Robust hypothesis testing and distribution estimation in Hellinger distance

Ananda Theertha Suresh
Google Research, New York

Abstract

We propose a simple robust hypothesis test that has the same sample complexity as that of the optimal Neyman-Pearson test up to constants, but robust to distribution perturbations under Hellinger distance. We discuss the applicability of such a robust test for estimating distributions in Hellinger distance. We empirically demonstrate the power of the test on canonical distributions.

1 Introduction

1.1 Simple hypothesis testing

Hypothesis testing and estimating unknown underlying distributions from samples are fundamental problems in statistics and learning theory respectively. The simplest hypothesis testing scenario is the following. Given two known distributions P and Q over a domain $\mathcal{X}$ and a set of n independent samples $X^n \triangleq X_1, X_2, \ldots, X_n$ generated from an unknown distribution $R \in \{P, Q\}$, simple hypothesis test asks which of the following two hypotheses is true:

$$\mathcal{H}_0 : R = P$$
$$\mathcal{H}_1 : R = Q.$$

The best known hypothesis test is the Neyman-Pearson test, which outputs $\mathcal{H}_0$ if

$$\frac{P(X^n)}{Q(X^n)} \geq t,$$

otherwise outputs $\mathcal{H}_1$ for a suitable threshold t (Neyman and Pearson, 1933; Cover and Thomas, 2012). There are two types of errors associated with hypothesis testing: type I error and type II error. Type I error is the probability that the test outputs $\mathcal{H}_1$ if $\mathcal{H}_0$ is true and type II error is the probability that the test outputs $\mathcal{H}_0$ if $\mathcal{H}_1$ is true. The Neyman-Pearson test achieves the best type II error for a given bound on the type I error.

For simplicity, let the error probability of a hypothesis test be the maximum of type I and type II errors. For a test T and distributions P and Q, let $N^T_\delta(P, Q)$ be the number of samples necessary to achieve error probability δ. Let the optimal sample complexity $N^*_\delta(P, Q)$ be the minimum number of samples necessary to achieve an error probability δ:

$$N^*_\delta(P, Q) = \min_T N^T_\delta(P, Q).$$

We need few definitions to state the optimal sample complexity. For a function $U : \mathcal{X} \to \mathbb{R}^+$, the p-norm of U is given by

$$\|U\|_p = \left(\int_{x \in \mathcal{X}} |U(x)|^p dx \right)^{1/p}.$$

For two distributions P and Q over $\mathcal{X}$[1], the Hellinger distance between P and Q is given by

$$\mathrm{H}(P, Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2.$$

The sample complexity of the optimal hypothesis test between P and Q is (Bar-Yossef and Papadimitriou, 2002; Canonne et al., 2019)

$$N^*_\delta(P, Q) = \Theta\left(\frac{\log(1/\delta)}{\mathrm{H}^2(P, Q)} \right). \tag{1}$$

1.2 Robust hypothesis testing

In many natural scenarios, the underlying distribution may not be either of P and Q, but close to one of them. This can happen due to a several reasons such as noisy samples, modelling error, or lack of expressivity in the class of distributions under consideration. For example, suppose we have the following two hypotheses:

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] We state the results for continuous distributions and the exact results hold for discrete distributions.

- $\mathcal{H}_0$: the number of submissions to a conference every year is Poi(5000), a Poisson distribution with mean 5000.
- $\mathcal{H}_1$: the number of submissions to the conference every year is Poi(6000).

It is plausible that in reality, the number of submissions every year is a Poisson mixture $(1-\epsilon)\cdot \text{Poi}(5000)+\epsilon\cdot \text{Poi}(10000)$ for a small ϵ. In this scenario, it is desirable for the hypothesis test to overcome the modelling error and output $\mathcal{H}_0$. It is also preferable for the proposed test to have the same sample complexity as the optimal simple hypothesis test. In this paper, we ask the following question:

Is there a test with the same sample complexity as that of the Neyman-Pearson test and is robust to a broad class of distribution perturbations?

We answer this question affirmatively. To define the broad class of distribution perturbations, we need a measure of closeness between distributions. Since Hellinger distance naturally characterizes the sample complexity of optimal hypothesis testing, we ask if there are robust hypothesis tests under the Hellinger distance.

