Strongly Replicable Distribution Testers imply High Probability Distribution Testers

Ilias Diakonikolas
University of Wisconsin-Madison
ilias@cs.wisc.edu

Jingyi Gao
University of Wisconsin-Madison
jingyig@cs.wisc.edu

Daniel M. Kane
University of California, San Diego
dakane@cs.ucsd.edu

Sihan Liu
University of California, San Diego
La Jolla, CA
sil046@ucsd.edu

Christopher Ye
University of California, San Diego
La Jolla, CA
czye@ucsd.edu

Abstract

We study connections between the sample cost of replicability, which requires an algorithm to produce consistent outputs in different runs, and the sample complexity in the high success probability regime for distribution testing algorithms. Specifically, we consider a stronger notion of replicability, which we term as strong replicability, and show that any ρ-strongly-replicable testing algorithm can be converted in a blackbox manner to a tester with at most $\exp(-\widetilde{\Omega}(\rho^{-2}))$ failure probability under the same sample budget. Our conversion is optimal for many natural testers. As applications of our conversion theorem, we obtain tight sample lower bounds for strongly replicable testers for coin-bias and closeness testing.

1 Introduction

In distribution testing, given sample access to a distribution $\mathbf{p}$, an algorithm is asked to test if $\mathbf{p}$ satisfies a property $\mathcal{P}$: it must outputs 1 with probability $2/3$ if $\mathbf{p} \in \mathcal{P}$, and outputs 0 with probability $2/3$ if $\mathbf{p}$ is far from $\mathcal{P}$. After the pioneering works formulating this field [7, 36] from a TCS perspective, a line of work has given efficient testers achieving information-theoretically optimal sample complexities for a variety of problems, such as uniformity/identity testing [49, 31, 53], and closeness testing [7, 54, 8, 22]. More broadly, substantial progress has been made on testing a wide range of natural properties; see, e.g., [6, 5, 9, 2, 44, 30, 3, 28, 17, 32, 24, 18, 19, 16, 29] for a sample of works, and [52, 15, 13] for surveys on the topic. The formal definition is as follows.

Definition 1.1 ((ε, δ)-testing of property $\mathcal{P}$ [26])**.** *Let $\varepsilon \in (0,1)$ be the tolerance and $\delta \in (0,1)$ be the correctness parameter, and let $\mathcal{P}$ be a property consisting of k-tuples of distributions over $\mathcal{X}$. We say that a (possibly randomized) algorithm $\mathcal{A} : \mathcal{X}^m \to \{0,1\}$ is an (ε, δ)-tester for $\mathcal{P}$ if, given a bag T of independent samples drawn from distributions $\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(k)}$, and denote by r the internal randomness of $\mathcal{A}$, it satisfies the following conditions:*

(1) ***Completeness:*** *if $(\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(k)}) \in \mathcal{P}$, then $\mathbb{E}_{r,T}[\mathcal{A}(T; r)] \geq 1 - \delta$.*

(2) ***Soundness:*** *if $k^{-1} \sum_{i=1}^{k} d_{TV}(\mathbf{p}^{(i)}, \mathbf{q}^{(i)}) \geq \varepsilon$ for every tuple $(\mathbf{q}^{(1)}, \dots, \mathbf{q}^{(k)}) \in \mathcal{P}$, then $\mathbb{E}_{r,T}[\mathcal{A}(T; r)] \leq \delta$.*

In particular: (a) uniformity testing over $[n]$ corresponds to the case $k = 1$, where $\mathcal{P}$ contains only the uniform distribution on $[n]$; (b) closeness testing over $[n]$ corresponds to the case $k = 2$, where

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

$\mathcal{P}$ consists of all pairs of the form $(\mathbf{p}, \mathbf{p})$ over $[n]$; and (c) independence testing over $[n_1] \times [n_2]$ corresponds to the case $k = 1$, where $\mathcal{P}$ consists of all product distributions over that domain.

Most works on distribution testing have focused on the setting of $(\varepsilon, 1/3)$-testing, and the statistical cost of testing for a variety of natural properties have long been well understood. In contrast, the sample complexity of (ε, δ)-testing is far less understood: prior to the work of [27], there were no known techniques beyond the standard trick to obtain an (ε, δ)-tester with $O(\log(1/\delta))$ overhead. Even now, in contrast to $(\varepsilon, 1/3)$-testing, the sample complexity of (ε, δ)-testing has been established for comparatively few distribution testing problems [27, 26].

Another shortcoming of $(\varepsilon, 1/3)$-testers (and in fact (ε, δ)-testers) can be seen when the underlying distribution is neither in $\mathcal{P}$ nor sufficiently far from $\mathcal{P}$. In this event, a tester may yield inconsistent answers. Given that distribution testing is a fundamental task and a common sub-routine (e.g. bias estimation) in statistical algorithms, the above shortcomings could hamper both the accuracy and stability of the resulting algorithms. To this end, researchers have proposed several stronger conditions on the tester. In this work, we will focus on two: high probability testing and replicability [40] (see Section 1.2 for a discussion of other variants of distribution testing).

High probability testing asks whether we can ensure the tester errs with probability δ for some arbitrarily small $\delta > 0$. Using a standard boosting trick (e.g. majority vote), we can obtain a high probability tester with $O(m \log(1/\delta))$ samples from a m-sample $(\varepsilon, 1/3)$-tester. However, for many fundamental distribution testing problems such as uniformity, closeness, and independence, this standard trick is provably sub-optimal! In particular, for all three tasks, improved (and provably optimal) testers were designed through the clever design and analysis of novel test statistics [27, 26].

Replicability [40] is the notion that algorithms should remain stable under a resampling of their inputs (given access to shared randomness). The formal notion is as follows.

Definition 1.2 (ρ-replicability [40])**.** *A randomized algorithm $\mathcal{A} : \mathcal{X}^m \to \mathcal{Y}$ is ρ-replicable if, for every distribution $\mathbf{p}$ on $\mathcal{X}$, $\Pr_{r,T,T'}\left[\mathcal{A}(T; r) = \mathcal{A}(T'; r)\right] \geq 1 - \rho$, where $T, T' \sim_{i.i.d.} \mathbf{p}^{\otimes m}$ and r denotes the internal randomness of $\mathcal{A}$.*

In particular, even for a distribution $\mathbf{p}$ that is neither in $\mathcal{P}$ or far from $\mathcal{P}$, we must ensure that our algorithm is consistent when run twice on $\mathbf{p}$. Similarly to high probability testing, previous works have established sample-optimal testers for uniformity [45], closeness and independence [1, 25].

Curiously, the optimal sample complexity of high probability testers and replicable testers seem to be intimately related. Consider as a motivating example, uniformity testing, perhaps the most fundamental distribution testing task: which asks us to determine whether $\mathbf{p}$ is uniform, or far from uniform in total variation distance. The optimal sample complexity for high probability uniformity testing [27] is:

$$\Theta\left(n^{1/2}\varepsilon^{-2}(\log(1/\delta))^{1/2} + \varepsilon^{-2}\log(1/\delta)\right).$$

In comparison the sample complexity of replicable uniformity testing [45, 25, 1] is

$$\Theta\left(n^{1/2}\varepsilon^{-2}\rho^{-1} + \varepsilon^{-2}\rho^{-2}\right).$$

That is, the sample complexities are identical if we replace ρ^{-1} with $\sqrt{\log(1/\delta)}$! Interestingly, this trend holds for all distribution testing problems studied so far in both the high probability and replicable settings, leading to the central question behind this work:

Are high probability testing and replicable testing statistically equivalent?

In particular, is the above relationship simply a coincidence, or can we show that any tester satisfying one notion can be efficiently transformed into a tester satisfying the other?

In this work, we make significant progress towards answering this question: we show that any tester that satisfies a strong notion of ρ-replicability (Definition 1.3) can be efficiently converted into a high probability tester with $\delta = \exp(-\tilde{\Omega}(\rho^{-2}))$. In the converse direction, we exhibit a simple task with a 1-sample high probability tester where any replicable tester requires $\Omega(\rho^{-2})$ samples, establishing that not every high probability tester can be converted into a replicable one.

1.1 Our Contributions

We now formally state our results, beginning with our definition of strong ρ-replicability, which we define below.

Definition 1.3 (Strong ρ-Replicability)**.** *For $\rho \in (0,1)$, a randomized algorithm $\mathcal{A} : \mathcal{X}^m \to \{0,1\}$ is strongly ρ-replicable if for any set $T \in \mathcal{X}^{2m}$ of size $2m$ and uniformly selected size m random subsets S and S' of T we have that*

$$\Pr_{r,S,S' \sim_u \binom{T}{m}} [\mathcal{A}(S;r) \neq \mathcal{A}(S';r)] < \rho$$

where r denotes the shared internal randomness of $\mathcal{A}$. Furthermore, for $\varepsilon, \delta \in (0,1)$ and a property $\mathcal{P}$ consisting of k-tuples of distributions over the domain $\mathcal{X}$, we say that $\mathcal{A}$ is a strongly $(\varepsilon, \delta, \rho)$-replicable $\mathcal{P}$ tester if it additionally an (ε, δ)-tester for $\mathcal{P}$.

In particular, strong replicability requires that our algorithm is consistent over subsamples from arbitrary size $2m$ subsets of $\mathcal{X}$, where (standard) replicability imposes this constraint on subsets obtained by taking i.i.d. samples from $\mathbf{p}$. We remark that strong replicability is indeed a stronger notion than replicability (see Lemma 3.4 which establishes that any strongly replicable tester implies a replicable tester, while Proposition 4.3 exhibits a replicable tester that is not strongly replicable). Our main result shows that *any* strongly replicable tester with a constant correctness of $1/3$ can be efficiently converted into a high-probability tester.

Theorem 1.4. *Suppose there is a strongly $(\varepsilon, 1/3, \rho)$-replicable tester using m samples. Then there is a high-probability tester using $2m$ samples and having error*

$$\delta = \exp\left(-\Omega\left(\frac{\rho^{-2}}{\log(1/\rho)\log(m)}\right)\right).$$

Moreover, the high probability tester runs in near linear time in the sample complexity with $O(\log(1/\delta))$ oracle calls to the strongly ρ-replicable tester.

We observe that our conversion is optimal up to polylogarithmic factors. The following table exhibits our lower bound for strong replicability for several classic problems.

