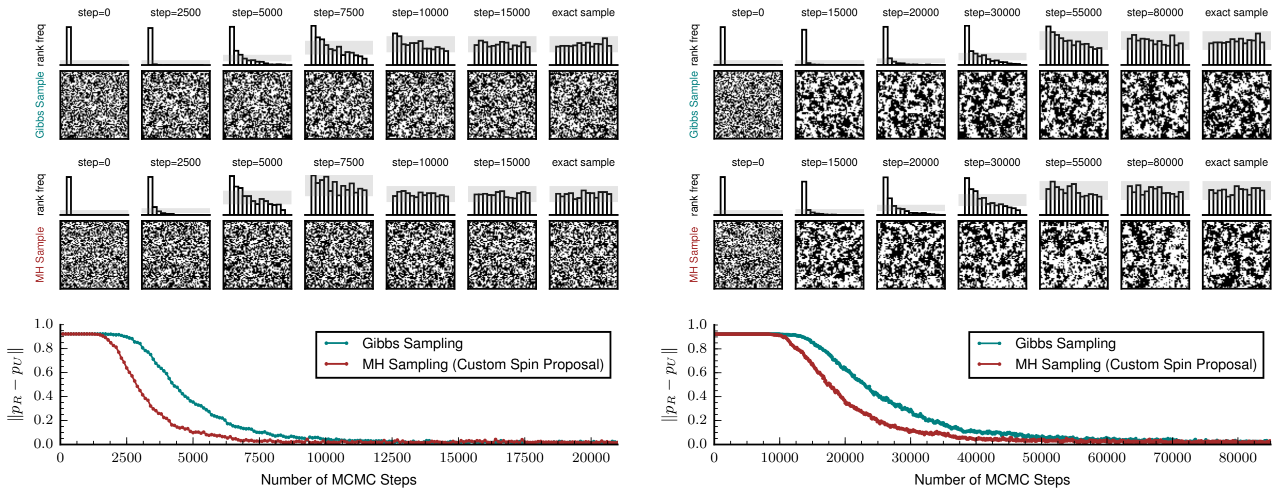


Figure 5: The uniformity of the SRS (bottom row) captures convergence behavior of MCMC sampling algorithms for Dirichlet process mixture models that are not captured by standard diagnostics such as the logscore (top row).

Eq. (11) indicates that by simulating n datasets $\{x_1, \dots, x_n\}$ i.i.d. from the marginal distribution, samples $\{\hat{z}_1, \dots, \hat{z}_n\}$ (where $z_i \approx \pi(z|x_i)$) from an approximate posterior should be i.i.d. samples from the prior $\pi(z)$. An approximate sampler can be thus be diagnosed by performing a goodness-of-fit test to check whether $\hat{z}_{1:n}$ are distributed according to π. Ranks of univariate marginals of a continuous parameter vector $z \in \mathbb{R}^d$ are used in [31]. We extend SBC to handle discrete latent variables z taking values in a large domain.

We sampled $n = 1000$ datasets $\{x_1, \dots, x_n\}$ independently from a Dirichlet process mixture model. Each dataset x_i has $k = 100$ observations and each observation is five-dimensional (i.e., $x_i \in \mathbb{R}^{k\times 5}$) with a Gaussian likelihood. From SBC, samples $\hat{z}_{1:n}$ (where $z_i \in \Pi_k$ and $|\Pi_k| \approx 10^{115}$) of the mixture assignment vector should be distributed according to the CRP prior $\pi(z)$. The top row of Figure 5 shows trace plots of the logscore (unnormalized posterior) of approximate samples from Rao–Blackwellized Gibbs, Auxiliary Variable Gibbs, and No-Gaps Gibbs samplers (Algorithms 3, 8, and 4 in [21]). Each line corresponds to an independent run of MCMC inference. The bottom row shows the evolution of the uniformity of the SRS using $m = 64$ and the ordering on partitions from Algorithm 2.

While logscores typically stabilize after 100 MCMC steps (one epoch through all observations in a dataset) and suggest little difference across the three samplers, the SRS shows that Rao-Blackwellized Gibbs is slightly more efficient than Auxiliary Variable Gibbs and that the sample quality from No-Gaps Gibbs is inferior to those from the other two algorithms up until roughly 5,000 steps. These results are consistent with the observation from [21] that No-Gaps has inefficient mixing (it excessively rejects proposals on singleton clusters).



(a) Temperature $T = 8$

(b) Temperature $T = 3$

Figure 6: Assessing the goodness-of-fit of approximate samples of a 64×64 Ising model for Gibbs sampling and Metropolis–Hastings sampling (with the custom spin proposal from [18]) at two temperatures using the SRS. In both cases, the SRS converges to its uniform distribution more rapidly for samples obtained from MH than for those from Gibbs sampling.

5.2 Ising models

In this application we use the SRS to assess the sample quality of approximate Ising model simulations. For a ferromagnetic $k \times k$ lattice with temperature T, the probability of a spin configuration $x \in \{-1, +1\}^{k \times k}$ is

$$P(x) \propto \exp\left(-1/T \sum_{i,j} x_i x_j\right). \qquad (12)$$

While Eq. (12) is intractable to compute for any x due to the unknown normalization constant, coupling-from-the-past [25] is a popular MCMC technique which can tractably obtain exact samples from the Ising model. For a 64×64 Ising model (domain size $2^{64 \times 64}$), we obtained 650 exact samples using coupling-from-the-past, and used these "ground-truth" samples to assess the goodness-of-fit of approximate samples obtained via Gibbs sampling and Metropolis–Hastings sampling (with a custom spin proposal [18, Section 31.1]).

For each temperature $T=3$ and $T=8$, we obtained 7,800 approximate samples using MH and Gibbs. The first two rows of Figure 6 each show the evolution of one particular sample (Gibbs, top; MH, bottom). Two exact samples are shown in the final column of each panel. All approximate and exact samples are independent of one another, obtained by running parallel Markov chains. The SRS of the exact samples with respect to the approximate samples was taken at checkpoints of 100 MCMC steps, using $m=12$ and an ordering based on the Hamiltonian energy, spin magnetization, and connected components. SRS histograms (and 99% variation bands) evolving at various steps are shown above the Ising model renderings.

The SRS is non-uniform (including in regimes where the difference between approximate and exact samples is too fine-grained to be detected visually) at early steps and more uniform at higher steps. The plots show that MH is a more efficient sampler than Gibbs at moderate temperatures, as its sample quality improves more rapidly. This characteristic was conjectured in [18], which noted that the MH sampler "has roughly double the the probability of accepting energetically unfavourable moves, so may be a more efficient sampler [than Gibbs]". In addition, the plots suggest that the samples become close to exact (in terms of their joint energy, magnetization, and connected components characteristics) after 20,000 steps for $T=8$ and 100,000 steps for $T=3$, even though obtaining exact samples using coupling-from-the-past requires between 500,000 and 1,000,000 MCMC steps for both temperatures.

6 Conclusion

This paper has presented a flexible, simple-to-implement, and consistent goodness-of-fit test for discrete distributions. The test statistic is based on the ranks of observed samples with respect to new samples from the candidate distribution. The key insight is to compute the ranks using an ordering on the domain that is able to detect differences in properties of interest in high dimensions. Unlike most existing statistics, the SRS is distribution-free and has a simple exact sampling distribution. Empirical studies indicate that the SRS is a valuable addition to the practitioner's toolbox for assessing sample quality in regimes which are not easily handled by existing methods.

Acknowledgments

The authors thank the anonymous referees for their helpful feedback. This research was supported by the DARPA SD2 program (contract FA8750-17-C-0239); the Ethics and Governance of Artificial Intelligence Initiative of the MIT Media Lab and Harvard's Berkman Klein Center; the Systems That Learn Initiative of MIT CSAIL; and an anonymous philanthropic gift.

References

[1] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pages 3591–3599. Curran Associates, 2015.

[2] Mohammad Ahsanullah, Valery B. Nevzorov, and Mohammad Shakil. *An Introduction to Order Statistics*. Atlantis Studies in Probability and Statistics. Atlantis Press, 2013.

[3] Erling B. Andersen. A goodness of fit test for the Rasch model. *Psychometrika*, 38(1):123–140, 1973.

[4] Taylor B. Arnold and John W. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2), 2011.

[5] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269. IEEE, 2000.

[6] Wray Buntine and Marcus Hutter. A Bayesian view of the Poisson–Dirichlet process. *arXiv preprint*, (arXiv:1007.0296), 2010.

[7] V. Choulakian, R. A. Lockhart, and M. A. Stephens. Cramér–von Mises statistics for discrete distributions. *Canadian Journal of Statistics*, 22 (1):125–137, 1994.

[8] W. J. Conover. A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67(339):591–596, 1972.

[9] J. Dehardt. Generalizations of the Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, 42(6):2050–2055, 1971.

[10] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[11] Phillip I. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. Springer Series in Statistics. Springer, 2004.

[12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(138):723–773, 2012.

[13] Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, 36(2):369–401, 1965.

[14] Susan Dadakis Horn. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, 33(1):237–247, 1977.

[15] Erich L. Lehmann and Howard J. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day Series in Probability and Statistics. Holden-Day, 1975.

[16] Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3rd edition, 2005.

[17] Cathryn M. Lewis and Jo Knight. Introduction to genetic association studies. In Ammar Al-Chalabi and Laura Almasy, editors, *Genetics of Complex Human Diseases: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, 2009.

[18] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[19] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[20] Alberto Maydeu-Olivares and Carlos Garcia-Forero. Goodness-of-fit testing. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education*, volume 7, pages 190–196. Elsevier, 2010.