Given two known distributions P and Q over a domain $\mathcal{X}$, and a set of n independent samples X^n generated from some distribution R, one can ask which of the following two hypotheses is true:

$$\begin{aligned}\mathcal{H}_0 &: \mathrm{H}(P,R) < \mathrm{H}(Q,R)\\ \mathcal{H}_1 &: \mathrm{H}(P,R) > \mathrm{H}(Q,R).\end{aligned}$$

For distributions such that $|\mathrm{H}(P,R)-\mathrm{H}(Q,R)|$ is arbitrarily small, differentiating between the two hypotheses with finitely many samples would not be possible. Hence we propose γ-robust hypothesis testing as follows: given two known distributions P and Q over a domain $\mathcal{X}$, and a set of n independent samples X^n generated from some distribution R, we ask which of the following two hypotheses is true:

$$\begin{aligned}\mathcal{H}_0 &: \gamma\cdot\mathrm{H}(P,R) \le \mathrm{H}(Q,R)\\ \mathcal{H}_1 &: \mathrm{H}(P,R) \ge \gamma\cdot\mathrm{H}(Q,R),\end{aligned}$$

for $\gamma > 1$, where γ is the slackness term. If neither of the hypotheses is true, then the test can output either of the hypotheses. As before, we define the error of the test as the maximum of type I and type II errors. For a test T, let $N^T_\delta(P,Q,\gamma)$, be the number of samples necessary to achieve error probability δ for γ-robust hypothesis testing and let $N^*_\delta(P,Q,\gamma)$ be the optimal sample complexity of the γ-robust hypothesis testing:

$$N^*_\delta(P,Q,\gamma) = \min_T N^T_\delta(P,Q,\gamma).$$

If a test cannot achieve error probability less than 1/2 asymptotically, then we say such a test is not γ-robust. A natural question is to ask if the Neyman-Pearson test is distributionally robust for some γ. We show that the Neyman-Pearson test is not robust for any $\gamma > 1$ by constructing P, Q and a set of distributions R_m such that $\lim_{m\to\infty}\mathrm{H}(P,R_m)=0$, but the Neyman-Pearson test outputs $\mathcal{H}_1$ with high probability. We provide the proof in Section 5.1.

Lemma 1. *There exists two distributions P and Q such that $N^*_{1/3}(P,Q)=\Theta(1)$ and the Neyman-Pearson test is not robust for any $\gamma>1$.*

1.3 Related works

We overview robust hypothesis tests with different measures. Let $\mathcal{T}$ denote the class of all hypothesis tests. For a pair of distributions P, Q, and test $T\in\mathcal{T}$, let $P_e(T,P,Q)$ be the maximum of type I and type II errors of T for distributions P and Q. The problem of finding optimal robust hypothesis test can be formulated as

$$\min_{T\in\mathcal{T}}\max_{P'\in C(P),Q'\in C(Q)} P_e(T,P',Q'),$$

for some convex sets $C(P)$ and $C(Q)$. For tests with n samples, $P_e(T,P',Q')$ is convex in both product spaces $(P')^n$ and $(Q')^n$ over X^n. Hence the above min-max problem is convex and the optimal test T can be obtained by computing the least favorable distributions. However this approach can be computationally inefficient.

The first closed form estimator is due to Huber (1965). They considered a Kolomogorov distance type metric and showed that a clipped log-likelihood test is optimal. Levy (2008); Gül and Zoubir (2017) studied robust distribution hypothesis testing with KL divergence, given by

$$\mathrm{KL}(P,R)=\int_x P(x)\log\frac{P(x)}{R(x)}dx.$$

However, KL divergence is not symmetric and simple examples such as the one in Lemma 1 do not have small KL divergence to the underlying true distributions. Gao et al. (2018) studied robust hypothesis tests in the Wasserstein metric. Gül and Zoubir (2014) designed robust hypothesis testing under Hellinger distance when $\mathcal{X}=\mathbb{R}$ and derived least favorable distributions and a decision rule. Dong (1999) considered the related problem of goodness of fit tests based on Hellinger distance.

Scheffé (1947) proposed robust hypothesis test in total variation distance, given by

$$\mathrm{TV}(P,Q)=\frac{1}{2}\left\|P-Q\right\|_1.$$

For any two distributions P and Q, Scheffe estimator uses $\mathcal{O}\left(\frac{\log(1/\delta)}{\mathrm{TV}^2(P,Q)}\right)$ samples to obtain an error probability of at most δ. It is easy to show that

$$\frac{1}{2}\mathrm{TV}^2(P,Q) \leq \mathrm{H}^2(P,Q) \leq \mathrm{TV}(P,Q). \tag{2}$$

We provide a simple proof of (2) in Section 6.1. By (2), the sample complexity of the Scheffe estimator can be worse than the sample complexity of the Neyman-Pearson test. Furthermore, if the upper bound in (2) is tight, then the sample complexity of the Scheffe estimator can be much higher than the optimal sample complexity as stated in the next lemma.