Table 1: Sample complexity bounds for some distribution testing problems. The upper bounds are known for ρ-replicable testers, while the lower bounds are for strongly ρ-replicable testers obtained via Theorem 1.4. The bounds match up to polylogarithmic factors.

Property	Upper Bound (Replicability)	Lower Bound (Strong Replicability)
Bias	$O\left(\frac{1}{\varepsilon^2\rho^2}\right)$ [40]	$\widetilde{\Omega}\left(\frac{1}{\varepsilon^2\rho^2}\right)$
Uniformity	$\widetilde{O}\left(\frac{\sqrt{n}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$ [45, 1]	$\widetilde{\Omega}\left(\frac{\sqrt{n}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$
Closeness	$O\left(\frac{n^{2/3}}{\varepsilon^{4/3}\rho^{2/3}} + \frac{\sqrt{n}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$ [1, 25]	$\widetilde{\Omega}\left(\frac{n^{2/3}}{\varepsilon^{4/3}\rho^{2/3}} + \frac{\sqrt{n}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$
Independence	$\widetilde{O}\left(\frac{n_1^{2/3}n_2^{1/3}}{\varepsilon^{4/3}\rho^{2/3}} + \frac{\sqrt{n_1 n_2}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$ [1, 25]	$\widetilde{\Omega}\left(\frac{n_1^{2/3}n_2^{1/3}}{\varepsilon^{4/3}\rho^{2/3}} + \frac{\sqrt{n_1 n_2}}{\varepsilon^2\rho} + \frac{1}{\varepsilon^2\rho^2}\right)$

In particular, we exhibit strongly replicable testers for bias estimation and closeness testing with optimal sample complexity. It remains a very interesting open question to determine whether we can obtain a similar transformation using only replicable testers as a starting point.

Theorem 1.5 (Theorem A.1 and Theorem A.2)**.** *There are $(\varepsilon, 1/3, \rho)$-strongly replicable testers for the following tasks:*

1. *Bias estimation, with $O(\rho^{-2}\varepsilon^{-2})$ samples.*
2. *Closeness testing with $O(n^{2/3}\rho^{-2/3}\varepsilon^{-4/3} + n^{1/2}\rho^{-1}\varepsilon^{-2} + \rho^{-2}\varepsilon^{-2})$ samples.*

As a corollary, we see that Theorem 1.4 is optimal up to logarithmic factors. For example, the first tester implies an (ε, δ)-tester for bias estimation with $\tilde{O}(\varepsilon^{-2}\log(1/\delta))$ samples. Finally, we exhibit

a simple property (identity testing) where high probability testing provably requires fewer samples than replicability.

Proposition 1.6. *There exists a property $\mathcal{P}$ (identity testing) for which there is an $(1, 0)$-tester with 1-sample but any $(1, \rho)$-replicable tester requires $\Omega(\rho^{-2})$ samples.*

The construction of Proposition 1.6 is simple: we test whether a distribution on $\{0, 1\}$ is fully supported on 0. The only distribution that is 1-far from the distribution is the distribution $\mathbf{q}$ fully supported on 1. There is a simple 1-sample tester with error 0: take one sample, accept if 0 and reject if 1. However, any replicable algorithm for this problem requires $\Omega(\rho^{-2})$ samples [38]. Although this rules out an efficient reduction from replicable testing to high probability testing in the most general sense, the provided counterexample is degenerate (i.e. $\varepsilon = 1$) in that it allows for an exceptionally simple high probability tester. Indeed, it may well be possible that there is an efficient reduction from replicable testing to high probability testing under some reasonable or natural assumptions. We leave the existence of such a reduction (and the appropriate accompanying assumptions) as an exciting direction for future research).

1.2 Related Work

We discuss various related work in this section.

Replicability Since its introduction, replicability has been studied in the context of a wide range of machine learning tasks, including multi-arm bandits [34], clustering [35], reinforcement learning [43, 33, 39], halfspace learning [41], and high-dimension statistics [37]. A related line of work explored the connection between replicability and other algorithmic stability notions such as differential privacy [12, 47], total variation indistinguishability [42], global stability [23], and one-way perfect generalization [12, 47].

Variants of Distribution Testing Several variants of distribution testing have been proposed, including but not limited to (1) tolerant testing where the algorithm must also accept samples from distributions close to $\mathcal{P}$ [50], (2) generalized testing, where the algorithm does not know the underlying domain [4], (3) robust distribution testing where the algorithm has access to noisy, corrupted distributions [11], and (4) conditional sampling, where the algorithm has additional access to conditional distributions [21, 14, 20, 10, 48].

Outline. In Section 2, we define the necessary preliminaries. In Section 3, we discuss (strong) ℓ_1-replicability and (strong) replicability. In Section 4, we prove our main result: Theorem 1.4. In Section A, we prove that Theorem 1.4 is optimal and provide a few examples of strongly replicable testers. In Section B and Section C, we provide omitted proofs for Section 3 and Section 4. In Section D, we prove Proposition 1.6.

2 Preliminaries

We write $[m] := \{1, \dots, m\}$ for a positive integer m. Throughout, a *bag* means a finite multiset, so repeated samples are allowed. If T is a bag, then $|T|$ denotes its cardinality counting multiplicity. For a finite bag T with $|T| \geq m$, we let $\binom{T}{m}$ denote the set of all size-m subbags of T, and $X \sim_u T$ denote a uniformly random element drawn from T.

For a distribution $\mathbf{p}$ on $\mathcal{X}$, we write $T \sim \mathbf{p}^{\otimes m}$ to mean that T is a bag of m independent samples drawn from $\mathbf{p}$, and similarly $(\mathbf{p}^{(1)}, \dots, \mathbf{p}^{(k)})$ denotes a k-tuple of distributions on $\mathcal{X}$. Expectations are written as $\mathbb{E}[\cdot]$, and $i.i.d.$ means "independent and identically distributed." We write $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ for total variation distance, namely

$$d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{p}(x) - \mathbf{q}(x)|.$$

When $\mathcal{X} = [n_1] \times [n_2]$, a product distribution means one of the form $\mathbf{p}_1 \otimes \mathbf{p}_2$ on that domain. Finally, unless stated otherwise, probabilities and expectations are taken over all sources of randomness appearing in the expression, including the random sample bags and the tester's internal random seed.

We will use the following concentration inequality to bound the deviation of a martingale from its starting expectation.

Fact 1 (Azuma-Hoeffding (see e.g. Thm 3.2.1 of [51]))**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a submartingale or supermartingale with respect to a sequence $\{Y_n\}_{n\in\mathbb{N}}$. If for all $n = 1, 2, \ldots$ it holds that $|X_n - X_{n-1}| \leq c_n$ almost surely, then for any $\varepsilon > 0$ and $n \in \mathbb{N}$:*

$$\Pr[X_n - X_0 \geq \varepsilon] \leq \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Definition 2.1 (High-probability tester)**.** *Let $\varepsilon, \delta \in (0, 1)$, and let $\mathcal{P}$ be a property of distributions. We say that a randomized algorithm is a high-probability tester for $\mathcal{P}$ with error δ if it solves the (ε, δ)-testing problem for $\mathcal{P}$. An algorithm is an ε-tester if it is a $(\varepsilon, 1/3)$-tester.*

3 Equivalence of Standard Replicability and ℓ_1-replicability

Throughout this work, it will be convenient to work with an alternative equivalent notion of replicability, which we call ℓ_1-replicability. Comparing to standard replicability, there are two main differences. First, an ℓ_1-replicable tester produces continuous outputs in $[0, 1]$ rather than discrete values of 0 or 1, so the correctness condition bounds the expected value of the tester's output rather than its failure probability. Second, instead of bounding the probability of specific outcomes, ℓ_1-replicability constrains the expected ℓ_1 distance between the tester's outputs. A similar characterization of replicability appears in [1]. Formally,

Definition 3.1 ($(\varepsilon, \delta, \rho)$-$\ell_1$-replicable $\mathcal{P}$ tester)**.** *For $\varepsilon, \delta, \rho \in (0, 1)$ and a property $\mathcal{P}$ consisting of k-tuples of distributions over the domain $\mathcal{X}$, an algorithm $\mathcal{A} : \mathcal{X}^m \to [0, 1]$ is a ρ-ℓ_1-replicable if given independent bags T and T', where each bag consists of m independent samples drawn from distributions within the arbitrary collection $\mathbf{p}^{(1)}, \cdots, \mathbf{p}^{(k)}$ of distributions over $\mathcal{X}$,*

$$\mathbb{E}_{T,T'}\left[|\mathcal{A}(T) - \mathcal{A}(T')|\right] < \rho.$$

Moreover, $\mathcal{A}$ is an $(\varepsilon, \delta, \rho)$-$\ell_1$-replicable tester for $\mathcal{P}$ if it additionally satisfies the conditions in Definition 1.1.

We will show that ℓ_1-replicability is equivalent to standard replicability, up to an arbitrarily small loss in correctness and replicability parameters during the conversion. Our main result will require an analogous notion of strong replicability, which we define below. Thus, we will interchangeable use the equivalent notions of replicability. Analogously to strong replicability (Definition 1.3) strong ℓ_1-replicability requires the algorithm's output to be stable across random subsets drawn from any fixed, worst-case pool of samples of size $2m$, whereas ℓ_1-replicability just requires stability across fresh sample sets drawn from the input distribution.

Definition 3.2 (Strong ρ-ℓ_1-Replicability)**.** *For $\rho \in (0, 1)$, an algorithm $\mathcal{A} : \mathcal{X}^m \to [0, 1]$ is strongly ρ-ℓ_1-replicable if for any set $T \in \mathcal{X}^{2m}$ of size $2m$ and uniformly selected size n random subsets S and S' of T we have that*

$$\mathbb{E}_{S,S' \sim_u \binom{T}{m}}\left[|\mathcal{A}(S) - \mathcal{A}(S')|\right] < \rho.$$

Furthermore, $\mathcal{A}$ is a strongly $(\varepsilon, \delta, \rho)$-$\ell_1$-replicable $\mathcal{P}$ tester if it additionally satisfies the conditions in Definition 1.1.

In what follows, we will show that there exists computationally tractable conversions between an (strong) ℓ_1-replicable distribution tester and a (strong) replicable distribution tester, with only an arbitrarily small loss in the replicability parameter and correctness parameter.