[21] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[22] J. A. Peacock. Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, 202(3):615–627, 1983.

[23] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 5:157–175, 1900.

[24] Anthony N. Pettitt and Michael A. Stephens. The Kolmogorov–Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2): 205–210, 1977.

[25] James G. Propp and David B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(2):223–252, 1996.

[26] Timothy R. C. Read and Noel A. C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Series in Statistics. Springer, 1988.

[27] Walter Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 1976.

[28] D. S. Starnes, D. Yates, and D. S. Moore. *The Practice of Statistics*. W. H. Freeman and Company, 2010.

[29] Michael Steele and Janet Chaseling. Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions. *Communications in Statistics—Simulation and Computation*, 35(4):1067–1075, 2006.

[30] Brett Stevens and Aaron Williams. The coolest order of binary strings. In *Proceedings of the 6th International Conference on Fun with Algorithms (FUN)*, pages 322–333. Springer, 2012.

[31] Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint*, (arXiv:1804.06788), 2018.

[32] Gregory Valiant and Paul Valiant. Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 685–694. ACM, 2011.

[33] D. A. Williams. Improved likelihood ratio tests for complete contingency tables. *Biometrika*, 63 (1):33–37, 1976.

[34] Michael Williams. How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics. *Journal of Instrumentation*, 5(09): P09004, 2010.

[35] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 5561–5570. PMLR, 2018.

A Appendix: Proofs

A.1 Uniformity of rank

Throughout this appendix, let $\mathcal{T}$ be a non-empty finite or countably infinite set, let $\prec$ be a total order on $\mathcal{T}$ (of any order type), and let $\mathbf{p}$ and $\mathbf{q}$ each be a probability distribution on $\mathcal{T}$. For $n \in \mathbb{N}$, let $[n]$ denote the set $\{0, 1, 2, \ldots, n-1\}$.

Given a positive integer m, define the following random variables:

$$X_0 \sim \mathbf{q} \tag{13}$$

$$U_0 \sim \mathsf{Uniform}(0,1) \tag{14}$$

$$X_1, X_2, \ldots, X_m \sim^{\mathrm{iid}} \mathbf{p} \tag{15}$$

$$U_1, U_2, \ldots, U_m \sim^{\mathrm{iid}} \mathsf{Uniform}(0,1) \tag{16}$$

$$R = \textstyle\sum_{j=1}^{m} \mathbb{I}\left[X_j \prec X_0\right] + \mathbb{I}\left[X_j = X_0, U_j < U_0\right]. \tag{17}$$

Our first main result is the following, which establishes necessary and sufficient conditions for uniformity of the rank statistic.

Theorem A.1 (Theorem 3.1 in the main text)**.** *We have $\mathbf{p} = \mathbf{q}$ if and only if for all $m \geq 1$, the rank statistic R is uniformly distributed on $[m+1] \coloneqq \{0, 1, \ldots, m\}$.*

Before proving Theorem A.1, we state and prove several lemmas. We begin by showing that an i.i.d. sequence yields a uniform rank distribution.

Lemma A.2. *Let $T_0, T_1, \ldots, T_m$ be an i.i.d. sequence of random variables. If $\Pr\{T_i = T_j\} = 0$ for all distinct i and j, then the rank statistics $S_i \coloneqq \sum_{j=0}^{m} \mathbb{I}[T_j \prec T_i]$ for $0 \leq i \leq m$ are each uniformly distributed on $[m+1]$.*

Proof. Since $T_0, T_1, \ldots, T_m$ is i.i.d., it is a finitely exchangeable sequence, and so the rank statistics $S_0, \ldots, S_m$ are identically (but not independently) distributed.

Fix an arbitrary $k \in [m+1]$. Then $\Pr\{S_i = k\} = \Pr\{S_j = k\}$ for all $i, j \in [m+1]$. By hypothesis, $\Pr\{T_i = T_j\} = 0$ for distinct i and j. Therefore the rank statistics are almost surely distinct, and the events $\{S_i = j\}$ (for $0 \leq i \leq m$) are mutually exclusive and exhaustive. Since these events partition the outcome space, their probabilities sum to 1, and so $\Pr\{S_i = k\} = 1/(m+1)$ for all $i \in [m+1]$.

Because k was arbitrary, S_i is uniformly distributed on $[m+1]$ for all $i \in [m+1]$. ☐

We will also use the following result about convergence of discrete uniform variables to a continuous uniform random variable.

Lemma A.3. *Let $(V_m)_{m\geq 1}$ be a sequence of discrete random variables such that V_m is uniformly distributed on $\{0, 1/m, 2/m, \ldots, 1\}$, and let U be a continuous random variable uniformly distributed on the interval $[0,1]$. Then $(V_m)_{m\geq 1}$ converges in distribution to U, i.e.,*

$$\lim_{m\to\infty} \Pr\{V_m < u\} = \Pr\{U < u\} = u. \tag{18}$$

for all $u \in [0,1]$.

Furthermore, the convergence (18) is uniform in u.

Proof. Let $\epsilon > 0$. The distribution function F_m of V_m is given by

$$F_m(u) = \begin{cases} 1/(m+1) & u \in [0, 1/m) \\ 2/(m+1) & u \in [1/m, 2/m) \\ \cdots & \\ (a+1)/(m+1) & u \in [a/m, (a+1)/m) \\ \cdots & \\ m/(m+1) & u \in [(m-1)/m, 1) \\ 1 & u = 1. \end{cases}$$

Observe that for $0 \leq a < m$, the value $F_m(u)$ lies in the interval $[a/m, (a+1)/m)$ since we have that $(a/m) < (a+1)/(m+1) < (a+1)/m$. Since u is also in this interval, $|F_m(u) - u| \leq (a+1)/m - a/m = 1/m < \epsilon$ whenever $m > 1/\epsilon$, for all u. ☐

The following intermediate value lemma for step functions on the rationals is straightforward. It makes use of sums defined over subsets of the rationals, which are well-defined, as we discuss in the next remark.

Lemma A.4. *Let $p\colon (\mathbb{Q} \cap [0,1]) \to [0,1]$ be a function satisfying $p(0) = 0$ and $\sum_{x\in\mathbb{Q}\cap[0,1]} p(x) = 1$. Then for each $\delta \in (0,1)$, there is some $w \in \mathbb{Q} \cap [0,1]$ such that*

$$\sum_{x\in\mathbb{Q}\cap(0,w)} p(x) \leq \delta \leq \sum_{x\in\mathbb{Q}\cap(0,w]} p(x).$$

Remark A.5. The infinite sums in Lemma A.4 taken over a subset of the rationals can be formally defined as follows: Consider an arbitrary enumeration $\{q_1, q_2, \ldots, q_n, \ldots\}$ of $\mathbb{Q} \cap [0,1]$, and define the summation over the integer-valued index $n \geq 1$. Since the series consists of positive terms, it converges absolutely, and so all rearrangements of the enumeration converge to the same sum (see, e.g., [27, Theorem 3.55]).

One can show that the Cauchy criterion holds in this setting. Namely, suppose that a sum $\sum_{a<x<c} p(x)$ of non-negative terms converges. Then for all $\epsilon > 0$ there is some rational $b \in (a,c)$ such that $\sum_{a<x\leq b} p(x) < \epsilon$.

We now prove both directions of Theorem A.1.

Proof of Theorem A.1. Because $\mathcal{T}$ is countable, by a standard back-and-forth argument the total order $(\mathcal{T}, \prec)$ is isomorphic to $(B, <)$ for some subset $B \subseteq \mathbb{Q} \cap (0,1)$. Without loss of generality, we may therefore take $\mathcal{T}$ to be $\mathbb{Q} \cap [0,1]$ and assume that $\mathbf{p}(0) = \mathbf{p}(1) = 0$.

Consider the unit square $[0,1]^2$ equipped with the dictionary order $\lhd_{\mathrm{d}}$. This is a total order with the least upper bound property. For each $i \in [m+1]$, define $T_i \coloneqq (X_i, U_i)$, which takes values in $[0,1]^2$, and observe that the rank R in Eq. (6) of Theorem A.1 is equivalent to the rank $\sum_{i=0}^{m} \mathbb{I}[T_i \lhd_{\mathrm{d}} T_0]$ of T_0 taken according to the dictionary order.

(Necessity) Suppose $\mathbf{p} = \mathbf{q}$. Then $T_0, \ldots, T_m$ are independent and identically distributed. Since $U_0, \ldots, U_m$ are continuous random variables, we have $\Pr\{T_i = T_j\} = 0$ for all $i \neq j$. Apply Lemma A.2.

(Sufficiency) Suppose that for all $m > 0$, we have that the rank R is uniformly distributed on $\{0, 1, 2, \ldots, m\}$. We begin the proof by first constructing a distribution function $F_{\mathbf{p}}$ on the unit square and then establishing several of its properties. First let $\tilde{\mathbf{p}} \colon [0,1] \to [0,1]$ be the "left-closed right-open" cumulative distribution function of $\mathbf{p}$, defined by

$$\tilde{\mathbf{p}}(x) \coloneqq \sum_{y \in \mathbb{Q} \cap [0,x)} \mathbf{p}(y)$$

for $x \in [0,1]$. Define $\mathbf{p}'$ to be the probability measure on $[0,1]$ that is equal to $\mathbf{p}$ on subsets of $\mathbb{Q} \cap [0,1]$ and is null elsewhere, and define the distribution function $F_{\mathbf{p}} \colon [0,1]^2 \to [0,1]$ on S by

$$F_{\mathbf{p}}(x,u) \coloneqq \tilde{\mathbf{p}}(x) + u\mathbf{p}'(x)$$

for $(x,u) \in [0,1]^2$. To establish that $F_{\mathbf{p}}$ is a valid distribution function, we show that its range is $[0,1]$; it is monotonically non-decreasing in each of its variables; and it is right-continuous in each of its variables.