Lemma 2. *Let $N^S_\delta(P,Q)$ denote the sample complexity of the Scheffe test for simple hypothesis testing. For any $K > 1$, there exists distributions P, Q such that*

$$N^S_\delta(P,Q) = \Omega\left(K \cdot N^*_\delta(P,Q)\right).$$

We relegate the proof of Lemma 2 in Section 5.2.

After the initial version of this paper, authors were made aware of (Baraud, 2011). Baraud (2011) studies the problem of estimator selection with respect to Hellinger distance. They propose an estimator based on variational formula for Hellinger distance. We note that their estimator also has similar guarantees as our proposed estimator for robust hypothesis testing under Hellinger distance.

2 Contributions

2.1 Test statistic

We propose a test statistic that is distributionally robust for $\gamma > \frac{\sqrt{2}}{\sqrt{2}-1}$ and further has same sample complexity as that of the Neyman-Pearson test up to multiplicative factors. Since our goal is to come up with a test whose performance guarantee is independent of the underlying domain, it is desirable to have a test of the form $\sum_{i=1}^n f(X_i)$.

Hellinger distance involves a square-root term in its definition and hence finding a f directly is difficult. Hence, we approximate the Hellinger distance by the symmetric chi-squared statistic, given by

$$\chi^2(P,Q) = \left\| \frac{P-Q}{\sqrt{P+Q}} \right\|_2^2 = \int_x \frac{(P(x)-Q(x))^2}{P(x)+Q(x)} dx.$$

The symmetric chi-squared statistic approximates the square of the Hellinger distance to a multiplicative factor of two:

$$\frac{1}{4}\chi^2(P,Q) \leq \mathrm{H}^2(P,Q) \leq \frac{1}{2}\chi^2(P,Q). \tag{3}$$

We provide a derivation of (3) in Section 6.2. The symmetric chi-square statistic can be written as

$$\begin{aligned} &\chi^2(P,Q) \\ &= \int_x \frac{(P(x)-Q(x))^2}{P(x)+Q(x)} dx \\ &= \mathbb{E}_{X\sim P}\left[\frac{P(X)-Q(X)}{P(X)+Q(X)}\right] - \mathbb{E}_{X\sim Q}\left[\frac{P(X)-Q(X)}{P(X)+Q(X)}\right]. \end{aligned} \tag{4}$$

(4) motivates the following test statistic. Given n samples X^n from an unknown distribution R, let

$$\mathrm{T}(P,Q,X^n) = \frac{1}{n}\sum_{i=1}^n \frac{P(X_i)-Q(X_i)}{P(X_i)+Q(X_i)}.$$

By (4),

$$\mathbb{E}_{X^n\sim P}[\mathrm{T}(P,Q,X^n)] - \mathbb{E}_{X^n\sim Q}[\mathrm{T}(P,Q,X^n)] = \chi^2(P,Q).$$

Furthermore,

$$\mathbb{E}_{X^n\sim P}[\mathrm{T}(P,Q,X^n)] + \mathbb{E}_{X^n\sim Q}[\mathrm{T}(P,Q,X^n)] = 0.$$

Hence,

$$\begin{aligned} \mathbb{E}_{X^n\sim P}[\mathrm{T}(P,Q,X^n)] &= -\mathbb{E}_{X^n\sim Q}[\mathrm{T}(P,Q,X^n)] \\ &= \frac{1}{2}\chi^2(P,Q). \end{aligned}$$

Hence, a natural test is to output P if $\mathrm{T}(P,Q,X^n) > 0$ and Q if $\mathrm{T}(P,Q,X^n) < 0$, while breaking ties randomly. We refer to this test as HellingerTest. HellingerTest does not have any tunable hyper-parameters and just depends on the underlying distributions P and Q. Instead of comparing to zero, one can compare to a threshold t to get precise trade offs between type I and type II errors.

2.2 Theoretical guarantees

We show that HellingerTest is also robust to distribution perturbations in Hellinger distance and has the optimal sample complexity of simple hypothesis testing. Thus HellingerTest guarantees robustness in Hellinger distance for free.

Theorem 1. *Let $N^H_\delta(P,Q,\gamma)$ be the sample complexity of the* HellingerTest. *For $\gamma > \frac{\sqrt{2}}{\sqrt{2}-1}$ and any δ,*

$$N^H_\delta(P,Q,\gamma) = \frac{c_\gamma \log(1/\delta)}{H^2(P,Q)} = \Theta\left(N^*_\delta(P,Q)\right),$$

where c_γ is a constant that depends on γ.