Lemma 3.3. *For $\varepsilon, \rho, \eta, \delta \in (0, 1)$, and a property $\mathcal{P}$ consisting of k-tuples of distributions over the domain $\mathcal{X}$, a $(\varepsilon, \delta, \rho)$-replicable $\mathcal{P}$ tester can be converted into a $(\varepsilon, \delta + \eta, \rho + 2\eta)$-$\ell_1$-replicable $\mathcal{P}$ tester. Conversely, a $(\varepsilon, \delta, \rho)$-$\ell_1$-replicable $\mathcal{P}$ tester can be converted into a $(\varepsilon, \delta, \rho)$-replicable $\mathcal{P}$ tester.*

Moreover, both conversions preserve computational efficiency and sample complexity. Furthermore, the equivalence holds if we replace ℓ_1-replicability and replicability with strong ℓ_1-replicability and strong replicability. The formal proof is deferred to Section B.

Finally, we show that strong replicability implies ℓ_1-replicability at the cost of an extra multiplicative factor of 2 in the stability parameter.

Lemma 3.4 (Properties of strong replicable testers)**.** *For a randomized algorithm $\mathcal{A} : \mathcal{X}^m \to \{0,1\}$, denote by r the internal randomness of $\mathcal{A}$. If $\mathcal{A}$ is ρ-strong-replicable, then $\mathcal{A}$ is (2ρ)-ℓ_1-replicable.*

4 Proof of the Main Theorem

In this section, we prove Theorem 1.4. Given a strongly replicable tester with constant correctness, we construct the high probability tester by evaluating the strongly replicable tester on independent random subbag of size m drawn from the input, and output the average. In this way, we can take advantage of strong replicability, and show a concentration bound on $F(T) := \mathbb{E}_{S\sim_u \binom{T}{m}}[\mathcal{A}(S)]$ for a randomly drawn T. Via Lemma 3.3, we may assume that $\mathcal{A}$ is in fact a strongly $(\varepsilon, 0.4, 2\rho)$-$\ell_1$-replicable tester (e.g. by choosing $\eta \le \min(0.01, \rho)$). The formal construction of the high probability tester is as follows.

Algorithm 1 HIGHPROBABILITYTESTER

Input: A set T_0 of $2m$ samples, a ρ-strong-replicable tester $\mathcal{A}$.
Parameters: Failure probability $\delta \in (0,1)$, sample parameter m.
Output: 1 for the completeness case or 0 for the soundness case.

1: $I \leftarrow \Theta(\log(1/\delta))$.
2: Draw I independent random size-m subsets $S_1, \ldots, S_I \sim_u \binom{T_0}{m}$.
3: Compute the estimate of the statistic:

$$\hat{F}(T_0) \leftarrow \frac{1}{I}\sum_{i\in[I]} \mathcal{A}(S_i).$$

4: **return** 0 if $\hat{F}(T_0) < 1/2$, and 1 otherwise.

We next proceed to show the concentration bound on $F(T)$, which is the key result in this section towards showing the correctness of Algorithm 1.

Proposition 4.1. *Let R be a set consisting of $2m$ pairs of elements (i.e. $4m$ elements in total). Let $T \subset R$ be selected by picking an element from each pair uniformly at random. Then $F(T)$ is within 0.01 of its expected value with probability $1 - \exp\left(-\Omega\left(\frac{\rho^{-2}}{\log(1/\rho)\log(m)}\right)\right)$.*

Proof of Proposition 4.1. For simplicity, denote by $R := \{R_1, \cdots, R_{2m}\}$, where R_i is a pair of elements for each $i = 1, \cdots, 2m$. The high level idea is that we first consider generating T in a particular way: starting with all of R we will consider one pair R_i at a time in ascending order in i, and randomly select which element of the pair should end up in T, and denote by $Y_1, \cdots, Y_m$ the Bernoulli random variables representing the choice of the element in the pair for each step. By the principle of deferred decisions, the resulting T is identically distributed as the T drawn as in premise. We will argue that this makes the sequence $\{\mathbb{E}[F]\} \cup \{\mathbb{E}[F(T)|X_1, \cdots, X_i]\}_{i=1}^m$ of expectation of $F(T)$ given our previous choices into a martingale. We will then use Azuma's Inequality to show concentration on the final value of $F(T)$. To do this, we will want to know that each decision contributes relatively little to the expectation. For this, we will first need an averaged concentration after having made some but not all of our decisions.

Lemma 4.2. *Let Q be a set consisting of $2m - k$ elements and k pairs of elements for some $0 \le k \le 2m$. Let $T \subset Q$ be randomly selected by using all singletons in Q and selecting a random element of each pair. Then if T' is generated independently by the same procedure as T, we have that*

$$\mathbb{E}_{T,T'}\left[|F(T) - F(T')|\right] \le 6\rho.$$

Proof of Lemma 4.2. Letting $S \subset T$ and $S' \subset T'$ be uniformly random subsets of size m, chosen independently conditional on T and T', we claim that it suffices to bound $\mathbb{E}_{S,S'}\left[|\mathcal{A}(S) - \mathcal{A}(S')|\right]$. Indeed, by Jensen's inequality we have that

$$\begin{aligned}\mathbb{E}_{T,T'}\left[|F(T) - F(T')|\right] &= \mathbb{E}_{T,T'}\left[\left|\mathbb{E}_{S\sim_u\binom{T}{m}, S'\sim_u\binom{T'}{m}}\left[\mathcal{A}(S) - \mathcal{A}(S')\right]\right|\right] \\ &\le \mathbb{E}_{T,T'}\left[\mathbb{E}_{S,S'}\left[|\mathcal{A}(S) - \mathcal{A}(S')|\right]\right].\end{aligned}$$

We index the k pairs and $2m-k$ singletons in Q by [2m], using the corresponsing indexing for T and T'. To bound the desired quantity, let us view S, S' as random size m subsets of $[2m]$ (i.e. the indices that define the respective subsets of T, T'.) Let $S_1 \subset T$ be the size m subset consisting of the elements of T whose indices are not selected by S' in T'.

We define the set T_1 as follows. T_1 contains all singletons in Q. For each index $j \in [2m]$ corresponding to a pair in Q, if the index $j \in S'$, we include the element of that pair in T' (and therefore S'), and otherwise if the index $j \in S_1$, we include the element of that pair in T (and therefore S_1). Finally, let $S_2 \subset T_1$ be a random subset of T_1 of size m (again, we may view S_2 as a random size m subset of $[2m]$.) We claim the following:

1. S, S_1 are random, independent size m subsets of T.
2. S_1, S_2 are random, independent size m subsets of T_1.
3. S_2, S' are random, independent size m subsets of T_1.

Towards the first point, we observe that S_1 is independent of S since it depends only on S' which is independent of S. Note that S is a random size m subset of T by definition. Furthermore, since the indices of S' is a random size m subset of $[2m]$, its complement S_1 is a random size m subset of $[2m]$, and therefore a random size m subset of T, as desired.

Towards the second and third points, we observe that S_2 is by definition a random size m subset of T_1, and thus it suffices to argue that S', S_1 are random (not necessarily independent) size m subsets of T_1. To see this, we view S', S_1 again as subsets of indices of $[2m]$, and note that S' is a random size m subset while S_1 is its complement. For any fixed choice of the indices of S', T_1 contains all singletons and an independent uniform choice from each pair in Q, so its distribution does not depend on those indices. Thus, conditional on T_1, the indices of S' and their complement defining S_1 remain uniform, and the fresh uniform choice defining S_2 is independent of both.

Thus, since $\mathcal{A}$ is 2ρ-strong-replicable, we have that

$$\mathbb{E}\left[|\mathcal{A}(S) - \mathcal{A}(S_1)|\right], \mathbb{E}\left[|\mathcal{A}(S_1) - \mathcal{A}(S_2)|\right], \mathbb{E}\left[|\mathcal{A}(S_2) - \mathcal{A}(S')|\right] \leq 2\rho.$$

Lemma 4.2 then follows by the triangle inequality. □

Suppose that we are at the stage where we have decided all but k of the pairs of R. Let X_i be a $\{\pm 1\}$ random variable based on which element of the i-th pair is eventually chosen. This makes $F(T)$ (conditioned on our $2m-k$ choices so far) a function on $\{\pm 1\}^k$ (with the uniform distribution over the hypercube). Since $\mathcal{A}$ has range $[0,1]$, we know that $F(T) \in [0,1]$ and Lemma 4.2 implies

$$\mathbb{E}_{T,T'}[|F(T) - F(T')|] \leq 6\rho.$$

We note that if we decide which element of the i-th pair to use, this changes the expectation of $F(T)$ by $\pm\,\mathbb{E}[X_i F(T)]$ since by law of total probability, we have that $\mathbb{E}[X_i F(T)] = \frac{1}{2}\,\mathbb{E}[F(T)|X_i = 1] + \frac{1}{2}\,\mathbb{E}[-F(T)|X_i = -1]$ and $\mathbb{E}[F(T)] = \frac{1}{2}\,\mathbb{E}[F(T)|X_i = 1] + \frac{1}{2}\,\mathbb{E}[F(T)|X_i = -1]$, then for either $z \in \{\pm 1\}$

$$\mathbb{E}[F(T)|X_i = z] - \mathbb{E}[F(T)] = \mathbb{E}[F(T)] - \mathbb{E}[F(T)|X_i = -z] = z\,\mathbb{E}[X_i F(T)].$$

We thus want to show that there is some i so that $|\,\mathbb{E}[X_i F(T)]|$ is reasonably small. First, without loss of generality, we may assume $\mathbb{E}[X_i F(T)]$ is non-negative for each i, since otherwise we can rename $X_i' = -X_i$, then X_i' is still a uniformaly distributed random variable over $\{\pm 1\}$, but with $\mathbb{E}[X_i' F(T)] > 0$. We then have that

$$\begin{aligned}
\sum_{i=1}^{k} \mathbb{E}[X_i F(T)] &= \mathbb{E}\left[\left(\sum_{i=1}^{k} X_i\right) F(T)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{k} X_i\right) (F(T) - \mu_F)\right] \\
&\leq \mathbb{E}\left[\left|\sum_{i=1}^{k} X_i\right| |F(T) - \mu_F|\right]
\end{aligned}$$

where μ_F is the expectation of $F(T)$ when over the random choice of $X \in \{\pm 1\}^k$, and the second equality follows from the fact that $\mathbb{E}\left[\left(\sum_{i=1}^k X_i\right)\right] = \sum_{i=1}^k \mathbb{E}[X_i] = 0$.