It is immediate that $F_{\mathbf{p}}(0,0) = 0$ and $F_{\mathbf{p}}(1,1) = 1$. Furthermore, To establish that $F_{\mathbf{p}}$ is monotonically non-decreasing, put $x < y$ and $u < v$ and observe that

$$\begin{aligned} F_{\mathbf{p}}(x,u) &= \tilde{\mathbf{p}}(x) + u\mathbf{p}'(x) \\ &\leq \tilde{\mathbf{p}}(x) + \mathbf{p}'(x) \\ &\leq \textstyle\sum_{z \in \mathbb{Q} \cap [0,y)} \mathbf{p}'(z) \\ &= \tilde{\mathbf{p}}(y) \\ &\leq F_{\mathbf{p}}(y,u) \end{aligned}$$

and

$$\begin{aligned} F_{\mathbf{p}}(x,u) &= \tilde{\mathbf{p}}(x) + u\mathbf{p}'(x) \\ &\leq \tilde{\mathbf{p}}(x) + v\mathbf{p}'(x) \\ &= F_{\mathbf{p}}(x,v). \end{aligned}$$

We now establish right-continuity. For fixed x, $F_{\mathbf{p}}(x,u)$ is a linear function of u and so continuity is immediate. For fixed u, we have shown that $F_{\mathbf{p}}(x,u)$ is non-decreasing so it is sufficient to show that for any x and for any $\epsilon > 0$ there exists $x' > x$ such that

$$\begin{aligned} \epsilon &> F(x',u) - F(x,u) \\ &= \tilde{\mathbf{p}}(x') + u\mathbf{p}'(x') - \tilde{\mathbf{p}}(x) - u\mathbf{p}(x) \\ &= \tilde{\mathbf{p}}(x') + u\mathbf{p}'(x') - \tilde{\mathbf{p}}(x) - u\mathbf{p}(x) \\ &= \sum_{y \in \mathbb{Q} \cap [x,x']} \mathbf{p}(y), \end{aligned}$$

which is immediate from the Cauchy criterion.

Finally, we note that Lemma A.4 and the continuity of $F_{\mathbf{p}}$ in u together imply that $F_{\mathbf{p}}$ obtains all intermediate values, i.e., for any $\delta \in [0,1]$ there is some (x,u) such that $F(x,u) = \delta$.

Next define the inverse $F_{\mathbf{p}}^{-1} \colon [0,1] \to [0,1]^2$ by

$$F_{\mathbf{p}}^{-1}(s) \coloneqq \inf\{(x,u) \mid F_{\mathbf{p}}(x,u) = s\} \tag{19}$$

for $s \in [0,1]$, where the infimum is taken with the respect to the dictionary order $\lhd_{\mathrm{d}}$. The set in Eq (19) is non-empty since $F_{\mathbf{p}}$ obtains all values in $[0,1]$. Moreover, $F_{\mathbf{p}}^{-1}(s) \in [0,1]^2$ since $\lhd_{\mathrm{d}}$ has the least upper bound property. (This "generalized" inverse is used since $F_{\mathbf{p}}$ is one-to-one only under the stronger assumption that $\mathbf{p}(x) > 0$ for all $x \in \mathbb{Q} \cap (0,1)$.) Analogously define $F_{\mathbf{q}}$ in terms of $\mathbf{q}$.

Now define the rank function

$$r(a_0, \{a_1, \ldots, a_m\}) \coloneqq \sum_{i=0}^{m} \mathbb{I}[a_i < a_0]$$

and note that $R \equiv r(T_0, \{T_1, \ldots, T_m\})$. By the hypothesis, $r(T_0, \{T_1, \ldots, T_m\})/m$ is uniformly distributed on $\{0, 1/m, 2/m, \ldots, 1\}$ for all $m > 0$. Applying Lemma A.3 gives

$$\begin{aligned} &\lim_{m \to \infty} \Pr\left\{ \frac{1}{m} \tilde{r}(T_0, \{T_1, \ldots, T_m\}) < s \right\} \\ &\qquad = \Pr\{U_0 < s\} \\ &\qquad = s. \end{aligned} \tag{20}$$

for $s \in [0,1]$.

For any $t \in [0,1]$ and $m \geq 1$, the random variable $\hat{F}_{\mathbf{p}}^{m}(t) \coloneqq \tilde{r}(t, \{T_1, \ldots, T_m\})/m$ is the empirical distribution of $F_{\mathbf{p}}$. Therefore, by the Glivenko–Cantelli theorem for empirical distribution functions on k-dimensional Euclidean space [9, Corollary of Theorem 4], the sequence of random variables $(\hat{F}_{\mathbf{p}}^{m}(t))_{m \geq 1}$ converges a.s. to the real number $F_{\mathbf{p}}(t)$ uniformly in t, Hence the sequence $(\hat{F}_{\mathbf{p}}^{m}(T_0))_{m \geq 1}$ converges a.s. to the random variable $\hat{F}_{\mathbf{p}}(T_0)$, so that for any $s \in [0,1]$,

$$\begin{aligned}
&\lim_{m\to\infty} \Pr\left\{\frac{1}{m}\tilde{r}(T_0, \{T_1,\dots,T_m\}) < s\right\} \\
&\quad= \lim_{m\to\infty} \Pr\left\{\hat{F}^m_{\mathbf{p}}(T_0) < s\right\} && (21)\\
&\quad= \Pr\left\{F_{\mathbf{p}}(T_0) < s\right\} && (22)\\
&\quad= \Pr\left\{T_0 \lhd_{\mathrm{d}} F^{-1}_{\mathbf{p}}(s)\right\} && (23)\\
&\quad= F_{\mathbf{q}}(F^{-1}_{\mathbf{p}}(s)). && (24)
\end{aligned}$$

The interchange of the limit and the probability in Eq. (22) follows from the bounded convergence theorem, since $\hat{F}^m_{\mathbf{p}}(T_0) \to F_{\mathbf{p}}(T_0)$ a.s. and for all $m \ge 1$ we have $|\hat{F}^m_{\mathbf{p}}(T_0)| \le 1$ a.s.

Combining Eq. (20) and Eq. (24), we see that

$$F_{\mathbf{q}}(F^{-1}_{\mathbf{p}}(s)) = s \implies F^{-1}_{\mathbf{p}}(s) = F^{-1}_{\mathbf{q}}(s),$$

for $s \in [0,1]$. Since $0 \le F_{\mathbf{p}}(x,u) \le 1$, for each $(x,u) \in [0,1]^2$ we have

$$\begin{aligned}
F^{-1}_{\mathbf{q}}(F_{\mathbf{p}}(x,u)) &= F^{-1}_{\mathbf{p}}(F_{\mathbf{p}}(x,u))\\
&= F^{-1}_{\mathbf{q}}(F_{\mathbf{q}}(x,u))\\
&= (x,u).
\end{aligned}$$

It follows that $F_{\mathbf{p}}(x,u) = F_{\mathbf{q}}(x,u)$ for all $(x,u) \in [0,1]^2$. Fixing $u = 0$, we obtain

$$\tilde{\mathbf{p}}(x) = F_{\mathbf{p}}(x,0) = F_{\mathbf{q}}(x,0) = \tilde{\mathbf{q}}(x) \qquad (25)$$

for $x \in [0,1]$.

Assume, towards a contradiction, that $\mathbf{p} \ne \mathbf{q}$. Let a be any rational such that $\mathbf{p}(a) \ne \mathbf{q}(a)$, and suppose without loss of generality that $\mathbf{q}(a) < \mathbf{p}(a)$. By the Cauchy criterion (Remark A.4), there exists some $b > a$ such that

$$\sum_{a<x<b} \mathbf{q}(x) < \mathbf{p}(a) - \mathbf{q}(a).$$

Then we have

$$\begin{aligned}
\tilde{\mathbf{q}}(b) &= \tilde{\mathbf{q}}(a) + \mathbf{q}(a) + \sum_{x\in\mathbb{Q}\cap(a,b)} \mathbf{q}(x)\\
&= \tilde{\mathbf{p}}(a) + \mathbf{q}(a) + \sum_{x\in\mathbb{Q}\cap(a,b)} \mathbf{q}(x)\\
&< \tilde{\mathbf{p}}(a) + \mathbf{q}(a) + (\mathbf{p}(a) - \mathbf{q}(a))\\
&= \tilde{\mathbf{p}}(a) + \mathbf{p}(a)\\
&\le \tilde{\mathbf{p}}(b),
\end{aligned}$$

and so $\tilde{\mathbf{p}} \ne \tilde{\mathbf{q}}$, contradicting Eq. (25). □

The following corollary is an immediate consequence.

Corollary A.6 (Corollary 3.3 in the main text)**.** *If $\mathbf{p} \ne \mathbf{q}$, then there is some m such that R is not uniformly distributed on $[m+1]$.*

The next theorem strengthens Corollary A.6 by showing that R is non-uniform for all but finitely many m.

Theorem A.7 (Theorem 3.4 in the main text)**.** *If $\mathbf{p} \ne \mathbf{q}$, then there is some $M \ge 1$ such that for all $m \ge M$, the rank R is not uniformly distributed on $[m+1]$.*

Before proving Theorem A.7, we show the following lemma.

Lemma A.8. *Suppose $Z_1,\dots,Z_{m+1}$ is a finitely exchangeable sequence of Bernoulli random variables. If*

$$S_m := \sum_{i=1}^{m} Z_i$$

is not uniformly distributed on $[m+1]$, then

$$S_{m+1} := \sum_{i=1}^{m+1} Z_i$$

is not uniformly distributed on $[m+2]$.