We also show a lower-bound on the performance of the HellingerTest.

Theorem 2. *For every $\gamma < \frac{1}{\sqrt{2}-1}$, there exists distributions P, Q, and R such that $\frac{H(Q,R)}{H(P,R)} \geq \gamma$ and*

$$\mathbb{E}_{X^n\sim R}[T(P,Q,X^n)] = 0.$$

2.3 Implications for distribution estimation

Distribution robust hypothesis testing can be rewritten as a test that given pair of distributions P, Q and samples from an unknown distribution R, finds a distribution $\hat{R} \in \{P, Q\}$ such that $\mathrm{H}(R, \hat{R}) \leq \gamma \cdot \min(\mathrm{H}(P, R), \mathrm{H}(Q, R))$.

Such a distribution robust hypothesis testing can be used as a subroutine for learning distributions. Consider the following learning problem: given n samples $X^n \sim P \in \mathcal{P}$, find an estimate $\hat{P}$ such that $\mathrm{H}(P, \hat{P}) \leq \epsilon$.

A natural algorithm is to obtain an ϵ-cover of $\mathcal{P}$, denoted by $\mathcal{P}_\epsilon$ and run the robust hypothesis test between every pair of distributions and output the distribution that wins in the maximum number of tests (Devroye and Lugosi, 2012). It can be shown that the overall algorithm selects a distribution that is at most $c \cdot \gamma \cdot \epsilon$ away from the true distribution, where c is a constant. We refer readers to (Devroye and Lugosi, 2012, Section 6.8) for a detailed description of this algorithm. The above algorithm can be further modified with improved run time (Acharya et al., 2014, 2018) and to provide differential privacy (Bun et al., 2019).

Perhaps the most popular estimator is the Scheffe test, which studied the problem under the total variation distance (Scheffé, 1947; Yatracos, 1985). Scheffe test has been used in variety of works including learning Gaussian mixtures (Daskalakis and Kamath, 2014; Suresh et al., 2014; Ashtiani et al., 2018), k-modal distributions (Daskalakis et al., 2012), log-concave distributions (Diakonikolas et al., 2017), and piece-wise polynimial distributions (Chan et al., 2014).

The proposed test can be used in place of the Scheffe's estimator in the above papers to obtain learning guarantees in the Hellinger distance.

2.4 Modifications for differential privacy

Differential privacy has become the standardized notion of privacy in statistics. We refer readers to (Dwork et al., 2014) for details on differential privacy. Optimal hypothesis test with differential privacy was proposed by Canonne et al. (2019). Let

$$\Delta(P, Q) = \max_{x \in \mathcal{X}} \frac{|P(x) - Q(x)|}{P(x) + Q(x)} \leq 1.$$

Changing one sample changes the proposed test statistic $T(P, Q, X^n)$ by at most $2\Delta(P, Q)/n$, hence HELLINGERTEST can be modified to a ϵ-differentially private test by adding Laplace noise,

$$T_\epsilon(P, Q, X^n) = T(P, Q, X^n) + \frac{Z}{n},$$

where Z is a Laplace random variable with parameter $2\Delta(P, Q)/\epsilon$. While this algorithm is not optimal in general, as we show below, it is optimal for $\epsilon > 1$ and further has the advantage that it is parameter-free and simple to use.

Corollary 1. *$T_\epsilon(P, Q, X^n)$ is an ϵ-DP algorithm. Furthermore, its sample complexity is optimal and is same as that of the non-private complexity up to constants for*

$$\epsilon \geq \max_{x \in \mathcal{X}} \frac{|P(x) - Q(x)|}{P(x) + Q(x)}.$$

If $\Delta(P, Q)$ is unknown, instead of adding Laplace noise with parameter $2\Delta(P, Q)/n$, one can add Laplace noise with parameter $2/n$ and it would be near-optimal for $\epsilon \geq 1$ for all P, Q.

We relegate the proof to Section 5.4. The above algorithm is amenable to the same clipping strategy proposed by Canonne et al. (2019) and can be modified to obtain the optimal sample complexity with differential privacy.