We break this expectation into cases based on whether $\left|\sum_{i=1}^k x_i\right| < \sqrt{Ck\log(1/\rho)}$ for some sufficiently large constant C.

Case 1: $\left|\sum_{i=1}^k X_i\right| < \sqrt{Ck\log(1/\rho)}$**.** We argue that

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^k X_i\right| |F(T) - \mu_F|\right] &\leq \sqrt{Ck\log(1/\rho)}\,\mathbb{E}_T[|F(T) - \mathbb{E}_{T'}[F(T')]|] \\
&= \sqrt{Ck\log(1/\rho)}\,\mathbb{E}_T[|\,\mathbb{E}_{T'}[F(T) - F(T')]|] \\
&\leq \sqrt{Ck\log(1/\rho)}\,\mathbb{E}_{T,T'}[|F(T) - F(T')|] \\
&= O(\rho\sqrt{k\log(1/\rho)}),
\end{aligned}$$

where the second line follows from the fact that T and T' are independent, the third line follows from Jensen's inequality, and the last line follows from Lemma 4.2.

Case 2: $\left|\sum_{i=1}^k X_i\right| \geq \sqrt{Ck\log(1/\rho)}$**.** In the second case, we bound the event that $\sum_{i=1}^k X_i$ has large absolute value. Since $\sum X_i$ is a sum of $\{\pm 1\}$ random variables with mean 0, a standard Hoeffding bound yields $\left|\sum_{i=1}^k X_i\right| \geq \sqrt{Ctk\log(1/\rho)}$ with probability at most $\exp(-2Ctk\log(1/\rho)/(4k)) = \exp(-5t\log(1/\rho)) < \rho^{5t}$ for any $t \geq 1$ and sufficiently large C. In particular, we can bound

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^k X_i\right| |F(T) - \mu_F|\right] &\leq O\left(\mathbb{E}\left[\mathbb{1}_{\left\{|\sum_{i=1}^k X_i| \geq \sqrt{Ck\log(1/\rho)}\right\}} \left|\sum_{i=1}^k X_i\right|\right]\right) \\
&\leq O\left(\sum_{t=2^i} \Pr\left(\left|\sum_{i=1}^k X_i\right| \geq \sqrt{Ctk\log(1/\rho)}\right)\sqrt{2Ctk\log(1/\rho)}\right) \\
&< O\left(\sum_{t=2^i} \sqrt{2Ctk\log(1/\rho)}\rho^{5t}\right) \\
&< O\left(\sqrt{k\log(1/\rho)}\rho\right)\sum_{i=1}^{\infty} 2^i \rho^{4\times 2^i} \\
&= O\left(\sqrt{k\log(1/\rho)}\rho\right).
\end{aligned}$$

In the first equality, we bound $|F(T) - \mu_F| = O(1)$. In the second, we write out the expectation and within each range of values, upper bound the probability that $\sum X_i$ lies in that range. In the final equality, we observe $\rho < 0.1$ implies that the series converges.

Combining with the fact that $\|F(T) - \mu_F\|_\infty = O(1)$, this range gives a smaller contribution. Thus:

$$\sum_{i=1}^k |\,\mathbb{E}[X_i F(T)]| = O(\rho\sqrt{k\log(1/\rho)}).$$

Thus, we can always find an i so that

$$|\,\mathbb{E}[X_i F(T)]| = O\left(\rho\sqrt{\frac{\log(1/\rho)}{k}}\right).$$

We then consider the pairs of R one at a time (always picking the one with $|\,\mathbb{E}[X_i F(T)]|$ smallest) and use Azuma's inequality to get a concentration bound for $F(T)$. In particular, the k-th to last step

has deviation at most $O\left(\rho\sqrt{\log(1/\rho)/k}\right)$, so Azuma's inequality implies that the probability that we deviate more than 0.01 from the mean is at most

$$\exp\left(-\frac{1}{O\left(\sum_{k=1}^{m}\rho^2\log(\rho)/k\right)}\right) = \exp\left(-\frac{\rho^{-2}}{\log(1/\rho)\log(m)}\right).$$

This completes the proof of Proposition 4.1. □

We are now ready to show the main theorem of the paper.

Proof of Theorem 1.4. We will show that Algorithm 1 is the desired high-probability tester.

Note that a random subset of T_0 can be obtained in $O(m)$ time (e.g. by randomly removing one element at a time), so that the running time of our algorithm can be bounded by

$$O\left(\log(1/\delta)(m + \mathrm{Time}(\mathcal{A}))\right)$$

where $\mathrm{Time}(\mathcal{A})$ is the time required to run $\mathcal{A}$. We observe that estimating $\hat{F}(T_0)$ does not require any new samples, as for each $i \in [I]$ we take a random subsample of T_0. It remains to prove the correctness of the algorithm.

For a given set T of size $2m$, let $F(T) := \mathbb{E}_{S\sim_u\binom{T}{m}}[\mathcal{A}(S)]$. We first note that $\hat{F}(T_0)$ is a decent approximation of $F(T_0)$, i.e. $|\hat{F}(T_0) - F(T_0)| \le 0.01$. This follows by a standard concentration bound. In particular, $\hat{F}(T_0)$ is a sum of I i.i.d. Bernoulli random variables with mean $F(T_0)$, so that a standard Hoeffding bound ensures that $|\hat{F}(T_0) - F(T_0)| \le 0.01$ with probability $1 - \delta/2$ whenever $I = C\log(1/\delta)$ for some sufficiently large constant C. Thus, conditioning on the event that $|\hat{F}(T_0) - F(T_0)| \le 0.01$, suppose that $\mathbf{p}$ is in $\mathcal{P}$ (an analogous argument applies in the soundness case). Let T and T' be two random sets of $2m$ samples from $\mathbf{p}$. Let $R = T \cup T'$, with the i-th sample of T paired with the i-th sample of T', and let T'' be obtained by picking a random element from each pair. By Proposition 4.1, with probability at least $1 - \exp\left(-\Omega\left(\frac{\rho^{-2}}{\log(1/\rho)\log(m)}\right)\right)$, both $|F(T) - F(T'')|$ and $|F(T') - F(T'')|$ are at most 0.01, since (T, T''), and (T', T'') are independent subsamples of R as specified in Proposition 4.1.

This means that for some $\delta = \exp\left(-\Omega\left(\frac{\rho^{-2}}{\log(1/\rho)\log(m)}\right)\right)$, with probability at least $1 - \delta/2$ over T, we have that with probability at least 0.99 over T', $|F(T) - F(T')| < 0.02$, which implies that $|F(T) - \mathbb{E}_{T'}[F(T')]| < 0.03$ (since either $|F(T) - F(T')| < 0.02$ or with probability at most 0.01, $|F(T) - F(T')| \le 1$). On the other hand, T' is independent of T and therefore satisfies

$$\mathbb{E}_{T'}[F(T')] = \mathbb{E}_{T'}[\mathbb{E}_{S'\subset T'}\,\mathcal{A}(S')] = \mathbb{E}[\mathcal{A}(S')] > 0.6.$$

Thus, with probability at least $1 - \delta$, we have that

$$\hat{F}(T_0) \ge F(T_0) - 0.01 \ge \mathbb{E}_{T'}[F(T')] - |F(T_0) - \mathbb{E}_{T'}[F(T')]| - 0.01 > \mathbb{E}_{T'}[F(T')] - 0.04 > \frac{1}{2},$$

where the second inequality follows from the reverse triangle inequality, which completes the proof. □

We conclude this section with some discussion regarding the result. First, we remark that a slightly stronger result can be obtained under an alternative notion of strong replicability. In particular, if $\mathcal{A}$ is strongly ρ-ℓ_2-replicable, (i.e. $\mathbb{E}\left[(\mathcal{A}(S) - \mathcal{A}(S'))^2\right] < \rho^2$ where S, S' are random size m subsets of T), then we can instead bound using Cauchy-Schwarz,

$$\mathbb{E}\left[\left(\sum_{i=1}^{k} X_i\right)(F(T) - \mu_F)\right] \le \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{k} X_i\right)^2\right]\mathbb{E}[(F(T) - \mu_F)^2]} = O(\rho\sqrt{k}).$$

Here, we use Cauchy-Schwarz, Lemma 4.2 with ℓ_2-replicability instead of ℓ_1-replicability, and then observe that since X_i are independent random variables in $\{\pm 1\}$, $\mathbb{E}[(\sum X_i)^2] = \sum_i \mathbb{E}[X_i^2] = k$.

Finally, we observe that we can construct testers that are replicable but neither strongly replicable testers nor high probability testers.

Proposition 4.3. *There exists a property $\mathcal{P}$ such that there exists a $(\varepsilon, 1/3, \rho)$-replicable tester $\mathcal{A}$ for $\mathcal{P}$ where $\mathcal{A}$ is not an $(\varepsilon, 1/3, \rho)$-strongly replicable tester for $\mathcal{P}$. Furthermore, $\mathcal{A}$ is not a (ε, δ)-tester for $\mathcal{P}$ where $\delta = \exp(-\tilde{\Omega}(\rho^{-2}))$. The formal proof is deferred to Section C.*

Note that Proposition 4.3 does not rule out an efficient transformation from replicable testers to high probability testers. In particular, for the given property $\mathcal{P}$ (all distributions), there is a simple 0-sample tester (always accept) that is both a replicable and high probability tester.

Conclusion. We introduce the notion of strong replicability and show that any distribution tester satisfying ρ-strong replicability can be efficiently converted into a high probability tester with error $\delta = \exp(-\tilde{\Omega}(\rho^{-2}))$ with the same sample complexity. As a corollary, we obtain lower bounds for strong replicability that match upper bounds for replicability for several distribution testing problems (bias, uniformity/identity, closeness, and independence). For bias and closeness testing, we further obtain sample-optimal strongly replicable testers. It remains an interesting open question to determine whether a similar reduction can be obtained for replicable testers.

In the reverse direction, we exhibit a task for which high probability testers can use 1 sample but ρ-replicable testers require $\Omega(\rho^{-2})$ samples. However, given that the task instance is somewhat contrived (it is designed so that an exceedingly simple high probability tester exists), we leave as an interesting open question to determine under what assumptions a reduction might exist.