Proof. By finite exchangeability, there is some $r \in [0,1]$ such that the distribution of every Z_i is $\mathsf{Bernoulli}(r)$. There are two cases.

Case 1: $r \ne 1/2$. For any $\ell \ge 1$, we have

$$\mathbb{E}[S_\ell] = \mathbb{E}\left[\sum_{i=1}^{\ell} Z_i\right] = \sum_{i=1}^{\ell} \mathbb{E}[Z_i] = \ell r \ne r/2 = \mathbb{E}[U_\ell],$$

and so S_ℓ is not uniformly distributed on $[\ell+1]$. In particular, this holds for ℓ equal to either m or $m+1$, and so both the hypothesis and conclusion are true.

Case 2: $r = 1/2$. We prove the contrapositive. Suppose that S_{m+1} is uniformly distributed on $[m+1]$.

Assume S_{m+1} is uniform and fix $k \in [m+1]$. By total probability, we have

$$\begin{aligned}
\Pr\{S_m = k\} = \Pr\{&S_m = k \text{ and } Z_{m+1} = 0\}\\
&+ \Pr\{S_m = k \text{ and } Z_{m+1} = 1\}.
\end{aligned} \qquad (26)$$

We consider the two events on the right-hand side of Eq. (26) separately.

First, the event $\{S_m = k\} \cap \{Z_{m+1} = 0\}$ is the union over all $\binom{m}{k}$ assignments of $(Z_1,\dots,Z_m)$ that have exactly k ones and $Z_{m+1} = 0$. All such assignments are disjoint events. Define the event

$$\begin{aligned}
A := \{&Z_1 = \cdots = Z_k = 1\\
&\text{and } Z_{k+1} = \cdots = Z_m = Z_{m+1} = 0\}.
\end{aligned}$$

By finite exchangeability, each assignment has probability $\Pr\{A\}$, and so

$$\Pr\{S_m = k \text{ and } Z_{m+1} = 0\} = \binom{m}{k} \Pr\{A\}. \quad (27)$$

Now, observe that the event $\{S_{m+1} = k\}$ is the union of all $\binom{m+1}{k}$ assignments of $(Z_1, \dots, Z_{m+1})$ that have exactly k ones. All the assignments are disjoint events and each has probability $\Pr\{A\}$, and so

$$\begin{aligned}\Pr\{S_{m+1} = k\} &= \binom{m+1}{k} \Pr\{A\} \\ &= \frac{1}{m+2}.\end{aligned} \quad (28)$$

Second, the event $\{S_m = k\} \cap \{Z_{m+1} = 1\}$ is the union over all $\binom{m}{k}$ assignments of $(Z_1, \dots, Z_m)$ that have exactly k ones and also $Z_{m+1} = 1$. All such assignments are disjoint events. Define the event

$$\begin{aligned}B \coloneqq \{&Z_1 = \dots = Z_k = Z_{m+1} = 1 \\ &\text{and } Z_{k+1} = \dots = Z_m = 0\}.\end{aligned}$$

Again by finite exchangeability, each assignment has probability $\Pr\{B\}$, and so

$$\Pr\{S_m = k \text{ and } Z_{m+1} = 1\} = \binom{m}{k} \Pr\{B\}. \quad (29)$$

Likewise, observe that the event $\{S_{m+1} = k+1\}$ is the union of all $\binom{m+1}{k+1}$ assignments of $(Z_1, \dots, Z_{m+1})$ that have exactly $k+1$ ones. All the assignments are disjoint events and each has probability $\Pr\{B\}$, and so

$$\begin{aligned}\Pr\{S_{m+1} = k+1\} &= \binom{m+1}{k+1} \Pr\{B\} \\ &= \frac{1}{m+2}.\end{aligned} \quad (30)$$

We now take Eq. (26), divide by $1/(m+2)$, and replace terms using Eqs. (27), (28), (29), and (30):

$$\begin{aligned}&\frac{\Pr\{S_m = k\}}{1/(m+2)} \\ &\quad = \frac{\Pr\{S_m = k \text{ and } Z_{m+1} = 0\}}{1/(m+2)} \\ &\qquad + \frac{\Pr\{S_m = k \text{ and } Z_{m+1} = 1\}}{1/(m+2)} \\ &\quad = \frac{\binom{m}{k}\Pr\{A\}}{\binom{m+1}{k}\Pr\{A\}} + \frac{\binom{m}{k}\Pr\{B\}}{\binom{m+1}{k+1}\Pr\{B\}} \\ &\quad = \frac{m!}{k!(m-k)!}\frac{k!(m+1-k)!}{(m+1)!} \\ &\qquad + \frac{m!}{k!(m-k)!}\frac{(k+1)!(m+1-(k+1))!}{(m+1)!} \\ &\quad = \frac{m+1-k}{m+1} + \frac{k+1}{m+1} \\ &\quad = \frac{m+2}{m+1} \\ &\quad = \frac{1/(m+1)}{1/(m+2)},\end{aligned}$$

and so we conclude that $\Pr\{S_m = k\} = 1/(m+1)$. □

We are now ready to prove Theorem A.7.

Proof of Theorem A.7. Suppose $\mathbf{p} \neq \mathbf{q}$. By Corollary A.6, there is some $M \geq 1$ such that the rank statistic $R = \sum_{i=1}^{M} \mathbb{I}[T_i \prec T_0]$ for $m = M$ is non-uniform over $[M+1]$. Observe that the rank statistic for $m = M+1$ is given by $\sum_{i=1}^{M+1} \mathbb{I}[T_i \prec T_0]$.

Now, each indicator $Z_i \coloneqq \mathbb{I}[T_i \prec T_0]$ is a Bernoulli random variable, and they are identically distributed since $(T_1, \dots, T_{M+1})$ is an i.i.d. sequence. Furthermore the sequence $(Z_1, \dots, Z_{M+1})$ is finitely exchangeable since the Z_i are conditionally independent given T_0. Then the sequence of indicators $(\mathbb{I}[T_1 \prec T_0], \mathbb{I}[T_2 \prec T_0], \dots, \mathbb{I}[T_{M+1} \prec T_0])$ satisfy the hypothesis of Lemma A.8, and so the rank statistic for $M+1$ is non-uniform. By induction, the rank statistic is non-uniform for all $m \geq M$. □

In fact, unless $\mathbf{p}$ and $\mathbf{q}$ satisfy an adversarial symmetry relationship under the selected ordering $\prec$, the rank is non-uniform for *any* choice of $m \geq 1$. Let $\lhd$ denote the lexicographic order on $\mathcal{T} \times [0,1]$ induced by $(\mathcal{T}, \prec)$ and $([0,1], <)$.

Corollary A.9 (Corollary 3.5 in the main text)**.** *Suppose* $\Pr\{(X, U_1) \lhd (Y, U_0)\} \neq 1/2$ *for* $Y \sim \mathbf{q}$, $X \sim \mathbf{p}$, *and* $U_0, U_1 \sim^{\text{iid}} \mathsf{Uniform}(0,1)$*. Then for all* $m \geq 1$*, the rank* R *is not uniformly distributed on* $[m+1]$*.*

Proof. If $\Pr\{(X, U_1) \lhd (Y, U_0)\} \neq 1/2$ then R is non-uniform for $m = 1$. The conclusion follows by Theorem A.7. □

A.2 An ordering that witnesses $\mathbf{p} \neq \mathbf{q}$ for $m = 1$

We now describe an ordering $\prec$ for which, when $m = 1$, we have $\Pr\{R = 0\} > 1/2$.

Define

$$A \coloneqq \{x \in \mathcal{T} \mid \mathbf{q}(x) > \mathbf{p}(x)\}$$

to be the set of all elements of $\mathcal{T}$ that have a greater probability according to $\mathbf{q}$ than according to $\mathbf{p}$, and let A^c denote its complement. Let $\mathbf{h}_{\mathbf{p},\mathbf{q}}$ be the signed measure given by the difference $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x) \coloneqq \mathbf{q}(x) - \mathbf{p}(x)$

between $\mathbf{q}$ and $\mathbf{p}$; for the rest of this subsection, we denote this simply by $\mathbf{h}$. Let $\prec$ be any total order on $\mathcal{T}$ satisfying

- if $\mathbf{h}(x) > \mathbf{h}(x')$ then $x \prec x'$; and
- if $\mathbf{h}(x) < \mathbf{h}(x')$ then $x \succ x'$.

The linear ordering $\prec$ may be defined arbitrarily for all pairs x and x' which satisfy $\mathbf{h}(x) = \mathbf{h}(x')$. As an immediate consequence, $x \prec x'$ whenever $x \in A$ and $x' \in A^c$. Intuitively, the ordering is designed to ensure that elements $x \in A$ are "small", and are ordered by decreasing value of $\mathbf{q}(x) - \mathbf{p}(x)$ (with ties broken arbitrarily); elements $x \in A^c$ are "large" and are ordered by increasing value of $\mathbf{p}(x) - \mathbf{q}(x)$ (again, with ties broken arbitrarily). The smallest element in $\mathcal{T}$ maximizes $\mathbf{q}(x) - \mathbf{p}(x)$ and the largest element in $\mathcal{T}$ maximizes $\mathbf{p}(x) - \mathbf{q}(x)$.

We first establish some easy lemmas.