2.5 Implications for other measures

If $\mathrm{H}(P, R) \leq \frac{1}{\gamma+1} \mathrm{H}(P, Q)$, then by the triangle inequality,

$$\begin{aligned} \mathrm{H}(Q, R) &\geq \mathrm{H}(P, Q) - \mathrm{H}(P, R) \\ &\geq (\gamma + 1 - 1)\mathrm{H}(P, R) \\ &= \gamma \mathrm{H}(P, R), \end{aligned}$$

Similarly, if $\mathrm{H}(Q, R) \leq \frac{1}{\gamma+1} \mathrm{H}(P, Q)$, then $\mathrm{H}(P, R) \geq \gamma \mathrm{H}(Q, R)$. Hence, if there is γ-robust hypothesis test, it can also differentiate between the following two hypotheses:

$$\mathcal{H}_0 : \mathrm{H}(P, R) \leq \frac{1}{\gamma + 1} \mathrm{H}(P, Q)$$
$$\mathcal{H}_1 : \mathrm{H}(Q, R) \leq \frac{1}{\gamma + 1} \mathrm{H}(P, Q),$$

such that the sample complexity is same as that of the Neyman-Pearson test. Furthermore if there is a measure d such that Hellinger distance is upper bounded by some function of d, then the test works for even that class of distributions. This observation yields the following corollary.

Corollary 2. *Let $\beta < \frac{\sqrt{2}-1}{2\sqrt{2}-1}$.* HELLINGERTEST *has the same complexity as the optimal simple hypothesis testing for the following composite hypothesis testing scenarios:*

1. *Hellinger distance:*

$$\mathcal{H}_0 : H(P, R) \leq \beta H(P, Q)$$
$$\mathcal{H}_1 : H(Q, R) \leq \beta H(P, Q).$$

2. *Total variation distance:*

$$\mathcal{H}_0 : TV(P,R) \le \beta^2 H^2(P,Q)$$
$$\mathcal{H}_1 : TV(Q,R) \le \beta^2 H^2(P,Q).$$

3. *KL distance $KL(\cdot, R)$:*

$$\mathcal{H}_0 : KL(P,R) \le 2\beta^2 H^2(P,Q)$$
$$\mathcal{H}_1 : KL(Q,R) \le 2\beta^2 H^2(P,Q).$$

4. *KL distance $KL(R, \cdot)$:*

$$\mathcal{H}_0 : KL(R,P) \le 2\beta^2 H^2(P,Q)$$
$$\mathcal{H}_1 : KL(R,Q) \le 2\beta^2 H^2(P,Q).$$

3 Experiments

We first evaluate Neyman-Pearson test and HellingerTest on few canonical distributions without distribution perturbations and demonstrate that they have similar performance. For these experiments, we set the threshold t such that the type I error is at most 0.05. The results are in Figure 1. The experiments are averaged over 1000 trials for statistical consistency. The behavior of Neyman-Pearson and HellingerTest are similar.

We then evaluate the effect of robustness for Gaussian distributions and Bernoulli distributions in Figures 2 and 3 respectively. In these experiments, we choose R such that $\mathrm{H}(P,R) \le \mathrm{H}(Q,R)$ and plot the performance of Neyman-Pearson test and HellingerTest. We set the threshold to one for the Neyman-Pearson test and and we set the threshold to zero for the HellingerTest. The experiments demonstrate that HellingerTest is robust to distribution perturbations, where as the Neyman-Pearson test is not.

4 Proof of Theorem 1

The analysis of the test statistic involves computing the variance and the expectation and using the Bernstein inequality. The next lemma bounds the variance in terms of Hellinger distance.

Lemma 3. *For any two distributions P and Q, if $X^n \sim R$, then*

$$Var(T(P,Q,X^n)) \le \frac{55}{n} \max\left(H^2(P,R), H^2(Q,R)\right).$$

Proof. Since $X_1, X_2, \ldots, X_n$, are i.i.d. samples from R,

$$\begin{aligned}\mathrm{Var}(\mathrm{T}(P,Q,X^n)) &= \frac{1}{n}\mathrm{Var}\left(\frac{P(X_1)-Q(X_1)}{P(X_1)+Q(X_1)}\right)\\ &\le \frac{1}{n}\mathbb{E}_{X\sim R}\left[\left(\frac{P(X)-Q(X)}{P(X)+Q(X)}\right)^2\right].\end{aligned}$$

For $\beta > 1$, let S be the set given by $\{x : R(X) > \beta(P(X)+Q(X))\}$. For $x \in S$,