References

[1] A. Aamand, M. Aliakbarpour, J. Y. Chen, S. Narayanan, and S. Silwal. On the structure of replicable hypothesis testers. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5771–5823. SIAM, 2026. 1, 1, 3

[2] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011. 1

[3] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. *Advances in Neural Information Processing Systems*, 28, 2015. 1

[4] T. Batu and C.L. Canonne. Generalized uniformity testing. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 880–889. IEEE, 2017. 1.2

[5] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *ACM Symposium on Theory of Computing*, pages 678–687, 2002. 1

[6] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451. IEEE, 2001. 1

[7] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science*, 2000. 1

[8] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):1–25, 2013. 1

[9] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004. 1

[10] R. Bhattacharyya and S. Chakraborty. Property testing of joint distributions using conditional samples. *ACM Transactions on Computation Theory (TOCT)*, 10(4):1–20, 2018. 1.2

[11] H. Black and C. Ye. Distribution testing in the presence of arbitrarily dominant noise with verification queries. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4410–4480. SIAM, 2026. 1.2

[12] M. Bun, M. Gaboardi, M. Hopkins, R. Impagliazzo, R. Lei, T. Pitassi, S. Sivakumar, and J. Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023. 1.2

[13] C. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, 19(6):1032–1198, 2022. 1

[14] C. Canonne, D. Ron, and R. A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1174–1192. SIAM, 2014. 1.2

[15] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020. 1

[16] C. L. Canonne, I. Diakonikolas, D. M. Kane, and S. Liu. Near-optimal bounds for testing histogram distributions. *CoRR*, abs/2207.06596, 2022. Conference version in NeurIPS'22. 1

[17] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing Bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 370–448, 2017. 1

[18] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, 2018*, pages 735–748. ACM, 2018. 1

[19] C. L. Canonne, A. Jain, G. Kamath, and J. Li. The price of tolerance in distribution testing. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 573–624. PMLR, 2022. 1

[20] C. L. Canonne, D. Ron, and R. A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015. 1.2

[21] S. Chakraborty, E. Fischer, Y. Goldhirsh, and A. Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 561–580, 2013. 1.2

[22] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014. 1

[23] Z. Chase, B. Chornomaz, S. Moran, and A. Yehudayoff. Local Borsuk-Ulam, Stability, and Replicability. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, 2024*, pages 1769–1780. ACM, 2024. 1.2

[24] C. Daskalakis, N. Dikkala, and G. Kamath. Testing ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, 2018*, pages 1989–2007. SIAM, 2018. 1

[25] I. Diakonikolas, J. Gao, D. Kane, S. Liu, and C. Ye. Replicable distribution testing. *arXiv preprint arXiv:2507.02814*, 2025. Conference version in NeurIPS'25. 1, 1, A, 5, A

[26] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price. Optimal testing of discrete distributions with high probability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 542–555, 2021. 1.1, 1

[27] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, pages 41–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2018. 1, 1

[28] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016. 1

[29] I. Diakonikolas, D. M. Kane, and S. Liu. Testing closeness of multivariate distributions via Ramsey theory. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, 2024*, pages 340–347. ACM, 2024. 1

[30] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1183–1202. IEEE, 2015. 1

[31] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 1841–1854. SIAM, 2015. 1

[32] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, volume 80 of *LIPIcs*, pages 8:1–8:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. 1

[33] E. Eaton, M. Hussing, M. Kearns, and J. Sorrell. Replicable reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023. 1.2

[34] H. Esfandiari, A. Kalavasis, Karbasi A, A. Krause, V. Mirrokni, and G. Velegkas. Replicable bandits. In *The Eleventh International Conference on Learning Representations, ICLR 2023, 2023*. OpenReview.net, 2023. 1.2

[35] H. Esfandiari, A. Karbasi, V. Mirrokni, G. Velegkas, and F. Zhou. Replicable clustering. *Advances in Neural Information Processing Systems*, 36:39277–39320, 2023. 1.2

[36] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75, 2011. 1

[37] M. Hopkins, R. Impagliazzo, D. Kane, S. Liu, and C. Ye. Replicability in high dimensional statistics. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1–8. IEEE, 2024. 1.2

[38] M. Hopkins, R. Impagliazzo, and C. Ye. Approximate replicability in learning. *arXiv preprint arXiv:2510.20200*, 2025. 1.1, D, D.1

[39] M. Hopkins, S. Liu, C. Ye, and Y. Yoshida. From generative to episodic: Sample-efficient replicable reinforcement learning. *arXiv preprint arXiv:2507.11926*, 2025. 1.2

[40] R. Impagliazzo, R. Lei, T. Pitassi, and J. Sorrell. Reproducibility in learning. In *Proceedings of the 54th annual ACM SIGACT symposium on theory of computing*, pages 818–831, 2022. 1, 1.2, 1

[41] A. Kalavasis, A. Karbasi, K. G. Larsen, G. Velegkas, and F. Zhou. Replicable learning of large-margin halfspaces. In *Forty-first International Conference on Machine Learning*, 2024. 1.2

[42] A. Kalavasis, A. Karbasi, S. Moran, and G. Velegkas. Statistical indistinguishability of learning algorithms. In *International Conference on Machine Learning, ICML*, 2023. 1.2

[43] A. Karbasi, G. Velegkas, L. Yang, and F. Zhou. Replicability in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:74702–74735, 2023. 1.2

[44] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011. 1

[45] S. Liu and C. Ye. Replicable uniformity testing. *Advances in Neural Information Processing Systems*, 37:32039–32075, 2024. 1, 1

[46] G. Lugosi. *Concentration of measure inequalities*. 2006. A

[47] S. Moran, H. Schefler, and J. Shafer. The bayesian stability zoo. *CoRR*, abs/2310.18428, 2023. 1.2

[48] S. Narayanan. On tolerant distribution testing in the conditional sampling model. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 357–373. SIAM, 2021. 1.2

[49] L. Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1

[50] M. Parnas, D. Ron, and R. Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006. 1.2

[51] S. Roch. *Modern discrete probability: An essential toolkit*. Cambridge University Press, 2024. 1

[52] R. Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012. 1

[53] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017. 1

[54] P. Valiant. Testing symmetric properties of distributions. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 383–392, 2008. 1

Supplementary Material

This appendix is organized as follows. Section A contains the missing proofs of Theorem 1.5 showing a strongly replicable (Definition 1.3) bias estimator and a strongly replicable closeness tester. Section B provides the omitted proofs of Section 3 regarding the connections between ℓ_1-replicability, standard replicability and strong replicability. In Section C, we provide the complete proof of Proposition 4.3, which demonstrates a replicable tester that fails to satisfy both strong replicability and high probability requirements. Finally, Section D provides the full proof of Proposition 1.6, showing that for identity testing, high probability is easier than replicability.

A Strongly Replicable Testers

In this section, we give examples of algorithms for distribution testing are strongly replicable (Definition 1.3). We begin with bias estimation: determine where an unknown Bernoulli distribution has bias less than p_0 or greater than $p_0 + \varepsilon$ for some $p_0 \in [0, 1]$.

Theorem A.1. *There is a strongly $(\varepsilon, 1/3, \rho)$-replicable algorithm for bias estimation with $O(\rho^{-2}\varepsilon^{-2})$ samples.*

Proof of Theorem A.1. Consider the algorithm that returns

$$\min\left(1, \max\left(0, \frac{\hat{p} - (p_0 + \varepsilon/4)}{\varepsilon/2}\right)\right)$$

where $\hat{p}$ denotes the empirical mean. Correctness of the empirical mean with $O(\varepsilon^{-2})$ samples follows from standard analysis, so it suffices to argue that the empirical mean with $O(\rho^{-2}\varepsilon^{-2})$ samples is strongly replicable.

Fix any set $T \in \{0,1\}^{2m}$ where a p-fraction of T is 1. Let $\mathcal{A}(S)$ denote a random variable representing the empirical mean computed from a random sample $S \subset T$ of size m. In particular, $\mathcal{A}(S)$ returns $(\frac{1}{m}\sum_{i=1}^{m} X_i - p_0 - \varepsilon/4)/(\varepsilon/2)$ clipped between $[0, 1]$ where X_i is the i-th sample included into S. Since clipping cannot increase variance, we bound the variance of $\mathcal{A}(S)$ as

$$\mathrm{Var}(\mathcal{A}(S)) \le \frac{4}{\varepsilon^2 m^2}\,\mathrm{Var}\left(\sum_i X_i\right) = \frac{4}{\varepsilon^2 m^2}\left(\sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)\right).$$

Now, we argue that X_i, X_j are negatively correlated. In particular

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] = \Pr(X_j = 1)(\Pr(X_i = 1 | X_j = 1) - \Pr(X_i = 1)) \le 0$$

since conditioned on $X_j = 1$, X_i is less likely to sample a 1 from T. Combined with the previous equation, we have

$$\mathrm{Var}(\mathcal{A}(S)) \le \frac{4\sum_{i=1}^m \mathrm{Var}(X_i)}{\varepsilon^2 m^2} \le \frac{1}{\varepsilon^2 m}$$

since X_i are bounded $\{0,1\}$ random variables. Thus, we have

$$\mathbb{E}[|\mathcal{A}(S) - \mathcal{A}(S')|] \le \sqrt{\mathbb{E}[|\mathcal{A}(S) - \mathcal{A}(S')|^2]} = \sqrt{2\,\mathrm{Var}(\mathcal{A}(S))} \le \sqrt{2}\varepsilon^{-1}m^{-1/2}.$$

Setting $m = O(\rho^{-2}\varepsilon^{-2})$ suffices to obtain the desired bound. □

Next, we study closeness testing: given two unknown distribution $\mathbf{p}, \mathbf{q}$ over $[n]$, determine whether $\mathbf{p} = \mathbf{q}$ or $d_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \ge \varepsilon$.

Theorem A.2. *There is a strongly $(\varepsilon, 1/3, \rho)$-replicable algorithm for closeness testing with sample complexity $m = \Theta(n^{2/3}\rho^{-2/3}\varepsilon^{-4/3} + n^{1/2}\rho^{-1}\varepsilon^{-2} + \rho^{-2}\varepsilon^{-2})$.*

We require a variant of the Efron-Stein inequality applied to variables sampled without replacement. Note that the following result crucially relies on the fact that the aggregation function f is symmetric, an assumption that is not required in the standard Efron-Stein inequality.