Lemma A.10. *$A = \emptyset$ if and only if $\mathbf{p} = \mathbf{q}$.*

Proof. Immediate. □

Lemma A.11.

$$\sum_{x \in A} [\mathbf{q}(x) - \mathbf{p}(x)] = \sum_{x \in A^c} [\mathbf{p}(x) - \mathbf{q}(x)].$$

Proof. We have

$$\begin{aligned}&\sum_{x \in A}[\mathbf{q}(x) - \mathbf{p}(x)] - \sum_{x \in A^c} [\mathbf{p}(x) - \mathbf{q}(x)]\\ &\quad = \sum_{x \in \mathcal{T}} \mathbf{q}(x) - \sum_{x \in \mathcal{T}} \mathbf{p}(x) = 0,\end{aligned}$$

as desired. □

Given a probability distribution $\mathbf{r}$, define its cumulative distribution function $\tilde{\mathbf{r}}$ by $\tilde{\mathbf{r}}(x) \coloneqq \sum_{y \prec x} \mathbf{r}(y)$.

Lemma A.12. *$\tilde{\mathbf{q}}(x) > \tilde{\mathbf{p}}(x)$ for all $x \in \mathcal{T}$.*

Proof. Let $\mathcal{T}_x \coloneqq \{y \in \mathcal{T} \mid y \prec x\}$. If $x \in A$ then $\mathcal{T}_x \subseteq A$, and so

$$\tilde{\mathbf{q}}(x) - \tilde{\mathbf{p}}(x) = \sum_{y \in \mathcal{T}_x} [\mathbf{q}(y) - \mathbf{p}(y)] > 0,$$

since all terms in the sum are positive.

Otherwise, $y \in A$ for all $y \prec x$, and so $A \subseteq \mathcal{T}_x$. Let $A^c_x \coloneqq \{y \in A^c \mid y \prec x\}$. Then

$$\begin{aligned}&\tilde{\mathbf{q}}(x) - \tilde{\mathbf{p}}(x)\\ &= \sum_{y \prec x}[\mathbf{q}(y) - \mathbf{p}(y)]\\ &= \sum_{y \in A}[\mathbf{q}(y) - \mathbf{p}(y)] + \sum_{y \in A^c_x} [\mathbf{q}(y) - \mathbf{p}(y)]\\ &= \sum_{y \in A_x} [\mathbf{q}(y) - \mathbf{p}(y)] - \sum_{y \in A^c_x} [\mathbf{p}(y) - \mathbf{q}(y)]\\ &> \sum_{y \in A_x} [\mathbf{q}(y) - \mathbf{p}(y)] - \sum_{y \in A^c} [\mathbf{p}(y) - \mathbf{q}(y)]\\ &= 0,\end{aligned}$$

establishing the lemma. □

We now analyze $\Pr\{R = 0\}$ in the case where $m = 1$. In this case, we may drop some subscripts and write Y in place of X_1, so that our setting reduces to the following random variables:

$$\begin{aligned}X_{\mathbf{p}} &\sim \mathbf{p}\\ Y_{\mathbf{q}} &\sim \mathbf{q}\\ R_{\mathbf{p},\mathbf{q}} \mid X_{\mathbf{p}}, Y_{\mathbf{q}} &\sim \begin{cases} 0 & \text{if } X_{\mathbf{p}} \succ Y_{\mathbf{q}},\\ 1 & \text{if } X_{\mathbf{p}} \prec Y_{\mathbf{q}},\\ \text{Bernoulli}(1/2) & \text{if } X_{\mathbf{p}} = Y_{\mathbf{q}}.\end{cases}\end{aligned}$$

(We have indicated $\mathbf{p}$ and $\mathbf{q}$ in the subscripts, for use in the next subsection.)

In other words, the procedure samples $X_{\mathbf{p}} \sim \mathbf{p}$ and $Y_{\mathbf{q}} \sim \mathbf{q}$ independently. Given these values, it then sets $R_{\mathbf{p},\mathbf{q}}$ to be 0 if $X_{\mathbf{p}} \succ Y_{\mathbf{q}}$, to be 1 if $X_{\mathbf{p}} \prec Y_{\mathbf{q}}$, and the outcome of an independent fair coin flip otherwise.

For the rest of this subsection, we will refer to these random variables simply as X, Y, and R, though later on we will need them for several choices of distributions $\mathbf{p}$ and $\mathbf{q}$ (and accordingly will retain the subscripts).

We now prove the following theorem.

Theorem A.13 (Theorem 3.6 in the main text)**.** *If $\mathbf{p} \neq \mathbf{q}$, then for $m = 1$ and the ordering $\prec$ defined above, we have* $\Pr\{R = 0\} > 1/2$.

Proof. From total probability and independence of X and Y, we have

$$\begin{aligned}&\Pr\{R = 0\}\\ &= \sum_{x,y \in \mathcal{T}} \Pr\{R{=}0 \mid X{=}x, Y{=}y\} \Pr\{Y = y\} \Pr\{X = x\}\\ &= \sum_{x,y \in \mathcal{T}} \Pr\{R{=}0 \mid X{=}x, Y{=}y\}\, \mathbf{q}(y)\mathbf{p}(x)\\ &= \sum_{x \in \mathcal{T}} \Pr\{R{=}0 \mid X{=}x, Y{=}x\}\, \mathbf{q}(x)\mathbf{p}(x)\\ &\qquad + \sum_{y \prec x \in \mathcal{T}} \Pr\{R{=}0 \mid X{=}x, Y{=}y\}\, \mathbf{q}(y)\mathbf{p}(x)\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{x \prec y \in \mathcal{T}} \Pr\{R{=}0 \,|\, X{=}x, Y{=}y\}\, \mathbf{q}(y)\mathbf{p}(x) \\
&= \frac{1}{2} \sum_{x \in \mathcal{T}} \mathbf{q}(x)\mathbf{p}(x) + 1 \sum_{y \prec x \in \mathcal{T}} \mathbf{q}(y)\mathbf{p}(x) \\
&\quad + 0 \sum_{x \prec y \in \mathcal{T}} \mathbf{q}(y)\mathbf{p}(x) \\
&= \frac{1}{2} \sum_{x \in \mathcal{T}} \mathbf{p}(x)\mathbf{q}(x) + \sum_{x \in \mathcal{T}} \tilde{\mathbf{q}}(x)\mathbf{p}(x).
\end{aligned}
$$

An identical argument establishes that

$$
\Pr\{R{=}1\} = \frac{1}{2} \sum_{x \in \mathcal{T}} \mathbf{p}(x)\mathbf{q}(x) + \sum_{x \in \mathcal{T}} \tilde{\mathbf{p}}(x)\mathbf{q}(x).
$$

Since $\Pr\{R{=}0\} + \Pr\{R = 1\} = 1$, it suffices to establish that $\Pr\{R = 0\} > \Pr\{R = 1\}$. We have

$$
\begin{aligned}
&\Pr\{R = 0\} - \Pr\{R = 1\} \\
&= \sum_{x \in \mathcal{T}} \tilde{\mathbf{q}}(x)\mathbf{p}(x) - \sum_{x \in \mathcal{T}} \tilde{\mathbf{p}}(x)\mathbf{q}(x) \\
&> \sum_{x \in \mathcal{T}} \tilde{\mathbf{p}}(x)\mathbf{p}(x) - \sum_{x \in \mathcal{T}} \tilde{\mathbf{p}}(x)\mathbf{q}(x) \\
&= \sum_{x \in \mathcal{T}} \tilde{\mathbf{p}}(x)[\mathbf{p}(x) - \mathbf{q}(x)] \\
&= \sum_{x \in A^c} \tilde{\mathbf{p}}(x)[\mathbf{p}(x) - \mathbf{q}(x)] - \sum_{x \in A} \tilde{\mathbf{p}}(x)[\mathbf{q}(x) - \mathbf{p}(x)] \\
&\geq \sum_{x \in A^c} \big(\max_{y \in A} \tilde{\mathbf{p}}(y)\big)[\mathbf{p}(x) - \mathbf{q}(x)] \\
&\quad - \sum_{x \in A} \tilde{\mathbf{p}}(x)[\mathbf{q}(x) - \mathbf{p}(x)] \\
&= \sum_{x \in A} \big(\max_{y \in A} \tilde{\mathbf{p}}(y)\big)[\mathbf{q}(x) - \mathbf{p}(x)] \\
&\quad - \sum_{x \in A} \tilde{\mathbf{p}}(x)[\mathbf{q}(x) - \mathbf{p}(x)] \\
&= \sum_{x \in A} \big(\max_{y \in A} \tilde{\mathbf{p}}(y) - \tilde{\mathbf{p}}(x)\big)[\mathbf{q}(x) - \mathbf{p}(x)] \\
&> 0.
\end{aligned}
$$

The first inequality follows from Lemma A.12; the second inequality follows from monotonicity of $\tilde{\mathbf{p}}$; the second-to-last equality follows from Lemma A.11; and the final inequality follows from the fact that all terms in the sum are positive. □

A.3 A tighter bound in terms of $L_\infty(\mathbf{p}, \mathbf{q})$

We have just exhibited an ordering such that when $\mathbf{p} \neq \mathbf{q}$ and $m = 1$, we have $\Pr\{R = 0\} > 1/2$. We are now interested in obtaining a tighter lower bound on this probability in terms of the L_∞ distance between $\mathbf{p}$ and $\mathbf{q}$.

In this subsection and the following one, we assume that $\mathcal{T}$ is finite. We first note the following immediate lemma.