$$(\sqrt{P}(x) - \sqrt{R}(x))^2 \ge R(x)(\sqrt{\beta}-1)^2/\beta. \quad (5)$$

Hence,

$$\begin{aligned}&\mathbb{E}_{X\sim R}\left[\left(\frac{P(X)-Q(X)}{P(X)+Q(X)}\right)^2\right]\\ &= \mathbb{E}_{X\sim R}\left[\left(\frac{P(X)-Q(X)}{P(X)+Q(X)}\right)^2 1_{x\in S}\right]\\ &+ \mathbb{E}_{X\sim R}\left[\left(\frac{P(X)-Q(X)}{P(X)+Q(X)}\right)^2 1_{x\notin S}\right]\\ &\overset{(a)}{\le} \mathbb{E}_{X\sim R}[1_{x\in S}] + \beta\chi^2(P,Q)\\ &\overset{(b)}{\le} \frac{2\beta}{(1-\sqrt{\beta})^2}\mathrm{H}^2(P,R) + \beta\chi^2(P,Q)\\ &\overset{(c)}{\le} \frac{2\beta}{(1-\sqrt{\beta})^2}\mathrm{H}^2(P,R) + 4\beta\mathrm{H}^2(P,Q)\\ &\overset{(d)}{\le} \frac{2\beta}{(1-\sqrt{\beta})^2}\mathrm{H}^2(P,R) + 16\beta\max\left(\mathrm{H}^2(P,R), \mathrm{H}^2(Q,R)\right)\\ &\le \left(\frac{2\beta}{(1-\sqrt{\beta})^2} + 16\beta\right)\cdot\max\left(\mathrm{H}^2(P,R), \mathrm{H}^2(Q,R)\right).\end{aligned}$$

(a) follows from the definition of set S and χ^2 statistic. (5) implies (b). (3) implies (c). (d) follows from triangle inequality and the fact that $(a+b)^2 \le 4\max(a^2, b^2)$. Minimizing over $\beta > 1$ yields the lemma. ☐

In the next lemma we bound the expectation, which is the crucial part of our proof.

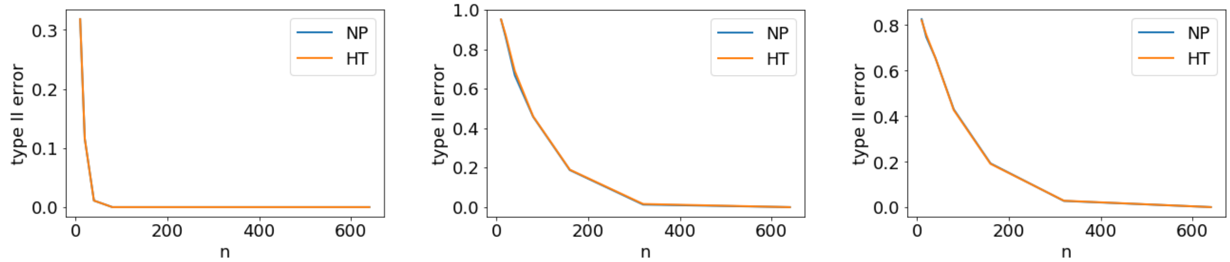
Lemma 4. *For distributions $P, Q,$ and R, if $H(Q,R) \ge \frac{\sqrt{2}}{\sqrt{2}\alpha - 1}H(P,R)$, for $\alpha \in (1/\sqrt{2}, 1)$, then*

$$\mathbb{E}[T(P,Q,X^n)] \ge 2(1-\alpha^2)H^2(Q,R).$$

Proof. Since $X_1, X_2, \ldots, X_n$ are i.i.d. samples from R,

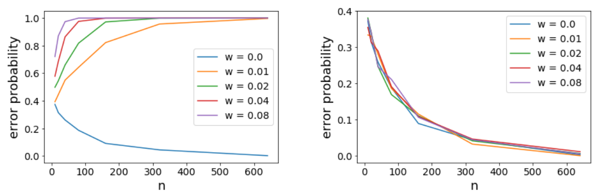
$$\begin{aligned}&2\mathbb{E}[\mathrm{T}(P,Q,X^n)]\\ &= \int_{x\in\mathcal{X}} \frac{(Q(x)-R(x))^2 - (P(x)-R(x))^2}{P(x)+Q(x))}\\ &+ (P(x) - Q(x)dx\\ &= \int_{x\in\mathcal{X}} \frac{(Q(x)-R(x))^2 - (P(x)-R(x))^2}{P(x)+Q(x)}dx,\end{aligned}$$

where the last equality follows from the fact that P and Q are probability distributions and hence integrates to 1. For any three non-negative numbers p, q, and r, it



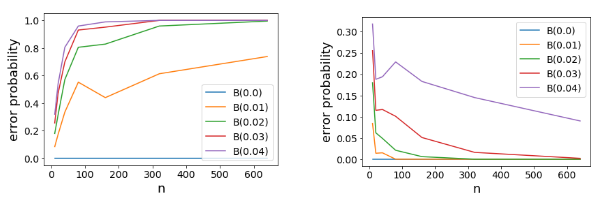
(a) $P = B(0)$, $Q = B(0.1)$. (b) $P = B(0.5)$, $Q = B(0.6)$. (c) $P = N(0,1)$, $Q = N(0.2,1)$.