Theorem A.3 (Efron-Stein without Replacement)**.** *Let T be a bag of $2m$ samples. Let $X_1, \dots, X_m$ be a random subsample from T of size m and $Z = f(X_1, \dots, X_m)$ for some symmetric f. For any $S = (X_1, \dots, X_m)$, denote $S^{A \mapsto B}$ the sample S with $A \in S$ replaced by $B \notin S$. Suppose $|f(S) - f(S^{a \mapsto b})| \le c$ for any S, A, B. Then,*

$$\mathrm{Var}(Z) \le m\,\mathbb{E}_{S, A \in S, B \notin S}\left[\left(f(S) - f\left(S^{A \mapsto B}\right)\right)^2\right].$$

Proof. We begin by following the proof of Theorem 7 of [46] for the standard Efron-Stein inequality, applying modifications where necessary. Define $V_i = \mathbb{E}[Z \mid X_1, \dots, X_i] - \mathbb{E}[Z \mid X_1, \dots, X_{i-1}]$ so that

$$V := \sum_{i=1}^m V_i = \mathbb{E}[Z \mid X_1, \dots, X_m] - \mathbb{E}[Z] = Z - \mathbb{E}[Z].$$

In particular, $\mathrm{Var}(V) = \mathrm{Var}(Z)$. Then, since V is a sum of martingale differences, we have

$$\mathrm{Var}(V) = \mathbb{E}\left[\left(\sum_i V_i\right)^2\right] = \mathbb{E}\left[\sum_i V_i^2\right] + 2\,\mathbb{E}\left[\sum_{i<j} V_i V_j\right] = \sum_i \mathbb{E}[V_i^2].$$

We note that up to this point, we have followed the proof of [46] exactly (no independence is required). In particular, we remark that for $i < j$,

$$\mathbb{E}[V_i V_j] = \mathbb{E}[\mathbb{E}[V_i V_j \mid X_1, \dots, X_i]] = \mathbb{E}[V_i\,\mathbb{E}[V_j \mid X_1, \dots, X_i]] = 0$$

where we use the tower rule in the first equality, pull out V_i since it is measurable w.r.t. $X_1, \dots, X_i$ in the second inequality, and observe that $\mathbb{E}[V_j \mid X_1, \dots, X_i] = 0$ in the final equality as a martingale difference.

Now, our goal is to bound $\mathbb{E}[V_i^2]$. Here, since X_i are no longer independent, we must make a different argument. Let $R_i = T \setminus \{X_1, \dots, X_{i-1}\}$. Then, for $x \in R_i$ define

$$h_i(x) := \mathbb{E}[Z \mid X_1, \dots, X_{i-1}, X_i = x]$$

so that $V_i = h_i(X_i) - \mathbb{E}[h(X_i)]$ given $X_1, \dots, X_{i-1}$. Then,

$$\mathbb{E}[V_i^2 \mid X_1, \dots, X_{i-1}] = \mathbb{E}[(h_i(X_i) - \mathbb{E}[h_i(X_i)])^2 \mid X_1, \dots, X_i] = \mathrm{Var}(h_i(X_i) \mid X_1, \dots, X_{i-1}).$$

In particular, it suffices to bound $\mathbb{E}[V_i^2]$ as follows. If X_i, X_i' are independent random draws from R_i, we have

$$\mathbb{E}[V_i^2 \mid X_1, \dots, X_{i-1}] = \frac{1}{2}\,\mathbb{E}_{X_i, X_{i'}}[(h_i(X_i) - h_i(X_i'))^2 \mid X_1, \dots, X_{i-1}]. \tag{1}$$

We discuss how to bound the right hand side. Next, for fixed $x \in R_i$, we may define S_x to be a random set of size m containing $\{X_1, \dots, X_{i-1}, x\}$, that is S_x is obtained by randomly sampling $m - i$ elements from $R_i \setminus \{x\}$. Our goal is now to bound $h_i(x) - h_i(y)$ for fixed $x \ne y \in R_i$. Consider the following coupling.

1. **Case 1:** $y \notin S_x$.
   Then, let $S_y = S_x \setminus \{x\} \cup \{y\}$. That is, $S_y = S_x^{x \mapsto y}$ is the set S_x with x swapped for y.
2. **Case 2:** $y \in S_x$. Then, let $S_y = S_x$.
   In particular, if $X_j = y$ for some $j > i$, set $X_i = y$ and $X_j = x$. We swap the indices of x and y but maintain the same final set.

Observe that this coupling still ensures that S_y is a random subset of size m. In particular, if we let Π denote this coupling, we have

$$\begin{aligned}(h_i(x) - h_i(y))^2 &= (\mathbb{E}_\Pi[f(S_x) \mid X_1, \ldots, X_{i-1}, X_i = x] - \mathbb{E}_\Pi[f(S_y) \mid X_1, \ldots, X_{i-1}, X_i = y])^2 \\ &= (\mathbb{E}_\Pi[f(S_x) - f(S_y) \mid X_1, \ldots, X_{i-1}])^2 \\ &\leq \mathbb{E}_\Pi[(f(S_x) - f(S_y))^2 \mid X_1, \ldots, X_{i-1}]\end{aligned}$$

where we conclude with Jensen's inequality. Since f is symmetric, we note that $f(S_x) = f(S_y)$ whenever $y \in S_x$, and the difference is $f(S_x) - f(S_x^{x \mapsto y})$ whenever $y \notin S_x$. Thus, if we sum over $y \notin S_x$, we have for any fixed x,

$$\sum_{y \in R_i \setminus \{x\}} (h_i(x) - h_i(y))^2 = \mathbb{E}_{S_x}\left[\sum_{b \notin S_x} \left(f(S_x) - f\left(S_x^{x \mapsto b}\right)\right)^2 \mid X_1, \ldots, X_{i-1}\right].$$

Summing over at most $2m - i + 1$ choices of x, we have

$$\sum_{x \neq y \in R_i} (h_i(x) - h_i(y))^2 \leq (2m - i + 1)\, \mathbb{E}_S\left[\sum_{b \notin S} \left(f(S) - f\left(S^{X_i \mapsto b}\right)\right)^2 \mid X_1, \ldots, X_{i-1}\right]$$

where S is an arbitrary subset of T of size m conditioned on $X_1, \ldots, X_{i-1}$. Plugging the above into Equation (1) and observing that X_i, X_i' are independently, randomly chosen from R_i, we have

$$\mathbb{E}[V_i^2 \mid X_1, \ldots, X_{i-1}] \leq \frac{1}{2(2m - i + 1)}\, \mathbb{E}_S\left[\sum_{b \notin S} \left(f(S) - f\left(S^{X_i \mapsto b}\right)\right)^2 \mid X_1, \ldots, X_{i-1}\right].$$

Taking expectation over $X_1, \ldots, X_{i-1}$ and applying the tower rule, we obtain

$$\mathbb{E}[V_i^2] \leq \frac{1}{2(2m - i + 1)}\, \mathbb{E}_S\left[\sum_{b \notin S} \left(f(S) - f\left(S^{X_i \mapsto b}\right)\right)^2\right].$$

Finally, we sum over i to bound $\mathrm{Var}(V)$ as

$$\mathrm{Var}(V) \leq \frac{1}{2} \sum_i \frac{1}{2m - i + 1}\, \mathbb{E}_S\left[\sum_{b \notin S} \left(f(S) - f\left(S^{X_i \mapsto b}\right)\right)^2\right].$$

Now, we again use the fact that f is symmetric to observe that $f(S^{X_i \to b})$ is identically distributed for all i: X_i is uniformly chosen from S while b is uniformly chosen from $T \setminus S$. In particular, we have

$$\mathbb{E}_S\left[\sum_{b \notin S} \left(f(S) - f\left(S^{X_i \mapsto b}\right)\right)^2\right] = m\, \mathbb{E}_{S, A \in S, B \notin S}\left[\left(f(S) - f\left(S^{A \mapsto B}\right)\right)^2\right]$$

so that

$$\begin{aligned}\mathrm{Var}(Z) &\leq \frac{m}{2}\, \mathbb{E}_{S, A \in S, B \notin S}\left[\left(f(S) - f\left(S^{A \mapsto B}\right)\right)^2\right] \sum_{i=1}^{m} \frac{1}{2m - i + 1} \\ &\leq m\, \mathbb{E}_{S, A \in S, B \notin S}\left[\left(f(S) - f\left(S^{A \mapsto B}\right)\right)^2\right].\end{aligned}$$

□

Given Theorem A.3, we prove Theorem A.2.

Proof of Theorem A.2. We will prove that the replicable closeness tester of [25] is in fact strongly replicable. We begin by recapping their algorithm below.

1. Fix $m = \Theta(n^{2/3}\rho^{-2/3}\varepsilon^{-4/3} + n^{1/2}\rho^{-1}\varepsilon^{-2} + \rho^{-2}\varepsilon^{-2})$ and draw $m_{\mathbf{p}}, m'_{\mathbf{p}}, m_{\mathbf{q}}, m'_{\mathbf{q}}$ from a multinomial distribution with parameters $(4m, (1/4, 1/4, 1/4, 1/4))$.
2. Draw two samples D_1, D_2 from $\mathbf{p}$ of size $m_{\mathbf{p}}, m'_{\mathbf{p}}$ and two samples D_3, D_4 from $\mathbf{q}$ of size $m_{\mathbf{q}}, m'_{\mathbf{q}}$. Let P_i, P'_i, Q_i, Q'_i be the frequency of i in D_1, D_2, D_3, D_4 respectively.
3. For all $j \in [n]$, compute $Z_j \leftarrow |P_j - Q_j| + |P'_j - Q'_j| - |P_j - P'_j| - |Q_j - Q'_j|$.
4. Compute $Z = \sum_{j=1}^{n} Z_j$.
5. Set a (randomized) threshold τ between R_0, R_1 (defined in Lemma 3.1 of [25]) and accept if $Z \le \tau$.