Lemma A.14. *Let $B, C \subseteq \mathcal{T}$. For all $\mathbf{p}, \mathbf{q}$ and all $\delta > 0$ there is an $\epsilon > 0$ such that for all distributions $\mathbf{p}'$ on $\mathcal{T}$ with $\sup_{x \in \mathcal{T}} |\mathbf{p}(x) - \mathbf{p}'(x)| < \epsilon$, we have*

$$
\begin{aligned}
&\big|\Pr(R_{\mathbf{p},\mathbf{q}} = 0 \,|\, X_{\mathbf{p}} \in B,\ Y_{\mathbf{q}} \in C) \\
&\quad - \Pr(R_{\mathbf{p}',\mathbf{q}} = 0 \,|\, X_{\mathbf{p}'} \in B,\ Y_{\mathbf{q}} \in C)\big| < \delta.
\end{aligned}
$$

Definition A.15. We say that $\mathbf{p}$ is **ϵ-discrete** (with respect to $\mathbf{q}$) if for all $a, b \in \mathcal{T}$ we have

$$
\big|\mathbf{h}_{\mathbf{p},\mathbf{q}}(a) - \mathbf{h}_{\mathbf{p},\mathbf{q}}(b)\big| \geq \epsilon.
$$

From Lemma A.14 we immediately obtain the following.

Lemma A.16. *For all $\mathbf{p}, \mathbf{q}$ and all $\delta > 0$ there is an $\epsilon > 0$ and an ϵ-discrete distribution $\mathbf{p}_\epsilon$ on $\mathcal{T}$ such that for all $B, C \subseteq \mathcal{T}$,*

$$
\begin{aligned}
&\big|\Pr(R_{\mathbf{p},\mathbf{q}} = 0 \,|\, X_{\mathbf{p}} \in B,\ Y_{\mathbf{q}} \in C) \\
&\quad - \Pr(R_{\mathbf{p}_\epsilon,\mathbf{q}} = 0 \,|\, X_{\mathbf{p}_\epsilon} \in B,\ Y_{\mathbf{q}} \in C)\big| < \delta.
\end{aligned}
$$

The next lemma will be crucial for proving our bound.

Lemma A.17. *Let $\mathbf{p}_0$ and $\mathbf{p}_1$ be probability measures on $\mathcal{T}$, and let $\lhd$ be a total order on $\mathcal{T}$ such that if $\mathbf{h}_{\mathbf{p}_0,\mathbf{q}}(x) > \mathbf{h}_{\mathbf{p}_0,\mathbf{q}}(x')$ then $x \lhd x'$ and if $\mathbf{h}_{\mathbf{p}_0,\mathbf{q}}(x) < \mathbf{h}_{\mathbf{p}_0,\mathbf{q}}(x')$ then $x \rhd x'$. Suppose that if $\mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) > 0$ and $\mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(y) \leq 0$, then $x \lhd y$. Then $\Pr(R_{\mathbf{p}_0,\mathbf{q}} = 0) \geq \Pr(R_{\mathbf{p}_1,\mathbf{q}} = 0)$.*

Proof. Note that

$$
\begin{aligned}
&\Pr(R_{\mathbf{p}_1,\mathbf{q}} = 0 \,|\, Y_{\mathbf{q}} = y) \\
&= \sum_{x \rhd y} \mathbf{p}_1(x) + \frac{1}{2}\mathbf{p}_1(y) \\
&= \sum_{x \rhd y} \mathbf{p}_0(x) + \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) + \frac{1}{2}[\mathbf{p}_0(y) + \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(y)] \\
&= \Pr(R_{\mathbf{p}_0,\mathbf{q}} = 0 \,|\, Y_{\mathbf{q}} = y) + \sum_{x \rhd y} \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) + \frac{1}{2}\mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(y) \\
&= \Pr(R_{\mathbf{p}_0,\mathbf{q}} = 0 \,|\, Y_{\mathbf{q}} = y) - \sum_{x \lhd y} \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) - \frac{1}{2}\mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(y),
\end{aligned}
$$

where the last equality holds because $\sum_{x \in \mathcal{T}} \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) = 0$. But by our assumption, we know that $\sum_{x \lhd y} \mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(x) + \frac{1}{2}\mathbf{h}_{\mathbf{p}_0,\mathbf{p}_1}(y)$ is non-negative and so $\Pr(R_{\mathbf{p}_1,\mathbf{q}} = 0 \,|\, Y_{\mathbf{q}} = y) \leq \Pr(R_{\mathbf{p}_0,\mathbf{q}} = 0 \,|\, Y_{\mathbf{q}} = y)$, from which the result follows. □

We will now provide a lower bound on $\Pr(R_{\mathbf{p},\mathbf{q}} = 0)$.

Proposition A.18.

$$\Pr(R_{\mathbf{p},\mathbf{q}} = 0) \geq \frac{1}{2} + \frac{1}{2} \max_{x \in \mathcal{T}} \mathbf{h}_{\mathbf{p},\mathbf{q}}(x)^2. \qquad (31)$$

Proof. Recall that $A \coloneqq \{x \in \mathcal{T} \mid \mathbf{q}(x) > \mathbf{p}(x)\}$. First note that by Lemma A.14, we may assume without loss of generality that $|A| = |\mathcal{T} \setminus A|$, by adding elements of mass arbitrarily close to 0. Let $k \coloneqq |A|$. Further, by Lemma A.16 we may assume without loss of generality that $\mathbf{p}, \mathbf{q}$ are an ϵ-discrete pair (for some fixed but small ϵ) with $|\mathcal{T}| \cdot \epsilon < L_\infty(\mathbf{p}, \mathbf{q})$. Let $(x_0^+, \dots, x_{k-1}^+)$ be the collection A listed in $\prec$-increasing order. Let $(x_0^-, \dots, x_{k-1}^-)$ be the collection $\mathcal{T} \setminus A$ listed in $\prec$-increasing order.

Let $\mathbf{p}^*$ be any probability measure such that

$$\mathbf{p}^*(x) = \begin{cases} \mathbf{p}(x) - e(\ell) & (x = x_\ell^-; e(\ell) \geq 0), \\ \mathbf{q}(x) - (k - \ell) \cdot \epsilon & (x = x_\ell^+; 0 \leq \ell < k - 1), \\ \mathbf{p}(x) & (x = x_0^+). \end{cases}$$

Note that for all $x, y \in \mathcal{T}$, we have $y \prec x$ if and only if $\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x) < \mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(y)$.

Now, for every $\ell < k - 1$ we have $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_\ell^+) \geq \ell \cdot \epsilon$ (as $\mathbf{p}, \mathbf{q}$ are an ϵ-discrete pair), and so we can always find such a $\mathbf{p}^*$. In particular the following are immediate.

(a) $x \prec y$ if and only if $\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x) > \mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(y)$,

(b) $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) = \mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_0^+)$,

(c) if $\mathbf{h}_{\mathbf{p},\mathbf{q}^*}(x) > 0$ and $\mathbf{h}_{\mathbf{p},\mathbf{p}^*}(y) \leq 0$ then $x \prec y$, and

(d) $(\mathbf{p}, \mathbf{q}^*)$ is an ϵ-discrete pair.

Note that $\Pr(R_{\mathbf{p},\mathbf{q}} = 0) \geq \Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0)$, by Lemma A.17 and (c). For simplicity, let $A_0 \coloneqq \{x_0^+\}$, $A_1 \coloneqq \{x_i^+\}_{1 \leq i \leq k-1}$ and $D \coloneqq \mathcal{T} \setminus A$.

We now condition on the value of $Y_\mathbf{q}$, in order to calculate $\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0)$.

Case 1: $Y_\mathbf{q} = x_i^-$. We have

$$\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0 \mid Y_\mathbf{q} = x_i^-) = \sum_{i<\ell<k} \mathbf{p}^*(x_\ell^-) + \frac{1}{2}\mathbf{p}^*(x_i^-).$$

Case 2: $Y_\mathbf{q} \in A_1$. We have

$$\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0 \mid Y_\mathbf{q} \in A_1) = \mathbf{p}^*(D) + \frac{1}{2}\mathbf{p}^*(A_1) + f_0(\epsilon),$$

where f_0 is a function satisfying $\lim_{\epsilon \to 0} f_0(\epsilon) = 0$.

Case 3: $Y_\mathbf{q} \in A_0$. We have

$$\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0 \mid Y_\mathbf{q} \in A_0) = \mathbf{p}^*(A_1) + \mathbf{p}^*(D) + \frac{1}{2}\mathbf{p}^*(A_0).$$

We may calculate these terms as follows:

$$\begin{aligned} \mathbf{p}^*(D) &= \mathbf{q}(D) + \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) + (k(k-1)/2)\epsilon, \\ \mathbf{p}^*(A_1) &= \mathbf{q}(A_1) - (k(k-1)/2)\epsilon, \\ \mathbf{p}^*(A_0) &= \mathbf{q}(A_0) - \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+). \end{aligned}$$