Figure 1: Comparision of Neyman-Pearson and HELLINGERTEST for different distributions.



(a) Neyman-Pearson test. (b) HELLINGERTEST.

Figure 2: Comparision of Neyman-Pearson and HELLINGERTEST with perturbed Gaussian distributions. $P = N(0,1)$, $Q = N(0.2,1)$, and $R = (1-w)P + wN(100,1)$.



(a) Neyman-Pearson test. (b) HELLINGERTEST.

Figure 3: Comparision of Neyman-Pearson and HELLINGERTEST with perturbed Bernoulli distributions. $P = B(0)$, $Q = B(0.1)$, and $R = B(r)$.

can be shown that

$$\frac{(q-r)^2}{p+q} - \frac{(p-r)^2}{p+q} = \frac{(q-r)^2}{q+r} - \frac{(p-r)^2}{p+r} + \frac{2r(q-r)(p-r)(p-q)}{(p+q)(q+r)(p+r)}.$$

Applying the above equality in the expectation, and substituting the definition of χ^2 statistic,

$$2\mathbb{E}[\mathrm{T}(P,Q,X^n)] \geq \chi^2(Q,R) - \chi^2(P,R) - 2\left\| \frac{(P-Q)(P-R)(Q-R)R}{(P+Q)(P+R)(Q+R)} \right\|_1$$

We first bound the last term.

$$\begin{aligned} &\left\| \frac{(P-Q)(P-R)(Q-R)R}{(P+Q)(P+R)(Q+R)} \right\|_1 \\ &\leq \left\| \frac{(P-R)(Q-R)R}{(P+R)(Q+R)} \right\|_1 \\ &\leq \left\| \frac{(P-R)(Q-R)}{\sqrt{(P+R)(Q+R)}} \right\|_1 \left\| \frac{R}{\sqrt{(P+R)(Q+R)}} \right\|_\infty \\ &\leq \left\| \frac{(P-R)(Q-R)}{\sqrt{(P+R)(Q+R)}} \right\|_1 \\ &\leq \left\| \frac{(P-R)}{\sqrt{(P+R)}} \right\|_2 \left\| \frac{(Q-R)}{\sqrt{(Q+R)}} \right\|_2 \\ &= \sqrt{\chi^2(Q,R)\chi^2(P,R)}, \end{aligned}$$

where the last inequality follows by the Cauchy-Schwarz inequality. Combining the above equations,

$$2\mathbb{E}[\mathrm{T}(P,Q,X^n)] \geq \chi^2(Q,R) - \chi^2(P,R) - 2\sqrt{\chi^2(Q,R)\chi^2(P,R)}.$$

Let $\gamma = \frac{\mathrm{H}(Q,R)}{\mathrm{H}(P,R)} \geq 2$. We now lower bound the above term in terms of Hellinger distances.

$$\begin{aligned} &\chi^2(Q,R) - \chi^2(P,R) - 2\sqrt{\chi^2(Q,R)\chi^2(P,R)} \\ &\stackrel{(a)}{\geq} \chi^2(Q,R) - 4\mathrm{H}^2(P,R) - 4\sqrt{\chi^2(Q,R)\mathrm{H}^2(P,R)} \\ &\stackrel{(b)}{\geq} 2\mathrm{H}^2(Q,R) - 4\mathrm{H}^2(P,R) - 4\sqrt{2\mathrm{H}^2(Q,R)\mathrm{H}^2(P,R)} \\ &\stackrel{(c)}{\geq} 2\mathrm{H}^2(Q,R) - \frac{4\mathrm{H}^2(Q,R)}{\gamma^2} - \frac{4\sqrt{2}\mathrm{H}^2(Q,R)}{\gamma}, \end{aligned}$$

where (a) follows by (3). $z - 4\mathrm{H}^2(P,R) - 4\sqrt{z\mathrm{H}^2(P,R)}$ is an increasing function of $z \in [4\mathrm{H}^2(P,R), \infty)$. Furthermore by (3), $\chi^2(Q,R) \geq 2\mathrm{H}^2(Q,R) \geq 2\gamma^2\mathrm{H}^2(P,R) \geq 4\mathrm{H}^2(P,R)$. Hence substituting a lower bound on $\chi^2(Q,R)$ yields (b). (c) follows from the definition of γ. Hence,

$$\mathbb{E}[\mathrm{T}(P,Q,X^n)] \geq \mathrm{H}^2(Q,R)\left(1-\frac{2}{\gamma^2}-\frac{2\sqrt{2}}{\gamma}\right)$$
$$= \mathrm{H}^2(Q,R)\left(2-\left(\frac{\sqrt{2}}{\gamma}+1\right)^2\right).$$

Substituting $\gamma = \frac{\sqrt{2}}{\sqrt{2}\alpha-1}$ yields the result. □

The proof of Theorem 1 uses the Bernstein inequality, which we state for completeness.