In our analysis, since correctness follows from [25], it suffices for us to argue strong replicability. Consider the algorithm that returns

$$\min\left(1, \max\left(0, \frac{Z - R_0}{R_1 - R_0}\right)\right).$$

To bound the variance, we observe that clipping cannot increase variance, so that it suffices to bound $\mathrm{Var}(Z)/(R_1 - R_0)^2$. From Lemma 3.1 of [25] we have

$$R_1 - R_0 = \Omega\left(\min\left(\varepsilon m, m^2\varepsilon^2 n^{-1}, m^{3/2}\varepsilon^2 n^{-1/2}\right)\right).$$

Now, consider the random variable Z when the samples D_1, D_2, D_3, D_4 are not drawn i.i.d. from $\mathbf{p}, \mathbf{q}$ but instead a random subset of size $4m$ from T, an arbitrary set of size $8m$. If we swap out any element in D_1, D_2, D_3, D_4, we observe that Z changes by at most 4, since a single sample appears in at most two terms. Thus, Theorem A.3 implies that $\mathrm{Var}(Z) \le 64m$. In particular, for two independent runs of the algorithm over subsets of T, with outputs Z_1, Z_2, we have

$$\mathbb{E}\left[\left|\frac{Z_1 - Z_2}{R_1 - R_0}\right|\right] \le \frac{\sqrt{2\,\mathrm{Var}(Z)}}{R_1 - R_0} = \frac{1}{\Omega\left(\min\left(\varepsilon m^{1/2}, m^{3/2}\varepsilon^2 n^{-1}, m\varepsilon^2 n^{-1/2}\right)\right)}.$$

Setting $m = \Theta(\rho^{-2}\varepsilon^{-2} + n^{2/3}\rho^{-2/3}\varepsilon^{-4/3} + n^{1/2}\rho^{-1}\varepsilon^{-2})$ suffices to upper bound the desired quantity by ρ. □

B Omitted Proofs of Section 3

We present the proof of Lemma 3.3.

Proof of Lemma 3.3. We prove the equivalence between replicability and ℓ_1-replicability and remark that the equivalence between strong replicability and strong ℓ_1-replicability hold identically. We will highlight the differences where necessary.

We first show the conversion from a ρ-replicable $\mathcal{P}$ tester $\mathcal{A}_1$ to a $(\varepsilon, \delta + \eta, \rho + 2\eta)$-$\ell_1$-replicable tester $\mathcal{A}'_1$. Let k be the same complexity of $\mathcal{A}_1$ and n the domain size. For $m = O(\eta^{-2}k \log n\eta^{-1})$, denote by $r := (r_1, \cdots, r_m)$ a vector of i.i.d. random seeds, each drawn uniformly at random from the set of possible seeds. Consider the algorithm $\mathcal{A}'_1$ that, upon receiving a bag T of samples, runs $\mathcal{A}_1$ independently m times using random seeds $r_1, \cdots, r_m$, respectively, and then outputs the average of the empirical outputs $\hat{f}(T; r) := \frac{1}{m}\sum_{i=1}^{m} \mathcal{A}_1(T; r_i)$. We claim that $\mathcal{A}'_1$ is $(\rho + 2\eta)$-ℓ_1-replicable and solves $(\varepsilon, \delta + \eta)$-$\mathcal{P}$ testing problem. Note that $\mathcal{A}'_1$ is efficient since it runs $\mathcal{A}_1$ m times and $m = \mathrm{poly}\,(1/\eta, k, \log n)$.

We first show correctness. Denote by $f(T) := \Pr_r\left[\mathcal{A}_1(T; r) = 1\right]$, then by a Chernoff bound, for any fixed bag T of samples and scalar $a = \eta/2$ we have that $\Pr_r\left[\left|\hat{f}(T; r) - f(T)\right| \ge a\right] \le 2e^{-2ma^2}$. Then, if we let n denote the underlying domain size, we can apply a union bound over all n^k possible bags of samples T to ensure that $2n^k e^{-2ma^2} < 0.1$ whenever $m = O(\eta^{-2}k \log n\eta^{-1})$ for some sufficiently large constant factor. In particular, there exists some r^* such that for every bag T that $|\hat{f}(T; r^*) - f(T)| \le a$. Thus, we fix this r^* and define $\hat{f}(T) := \hat{f}(T, r^*)$. Our algorithm $\mathcal{A}'_1$ will output $\hat{f}(T)$ (note that now $\mathcal{A}'_1$ is a deterministic algorithm). Thus, we bound:

$$\left|\mathbb{E}_T\left[f(T) - \hat{f}(T)\right]\right| \le \mathbb{E}_T\left[\left|f(T) - \hat{f}(T)\right|\right] \le a = \eta/2, \tag{2}$$

where the first equality follows from Jensen's inequality, the second by our observation that $|\hat{f}(T) - f(T)| \leq a$ for all T. In the completeness case, we have that

$$\mathbb{E}_T[\mathcal{A}_1'(T)] = \mathbb{E}_T\left[\hat{f}(T)\right] \geq \mathbb{E}_T[f(T)] - \left|\mathbb{E}_T\left[f(T) - \hat{f}(T)\right]\right| \geq 1 - \delta - \eta,$$

where the first inequality follows from the reverse triangle inequality, the second inequality follows from Equation (2) and the correctness guarantee of $\mathcal{A}_1$. The correctness of $\mathcal{A}_1'$ in the soundness case can be shown using the same argument.

We next show ℓ_1-replicability (resp. strong ℓ_1-replicability). To show an upper bound on $\mathbb{E}_{S,S'}[|\hat{f}(S) - \hat{f}(S')|]$, we first notice that

$$\begin{aligned}
&\mathbb{E}_{S,S'}\left[\left|\hat{f}(S) - \hat{f}(S')\right|\right] \leq \mathbb{E}_{S,S'}\left[\left|\hat{f}(S) - f(S) + f(S) - \hat{f}(S') - f(S') + f(S')\right|\right] \\
&\leq \mathbb{E}_S\left[\left|\hat{f}(S) - f(S)\right|\right] + \mathbb{E}_{S'}\left[\left|\hat{f}(S') - f(S')\right|\right] + \mathbb{E}_{S,S'}\left[|f(S) - f(S')|\right],
\end{aligned}$$

where the second inequality follows from triangle inequality. Notice that we can bound the first two terms using Equation (2) by η, and the last term can be upper bounded by ρ as follows,

$$\begin{aligned}
&\mathbb{E}_{S,S'}\left[|f(S) - f(S')|\right] = \mathbb{E}_{S,S'}\left[\left|\Pr_r[\mathcal{A}_1(S;r) = 1] - \Pr_r[\mathcal{A}_1(S';r) = 1]\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\left|\Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 1, \mathcal{A}_1(S';r) = 1] + \Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 1, \mathcal{A}_1(S';r) = 0]\right.\right. \\
&\qquad\qquad \left.\left. - \Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 1, \mathcal{A}_1(S';r) = 1] - \Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 0, \mathcal{A}_1(S';r) = 1]\right|\right] \\
&\leq \mathbb{E}_{S,S'}\left[\Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 1, \mathcal{A}_1(S';r) = 0] + \Pr_{r,S,S'}[\mathcal{A}_1(S;r) = 0, \mathcal{A}_1(S';r) = 1]\right] \\
&= \Pr_{r,S,S'}\left[\mathcal{A}_1(S;r) \neq \mathcal{A}_1(S';r)\right] < \rho,
\end{aligned}$$

where the second equality follows from law of total probability and the fact that $\mathcal{A}_1$ only returns 0 or 1. Note that we only use that $\mathcal{A}_1$ is replicable in the final equality. If $\mathcal{A}_1$ is strongly replicable, then an identical argument holds where S, S' are instead sampled from an arbitrary set of size $2m$. Hence, we have that $\mathbb{E}\left[\left|\hat{f}(S) - \hat{f}(S')\right|\right] = \rho + 2\eta$, which concludes the proof of ℓ_1-replicability of $\mathcal{A}_1'$.

On the other hand, we proceed to show how to obtain a ρ-replicable tester $\mathcal{A}_2'$ given a ρ-ℓ_1-replicable $\mathcal{P}$ tester $\mathcal{A}_2$. Consider the algorithm $\mathcal{A}_2'$ that upon receiving input T, runs $\mathcal{A}_2$ on T, and compare the result against the threshold τ that is uniformly at randomly selected from $(0, 1)$. If $\mathcal{A}_2(T) \leq \tau$, $\mathcal{A}_2'$ rejects, otherwise accepts. We claim that $\mathcal{A}_2'$ is a ρ-replicable $\mathcal{P}$ tester, and note that $\mathcal{A}_2'$ is efficient since it only requires running $\mathcal{A}_2$ once.

We first show correctness. Indeed, in the completeness case we have that

$$\begin{aligned}
\Pr_{T,\tau}[\mathcal{A}_2'(T;\tau) = 0] &= \Pr_{T,\tau}\left[\mathcal{A}_2(T) \leq \tau\right] \\
&= \mathbb{E}_T\left[\Pr_\tau\left[\mathcal{A}_2(T) \leq \tau \mid T\right]\right] = \mathbb{E}_T\left[1 - \mathcal{A}_2(T)\right] \\
&= 1 - \mathbb{E}_T\left[\mathcal{A}_2(T)\right] < \delta,
\end{aligned}$$

where the first equality follows from law of total expectation, the second equality follows from the fact that the probability of a uniformly distributed random variable between $[0, 1]$ is greater than a constant in $[0, 1]$ is simply one minus that constant, and the third equality follows from the correctness guarantee of $\mathcal{A}_2$. Similarly, we have that in the soundness case

$$\Pr_{T,\tau}[\mathcal{A}_2'(T) = 1] = \Pr_{T,\tau}[\mathcal{A}_2(T) > \tau] < \delta,$$

which concludes the proof of correctness of $\mathcal{A}_2'$. We next show ρ-replicability of $\mathcal{A}_2'$. Indeed, $\mathcal{A}_2'$ producing inconsistent outputs when run on two different sample sets from the same distribution happens exactly when threshold τ falls between the result of $\mathcal{A}_2$ in two different runs. Denote by $X_1 := \min\{\mathcal{A}_2(T), \mathcal{A}_2(T')\}$ and $X_2 := \max\{\mathcal{A}_2(T), \mathcal{A}_2(T')\}$ the random variables with

randomness in T, T', then we have that

$$\Pr_{T,T',\tau}\left[\mathcal{A}_2'(T) \neq \mathcal{A}_2'(T')\right] = \Pr_{T,T',\tau}\left[\tau \in [X_1, X_2)\right] = \mathbb{E}_{T,T'}\left[\Pr_{\tau}\left[\tau \in (X_1, X_2) | \, T, T'\right]\right]$$
$$= \mathbb{E}_{T,T'}\left[|\mathcal{A}_2(T) - \mathcal{A}_2(T')|\right] < \rho,$$

where the second equality follows from the law of total expectation and the fact that the point mass of a continuous distribution is 0, which concludes the proof. Again, we only use ℓ_1-replicability of $\mathcal{A}_2$ in the final inequality, and observe that an identical argument holds for strong ℓ_1-replicable testers. □

We present the proof of Lemma 3.4.