Putting all of this together, we obtain

$$\begin{aligned} &\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0) \\ &= \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{p}^*(x_\ell^-) + \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)\mathbf{p}^*(x_i^-) \\ &\quad + \mathbf{q}(A_1)\mathbf{p}^*(D) + \frac{1}{2}\mathbf{q}(A_1)\mathbf{p}^*(A_1) + \mathbf{q}(A_1)f_0(\epsilon) \\ &\quad + \mathbf{q}(A_0)\mathbf{p}^*(A_1) + \mathbf{q}(A_0)\mathbf{p}^*(D) + \frac{1}{2}\mathbf{q}(A_0)\mathbf{p}^*(A_0) \\ &= \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)[\mathbf{q}(x_\ell^-) - \mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_\ell^-)] \\ &\quad + \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)[\mathbf{q}(x_i^-) - \mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_i^-)] \\ &\quad + \mathbf{q}(A_1)[\mathbf{q}(D) + \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+)] + \frac{1}{2}\mathbf{q}(A_1)\mathbf{q}(A_1) \\ &\quad + \mathbf{q}(A_0)\mathbf{q}(A_1) + \mathbf{q}(A_0)[\mathbf{q}(D) + \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+)] \\ &\quad + \frac{1}{2}\mathbf{q}(A_0)[\mathbf{q}(A_0) - \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+)] + f_1(\epsilon) \\ &= \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{q}(x_\ell^-) + \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)\mathbf{q}(x_i^-) \\ &\quad + \mathbf{q}(A_1)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_1)\mathbf{q}(A_1) + \mathbf{q}(A_0)\mathbf{q}(A_1) \\ &\quad + \mathbf{q}(A_0)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_0)\mathbf{q}(A_0) \\ &\quad - \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_\ell^-) \\ &\quad - \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_i^-) + \mathbf{q}(A_1)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) \\ &\quad + \mathbf{q}(A_0)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) - \frac{1}{2}\mathbf{q}(A_0)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) + f_1(\epsilon) \\ &= \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{q}(x_\ell^-) + \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)\mathbf{q}(x_i^-) \\ &\quad + \mathbf{q}(A_1)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_1)\mathbf{q}(A_1) + \mathbf{q}(A_0)\mathbf{q}(A_1) \\ &\quad + \mathbf{q}(A_0)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_0)\mathbf{q}(A_0) \\ &\quad - \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_\ell^-) \\ &\quad - \frac{1}{2} \sum_{i<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_i^-) + \mathbf{q}(A_1)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) \\ &\quad + \frac{1}{2}\mathbf{q}(A_0)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) + f_1(\epsilon), \end{aligned}$$

where f_1 is a function satisfying $\lim_{\epsilon \to 0} f_1(\epsilon) = 0$.

We also have

$$\begin{aligned}\frac{1}{2} &= \Pr(R_{\mathbf{q},\mathbf{q}} = 0) \\ &= \sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{q}(x_\ell^-) + \frac{1}{2}\sum_{i<k} \mathbf{q}(x_i^-)\mathbf{q}(x_i^-) \\ &\quad + \mathbf{q}(A_1)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_1)\mathbf{q}(A_1) \\ &\quad + \mathbf{q}(A_0)\mathbf{q}(A_1) + \mathbf{q}(A_0)\mathbf{q}(D) + \frac{1}{2}\mathbf{q}(A_0)\mathbf{q}(A_0).\end{aligned}$$

Putting these two equations together, we obtain

$$\begin{aligned}&\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0) - \frac{1}{2} \\ &= \Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0) - \Pr(R_{\mathbf{q},\mathbf{q}} = 0) \\ &= -\sum_{i<k} \sum_{i<\ell<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_\ell^-) \\ &\quad - \frac{1}{2}\sum_{i<k} \mathbf{q}(x_i^-)\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_i^-) + \mathbf{q}(A_1)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) \\ &\quad + \frac{1}{2}\mathbf{q}(A_0)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) + f_1(\epsilon) \\ &\geq \frac{1}{2}\mathbf{q}(A_0)\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) + f_1(\epsilon),\end{aligned}$$

as $\mathbf{h}_{\mathbf{p}^*,\mathbf{q}}(x_\ell^-) \leq 0$ for all $\ell < k$ and $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) \leq 0$.

But we know that

$$\mathbf{q}(A_0) = \mathbf{q}(x_0^+) = \mathbf{p}^*(x_0^+) + \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+) \geq \mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+).$$

Therefore, as $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x_0^+)$ is the maximal value of $\mathbf{h}_{\mathbf{p},\mathbf{q}}$, by taking the limit as $\epsilon \to 0$ we obtain

$$\Pr(R_{\mathbf{p}^*,\mathbf{q}} = 0) \geq \frac{1}{2} + \frac{1}{2}\max_{x\in\mathcal{T}} \mathbf{h}_{\mathbf{p},\mathbf{q}}(x)^2,$$

as desired. □

Finally, we arrive at the following theorem.

Theorem A.19. *Given probability measure $\mathbf{p}, \mathbf{q}$ on $\mathcal{T}$ there is a linear ordering $\sqsubset$ of $\mathcal{T}$ such that if $X_\mathbf{p}$ and $Y_\mathbf{q}$ are sampled independently from $\mathbf{p}$ and $\mathbf{q}$ respectively then*

$$\Pr(X_\mathbf{q} \sqsubset Y_\mathbf{p}) \geq \frac{1}{2} + \frac{1}{2}L_\infty(\mathbf{p},\mathbf{q})^2. \tag{32}$$

Proof. Note that

$$L_\infty(\mathbf{p},\mathbf{q}) = \max\{\max_{x\in\mathcal{T}} \mathbf{h}_{\mathbf{p},\mathbf{q}}(x), \max_{x\in\mathcal{T}} \mathbf{h}_{\mathbf{q},\mathbf{p}}(x)\}.$$

If $L_\infty(\mathbf{p},\mathbf{q}) = \max_{x\in\mathcal{T}} \mathbf{h}_{\mathbf{p},\mathbf{q}}(x)$, then the theorem follows from Proposition A.18 using the ordering $x \sqsubset y$ if and only if $\mathbf{h}_{\mathbf{p},\mathbf{q}}(x) > \mathbf{h}_{\mathbf{p},\mathbf{q}}(y)$.

If, however, $L_\infty(\mathbf{p},\mathbf{q}) = \max_{x\in\mathcal{T}} \mathbf{h}_{\mathbf{q},\mathbf{p}}(x)$, then the theorem follows from Proposition A.18 by interchanging $\mathbf{p}$ and $\mathbf{q}$, i.e., by using the ordering $x \sqsubset y$ if and only if $\mathbf{h}_{\mathbf{q},\mathbf{p}}(x) > \mathbf{h}_{\mathbf{q},\mathbf{p}}(y)$. □

A.4 Sample complexity

We now show how to amplify this result by repeated trials to obtain a bound on the sample complexity of the main algorithm for determining whether $\mathbf{p} = \mathbf{q}$.

Let $\sqsubset$ be the linear ordering defined in Theorem A.19.

Theorem A.20 (Theorem 3.7 in the main text)**.** *Given significance level $\alpha = 2\Phi(-c)$ for $c > 0$, the proposed test with ordering $\sqsubset$ and $m = 1$ achieves power $\beta \geq 1 - \Phi(-c)$ using*

$$n \approx 4c^2/L_\infty(\mathbf{p},\mathbf{q})^4 \tag{33}$$

samples from $\mathbf{q}$, where Φ is the cumulative distribution function of a standard normal.

Proof. Assume without loss of generality that the order $\sqsubset$ from Theorem A.19 is such that $L_\infty = \max_{s\in\mathcal{T}}(\mathbf{q}(x) - \mathbf{p}(x))$. Let $(Y_1, \ldots, Y_n) \sim^{\text{iid}} \mathbf{q}$ be the n samples from $\mathbf{q}$. With $m = 1$, the testing procedure generates n samples $(X_1, \ldots, X_n) \sim^{\text{iid}} \mathbf{p}$, and $2n$ uniform random variables $(U_1^Y, \ldots, U_n^Y, U_1^X, \ldots, U_n^X) \sim^{\text{iid}}$ $\mathsf{Uniform}(0,1)$ to break ties. Let $\lhd$ denote the lexicographic order on $\mathcal{T} \times [0,1]$ induced by $(\mathcal{T}, \lhd)$ and $([0,1], <)$. Define $W_i \coloneqq \mathbb{I}\left[(Y_i, U_i^Y) \lhd (X_i, U_i^X)\right]$, for $1 \leq i \leq n$, to be the rank of the i-th observation from $\mathbf{q}$.

Under the null hypothesis H_0, each rank W_i has distribution $\mathsf{Bernoulli}(1/2)$ by Lemma A.2. Testing for uniformity of the ranks on $\{0,1\}$ is equivalent to testing whether a coin is unbiased given the i.i.d. flips $\{W_1, \ldots, W_n\}$. Let $\hat{B} \coloneqq \sum_{i=1}^n (1 - W_i)/n$ denote the empirical proportion of zeros. By the central limit theorem, for sufficiently large n, we have that $\hat{B}$ is approximately normally distributed with mean $1/2$ and standard deviation $1/(2\sqrt{n})$. For the given significance level $\alpha = 2\Phi(-c)$, we form the two-sided reject region $F = (-\infty, \gamma) \cup (\gamma, \infty)$, where the critical value γ satisfies

$$c = \frac{\gamma - 1/2}{1/(2\sqrt{n})} = 2\sqrt{n}(\gamma - 1/2). \tag{34}$$

Replacing n in Eq. (7), we obtain

$$\begin{aligned}\gamma &= 1/2 + c/(2\sqrt{n}) \\ &= 1/2 + c/(2(2c/L_\infty(\mathbf{p},\mathbf{q})^2)) \\ &= 1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/4. \end{aligned} \tag{35}$$

This construction ensures that $\Pr\{\mathsf{reject} \mid \mathsf{H}_0\} = \alpha$.