Proof of Lemma 3.4. We must bridge the gap between ℓ_1-replicability and strong-replicability. For random sample bags $S, S' \sim_{i.i.d.} \mathbf{p}^{\otimes m}$ where $\mathbf{p}$ is the input distribution, denote by $T = S \uplus S'$ the combined sample pool. At a high level, we use the law of total expectation to condition on a fixed realization of T. Under this conditioning, S or S' are distributed as uniform random sub-bags of T. By introducing a third auxiliary sub-bag S'' that is also an uniformly chosen from T of size m, and applying triangle inequality, we can use the strong-replicability to obtain the desired result.

More concretely, for random sample bags $S, S' \sim_{i.i.d.} \mathbf{p}^{\otimes m}$ where $\mathbf{p}$ is the input distribution, denote by $T = S \uplus S'$ the random variable of the combined sample pool, and let $S'' \sim_u \binom{T}{m}$ be an independent subset of T of size m, then, by triangle inequality and linearity of expectation we have that

$$\mathbb{E}\left[|\mathcal{A}(S) - \mathcal{A}(S')|\right] \leq \mathbb{E}\left[|\mathcal{A}(S) - \mathcal{A}(S'')|\right] + \mathbb{E}\left[|\mathcal{A}(S') - \mathcal{A}(S'')|\right].$$

It suffices to upper-bound $\mathbb{E}\left[|\mathcal{A}(S) - \mathcal{A}(S'')|\right]$, and $\mathbb{E}\left[|\mathcal{A}(S') - \mathcal{A}(S'')|\right]$ can be bounded using the same argument by symmetry between S and S'. Indeed,

$$\mathbb{E}_{S\sim_{i.i.d.}\mathbf{p}^{\otimes m};T;S''\sim_u\binom{T}{m}}\left[|\mathcal{A}(S) - \mathcal{A}(S'')|\right] = \mathbb{E}_T\left[\mathbb{E}_{S;S''}\left[|\mathcal{A}(S;r) - \mathcal{A}(S'')||T\right]\right]$$
$$= \mathbb{E}_T\left[\mathbb{E}_{S,S''\sim_u\binom{T}{m}}\left[|\mathcal{A}(S) - \mathcal{A}(S')|\right]\right] < \rho,$$

where the first equality follows from the law of total expectation, and the second equality follows from the fact that the sequence of random variables in T is $i.i.d.$ and hence, by symmetry, any subset of variables from T, in particular S, is distributed identically to a uniform random subset of T, which concludes the proof. □

C Omitted Proofs of Section 4

In this section, we prove Proposition 4.3. We will require Maclaurin's inequality.

Theorem C.1 (Maclaurin's Inequality)**.** *Let $a_1, \ldots, a_n \geq 0$. For $k \geq 1$, define*

$$S_k = \frac{\sum_{i_1 < \cdots < i_k} \prod_{j=1}^{k} a_{i_j}}{\binom{n}{k}}.$$

Then, $S_1 \geq S_2^{1/2} \geq \cdots \geq S_n^{1/n}$ with equality if and only if $a_1 = \cdots = a_n$.

We now prove Proposition 4.3.

Proof of Proposition 4.3. Consider $\mathcal{P}$ to be the set of all distributions. Let $Y \subset \mathcal{X}$ be a subset whose size is $C\log(1/\rho)$ for an appropriately large constant C. We design $\mathcal{A}$ to be the following algorithm with sample complexity $m = |Y|$. Given samples S, accept unless both of the following are satisfied:

1. No element of Y appears more than once in S.
2. At least a third of the elements of Y appear in S.

First, we claim that $\mathcal{A}$ is $(\varepsilon, 1/3, \rho)$-replicable, i.e. for any distribution $\mathbf{p}$, we have $\Pr(\mathcal{A}(S) = 0) < \rho$. To see this, consider two cases. We bound the probability that the algorithm rejects. If our sample does not contain $|Y|/3$ elements from Y, then clearly the second condition is not met, and the algorithm accepts. Conditioned on the event that the algorithm receives at least $m := |Y|/3$ samples in Y, our goal is to upper bound the probability that no collision occurs. Note that it suffices to condition on the event that the algorithm receives exactly m samples in Y, since the probability of no collision occurring decreases with more samples in Y. In particular, since samples in Y are drawn i.i.d. from $\mathbf{p}_{|Y}$ (the distribution $\mathbf{p}$ conditioned on Y), the probability of no collision decreases as the number of samples in Y increase. Formally, the above discussion can be summarized as:

$$\begin{aligned}\Pr(\mathcal{A} \text{ rejects}) &= \Pr(\text{no collision in } Y \text{ and at least } |Y|/3 \text{ samples in } Y) \\ &\leq \Pr(\text{no collision in } Y \mid \text{at least } |Y|/3 \text{ samples in } Y) \\ &\leq \Pr(\text{no collision in } Y \mid \text{exactly } |Y|/3 \text{ samples in } Y).\end{aligned}$$

To bound this quantity, we observe that the probability that there are no collisions among the m samples can be expressed as

$$m! \sum_{i_1 < \cdots < i_m} \prod_{j=1}^{m} \mathbf{p}(y_{i_j}) \leq \frac{|Y|!}{|Y|^m (|Y| - m)!} = \prod_{j=1}^{m} \frac{|Y| - j + 1}{|Y|}.$$

The first expression is the probability of drawing m distinct samples, while the inequality follows from Maclaurin's inequality (Theorem C.1) where we use $S_m \leq S_1^m \leq (1/|Y|)^m$. To upper bound this final quantity, we have

$$\prod_{j=1}^{m} \frac{|Y| - j + 1}{|Y|} = \prod_{j=1}^{m} \left(1 - \frac{j-1}{|Y|}\right) \leq \exp\left(-\sum_{j=1}^{m-1} \frac{j}{|Y|}\right) = \exp\left(-m^2/3|Y|\right).$$

Since we set $m \geq |Y|/3$, we have that the above quantity is $\exp(-\Omega(|Y|)) < \rho$ as desired.

Next, we show that $\mathcal{A}$ is not $(\varepsilon, 1/3, \rho)$-strongly replicable (regardless of sample complexity). In particular, choose T that contains exactly a $2/3$-fraction of the elements in Y, each appearing once, and the remaining elements outside of Y. Clearly, for any subset $S \subset T$, no element of T can appear more than once. Furthermore, the number of elements of Y that appear in S is less than $|Y|/3$ with probability $\frac{1}{2}$, so that the algorithm accepts with probability $\frac{1}{2}$. Thus, $\mathbb{E}[|\mathcal{A}(S) - \mathcal{A}(S')|] = \Omega(\mathrm{Var}(\mathcal{A}(S))) = \Omega(1)$.

Finally, we show that $\mathcal{A}$ is not a high probability tester with error probability $\exp(-\tilde{\Omega}(\rho^{-2}))$. Consider the distribution that assigns probability $1/|Y|$ to each element of Y. Then, the probability that we produce a sample of size $m = |Y|/3$ containing each element of Y once is exactly (using the equality case of Maclaurin's inequality)

$$\frac{m!}{|Y|^m} = \prod_{j=1}^{m-1} \left(1 - \frac{j}{|Y|}\right) \geq \exp\left(-\Omega\left(\frac{m^2}{|Y|}\right)\right) = \mathrm{poly}\,(\rho).$$

□

D High Probability Testers may not be Replicable

In this section, we exhibit a simple property (identity testing) where high probability testing is easier than replicability.

Proposition 1.6. *There exists a property $\mathcal{P}$ (identity testing) for which there is an $(1, 0)$-tester with 1-sample but any $(1, \rho)$-replicable tester requires $\Omega(\rho^{-2})$ samples.*

Proof of Proposition 1.6. Let $\mathcal{X} = \{0, 1\}$. Consider the property $\mathcal{P}$ which consists of the distribution $\mathbf{p}$ supported entirely on 0. Note that the unique distribution that is 1-far from $\mathcal{P}$ in total variation distance is the distribution $\mathbf{q}$ supported entirely on 1. Then, it is easy to see that $\mathcal{A}$ which takes 1 sample and returns $\mathbf{p}$ if the sample is 0 and $\mathbf{q}$ if the sample is 1 is a $(1, 0)$-tester for $\mathcal{P}$.

On the other hand, any $(1, \rho)$-replicable tester for $\mathcal{P}$ requires $\Omega(\rho^{-2})$ samples, which follows from Theorem 3.12 of [38], which we restate here for completeness.

Theorem D.1 (Theorem 3.12 of [38]). *Any ρ-replicable algorithm for $\{\pm 1\}$-bias estimation requires $\Omega(\rho^{-2})$ samples.*

The $\{\pm 1\}$-bias estimation problem asks an algorithm to distinguish a coin with bias 1 from a coin with bias 0 with error at most $1/3$, which is exactly the problem of $(1, 1/3)$-testing for $\mathcal{P}$. Thus, any ρ-replicable algorithm that is a 1-tester for $\mathcal{P}$ is also a ρ-replicable algorithm for $\{\pm 1\}$-bias estimation. $\square$

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction concisely demonstrate our contributions and scope.
   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: Our reduction from high probability testing requires strong replicability instead of replicability; we discuss this in our paper. Furthermore, we exhibit some barriers to obtaining a reduction in the reverse direction.
   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: All assumptions were clearly provided in the statements and each statement has a corresponding proof in either the main body or appendix.
   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [N/A]
   Justification: The paper does not include experiments.
   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [N/A]
   Justification: The paper does not include experiments requiring code.
   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [N/A]
   Justification: The paper does not include experiments.
   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [N/A]
   Justification: The paper does not include experiments.
   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: [N/A]
Justification: The paper does not include experiments.
Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
Answer: [Yes]
Justification: Yes, the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.
Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [N/A]
Justification: The work is theoretical and we do not see any major or immediate implications.
Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
Answer: [N/A]
Justification: The paper poses no such risks.
Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [N/A]
Justification: the paper does not use existing assets.
Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [N/A]
Justification: The paper does not release new assets.
Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [N/A]
Justification: The paper does not involve crowdsourcing nor research with human subjects.
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
Answer: [N/A]
Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.
Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.