We now show that the test with this rejection region has power $\beta \geq \Pr\{\mathsf{reject} \mid \mathsf{H}_1\} = 1 - \Phi(-c)$. Under the alternative hypothesis H_1, each W_i has (in the worst case) distribution $\mathsf{Bernoulli}(1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/2)$ by Theorem A.19, so that the empirical proportion

$\hat{B}$ is approximately normally distributed with mean at least $1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/2$ and standard deviation at most $1/(2\sqrt{n})$. Under the alternative distribution of $\hat{B}$, the standard score c' of the critical value γ is

$$\begin{aligned} c' &= \frac{\gamma - (1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/2)}{1/(2\sqrt{n})} \\ &= 2\sqrt{n}((1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/4) - (1/2 + L_\infty(\mathbf{p},\mathbf{q})^2/2)) \\ &= -2\sqrt{n}(L_\infty(\mathbf{p},\mathbf{q})^2/4) \\ &= -\sqrt{n}L_\infty(\mathbf{p},\mathbf{q})^2/2 \\ &= -c, \end{aligned} \tag{36}$$

where the second equality follows from Eq. (35). Observe that the not reject region $F^c = [-\gamma, \gamma] \subset (-\infty, \gamma]$, and so the probability that $\hat{B}$ falls in F^c is at most the probability that $\hat{B} < \gamma$, which by Eq. (36) is equal to $\Phi(-c)$. It is then immediate that $\beta \geq 1 - \Phi(-c)$. $\square$

The following corollary follows directly from Theorem 3.7.

Corollary A.21. *As the significance level α varies, the proposed test with ordering $\sqsubset$ and $m = 1$ achieves an overall error $(\alpha + (1-\beta))/2 \leq 3\Phi(-c)/2$ using $n = 4c^2/L_\infty(\mathbf{p},\mathbf{q})^4$ samples.*

A.5 Distribution of the test statistic under the alternative hypothesis

In this subsection we derive the distribution of R under the alternative hypothesis $\mathbf{p} \neq \mathbf{q}$. As before, write $\tilde{\mathbf{p}}(x) \coloneqq \sum_{x' < x} \mathbf{p}(x)$.

Theorem A.22. *The distribution of R is given by*

$$\Pr\{R = r\} = \sum_{x \in \mathcal{T}} H(x, m, r)\, \mathbf{q}(x) \tag{37}$$

for $0 \leq r \leq m$, where $H(x, m, r) \coloneqq$

$$\begin{cases} \dbinom{r}{m} [\tilde{\mathbf{p}}(x)]^r [1 - \tilde{\mathbf{p}}(x)]^{m-r} & (\mathbf{p}(x) = 0) \\ \dfrac{1}{m+1} & (\mathbf{p}(x) = 1) \\ \displaystyle\sum_{e=0}^{m} \Bigg\{ \Bigg[\sum_{j=0}^{e} \binom{m-e}{r-j} \left[\frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{r-j} \\ \quad \left[1 - \frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{(m-e)-(r-j)} \left(\frac{1}{e+1}\right) \Bigg] \\ \quad \dbinom{m}{e} [\mathbf{p}(x)]^m [1 - \mathbf{p}(x)]^{e-m} \Bigg\} & (0 < \mathbf{p}(x) < 1) \end{cases}$$

Proof. Define the following random variables:

$$L \coloneqq \sum_{i=1}^{m} \mathbb{I}[X_i \prec X_0], \tag{38}$$

$$E \coloneqq \sum_{i=1}^{m} \mathbb{I}[X_i = X_0], \tag{39}$$

$$G \coloneqq \sum_{i=1}^{m} \mathbb{I}[X_i \succ X_0]. \tag{40}$$

We refer to L, E, and G as "bins", where L is the "less than" bin, E is the "equal to" bin, and G is the "greater than" bin (all with respect to X_0). Total probability gives

$$\begin{aligned} \Pr\{R = r\} &= \sum_{x \in \mathcal{T}} \Pr\{R = r, X_0 = x\} \\ &= \sum_{\substack{x \in \mathcal{T} \\ \mathbf{q}(x) > 0}} \Pr\{R = r \mid X_0 = x\}\, \mathbf{q}(x). \end{aligned}$$

Fix $x \in \mathcal{T}$ such that $\mathbf{q}(x) > 0$. Consider $\Pr\{R = r \mid X_0 = s\}$. The counts in bins L, E, and G are binomial random variables with m trials, where the bin L has success probability $\tilde{\mathbf{p}}(x)$, the bin E has success probability $\mathbf{p}(x)$, and the bin G has success probability $1 - (\tilde{\mathbf{p}}(x) + \mathbf{p}(x))$. We now consider three cases.

Case 1: $\mathbf{p}(x) = 0$. The event $\{E = 0\}$ occurs with probability one since each X_i, for $1 \leq i \leq m$, cannot possibly be equal to x. Therefore, conditioned on $\{X_0 = x\}$, the event $\{R = r\}$ occurs if and only if $\{L = r\}$. Since L is binomially distributed,

$$\begin{aligned} \Pr\{R = r \mid X_0 = x\} &= \Pr\{L = r \mid X_0 = x\} \\ &= \binom{m}{r} [\tilde{\mathbf{p}}(x)]^r [1 - \tilde{\mathbf{p}}(x)]^{m-r}. \end{aligned}$$

Case 2: $\mathbf{p}(x) = 1$. Then the event $\{E = m\}$ occurs with probability one since each X_i, for $1 \leq i \leq m$, can only equal s. The uniform numbers $U_0, \dots, U_m$ used to break the ties will determine the rank R of X_0. Let B be the rank of U_0 among the m other uniform random variables $U_1, \dots, U_m$. The event $\{R = r\}$ occurs if and only if $\{B = r\}$. Since the U_i are i.i.d., B is uniformly distributed over $\{0, 1, 2, \dots, m\}$ by Lemma A.2. Hence

$$\Pr\{R = r \mid X_0 = x\} = \Pr\{B = r \mid X_0 = x\} = \frac{1}{m+1}.$$

Case 3: $0 < \mathbf{p}(x) < 1$. By total probability,

$$\begin{aligned} &\Pr\{R = r \mid X_0 = x\} \\ &= \sum_{e=0}^{m} \Pr\{R = r \mid X_0 = x, E = e\} \Pr\{E = e \mid X_0 = x\}. \end{aligned}$$

Since E is binomially distributed,

$$\Pr\{E = e \mid X_0 = x\} = \binom{m}{e} [\mathbf{p}(x)]^e [1 - \mathbf{p}(x)]^{m-e}.$$

We now tackle the event $\{R=r \mid X_0=x, E=e\}$. The uniform numbers $U_0,\dots,U_m$ used to break the ties will determine the rank R of X_0. Define B to be the rank of U_0 among the e other uniform random variables assigned to bin E, i.e., those U_i for $1 \le i \le m$ such that $X_i = s$. The random variable B is independent of all the X_i, but is dependent on E. Given $\{E=e\}$, B is uniformly distributed on $\{0,1,\dots,e\}$. By total probability,

$$
\begin{aligned}
&\Pr\{R=r \mid X_0=x, E=e\}\\
&=\sum_{b=0}^{e}\big[\Pr\{R=r \mid X_0=x, E=e, B=b\}\\
&\qquad\qquad \Pr\{B=b \mid E=e\}\big]\\
&=\sum_{b=0}^{e}\Pr\{R=r \mid X_0=x, E=e, B=b\}\,\frac{1}{e+1}.
\end{aligned}
$$

Conditioned on $\{E=e\}$ and $\{B=0\}$, the event $\{R=r\}$ occurs if and only if $\{L=r\}$, since exactly 0 random variables in bin E “are less” than X_0, so exactly r random variables in bin L are needed to ensure that the rank of X_0 is r. By the same reasoning, for $0 \le b \le e$, conditioned on $\{E=e, B=b\}$ we have $\{R=r\}$ if and only if $\{L=r-b\}$.

Now, conditioned on $\{E=e\}$, there are $m-e$ remaining assignments to be split among bins L and G. Let i be such that $X_i \ne x$. Then the relative probability that X_i is assigned to bin L is $\tilde{\mathbf{p}}(x)$ and to bin G is $1-(\tilde{\mathbf{p}}(x)+\mathbf{p}(x))$. Renormalizing these probabilities, we conclude that L is conditionally (given $\{E=e\}$) a binomial random variable with $m-e$ trials and success probability $\tilde{\mathbf{p}}(x)/(\tilde{\mathbf{p}}(x)+(1-(\tilde{\mathbf{p}}(x)+\mathbf{p}(x)))) = \tilde{\mathbf{p}}(x)/(1-\mathbf{p}(x))$. Hence

$$
\begin{aligned}
&\Pr\{R=r \mid X_0=x, E=e, B=b\}\\
&=\Pr\{L=r-b \mid X_0=x, E=e\}\\
&=\binom{m-e}{r-j}\left[\frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{r-j}\left[1-\frac{\tilde{\mathbf{p}}(x)}{1-\mathbf{p}(x)}\right]^{(m-e)-(r-j)},
\end{aligned}
$$

completing the proof. □

Remark A.23. The sum in Eq. (37) of Theorem A.22 converges since $H(x,m,r) \le 1$.

Remark A.24. Theorem A.22 shows that it is not the case that we must have $\mathbf{p}=\mathbf{q}$ whenever there exists some m for which the rank R is uniform on $[m+1]$. For example, let $m=1$, let $\mathcal{T} \coloneqq \{0,1,2,3\}$, let $\prec$ be the usual order $<$ on $\mathcal{T}$, and let $\mathbf{p} \coloneqq \frac{1}{2}\delta_0+\frac{1}{2}\delta_3$ and $\mathbf{q} \coloneqq \frac{1}{2}\delta_1+\frac{1}{2}\delta_2$. Let $X\sim\mathbf{p}$ and $Y\sim\mathbf{q}$. Then we have $\Pr\{R=0\} = \Pr\{X>Y\} = 1/2 = \Pr\{Y<X\} = \Pr\{R=1\}$. Rather, Theorem A.1 tells us merely if R is not uniform on $\{0,\dots,m\}$ for *some* m, then $\mathbf{p}\ne\mathbf{q}$. In the example given above, $m=2$ (and so by Theorem A.7 all $m\ge 2$) provides such a witness.