Lemma 5 (Bernstein inequality)**.** *Let $Z_1, Z_2, \ldots Z_n$ are i.i.d. random variables and $M = \max_Z |Z|$ and $\sigma(Z)$ denote its variance. Then with probability at least $1-\delta$,*

$$\mathbb{E}[Z] - \frac{1}{n}\sum_i Z_i \leq\ 2\sigma(Z)\sqrt{\frac{\log\frac{1}{\delta}}{n}} + \frac{4M}{3n}\log\frac{1}{\delta}.$$

Proof of Theorem 1. Without loss of generality, we assume $\mathrm{H}(Q,R) \geq \gamma \mathrm{H}(P,R)$. Let $\gamma = \frac{\sqrt{2}}{\sqrt{2}\alpha-1}$ for $\alpha \in (1/\sqrt{2}, 1)$. We apply Bernstein theorem based on our bounds on expectations and variances. In particular, let $Z = \frac{P(x)-Q(x)}{P(x)+Q(x)}$. Hence by Lemma 4,

$$\mathbb{E}[Z] \geq 2(1-\alpha^2)\mathrm{H}^2(Q,R).$$

By Lemma 3,

$$\sigma(Z) \leq 8\mathrm{H}(Q,R),$$

and $M = \max_Z |Z| \leq 1$. Hence, with probability at least $1-\delta$,

$$\mathrm{T}(P,Q,X^n) \geq 2(1-\alpha^2)\mathrm{H}^2(Q,R) - \frac{c\mathrm{H}(Q,R)\sqrt{\log\frac{1}{\delta}}}{\sqrt{n}} - \frac{c\log\frac{1}{\delta}}{n}, \quad (6)$$

for some constant $c > 1$. Hence if $n \geq \frac{100c^2\log\frac{1}{\delta}}{\mathrm{H}^2(Q,R)(1-\alpha^2)^2}$, then with probability at least $1-\delta$,

$$\mathrm{T}(P,Q,X^n) \geq (1-\alpha^2)\mathrm{H}^2(Q,R) > 0.$$

The theorem follows by observing that

$$\mathrm{H}(P,Q) \leq \mathrm{H}(P,R) + \mathrm{H}(Q,R) \leq (\gamma+1)\mathrm{H}(Q,R).$$

□

5 Proofs of other results

5.1 Proof of Lemma 1

We give a simple example with Bernoulli distributions. Similar results hold for other distributions such as Gaussian mixtures. Let $B(p)$ be the Bernoulli distribution with parameter p. Let $P = B(0)$ and $Q = B(1/2)$. By (1), $N^*_{1/3}(P,Q) = \Theta(1)$.

Let $R = B(1/(16\gamma^2))$. It can be shown that $\mathrm{H}(P,R) \leq 1/(4\gamma)$ and $\mathrm{H}(Q,R) \geq 1/3$ for $\gamma > 1$. Hence,

$$\frac{\mathrm{H}(Q,R)}{\mathrm{H}(P,R)} \geq \frac{\frac{1}{3}}{\frac{1}{4\gamma}} \geq \gamma.$$

Let $\delta > 0$. Given $n \geq 16\gamma^2 \log\frac{1}{\delta}$ samples from R, then with probability at least $1-\delta$, at least one of the symbols is 1. Then, $P(X^n) = 0$ and $Q(X^n) = 1/2^n$ and for any finite threshold t, the test outputs $\mathcal{H}_1$. Hence, the error probability of the Neyman-Pearson test is at least $1-\delta$. Taking the limit as $\delta \to 0$ shows that Neyman-Pearson test is not robust.

5.2 Proof of Lemma 2

Let $\mathcal{X} = \{0,1,2\}$ and $\epsilon = 1/(2K)$. Let P be given by $P(0) = 1/2$, $P(1) = 1/2-\epsilon$, $P(2) = \epsilon$. Let Q be given by $Q(0) = 1/2-\epsilon$, $Q(1) = 1/2+\epsilon$, and $Q(2) = 0$.

The Hellinger distance between P and Q is $\Theta(\sqrt{\epsilon})$. By (1), $N^*_\delta(P,Q) = \Theta(\log(1/\delta)/\epsilon)$.

Scheffe's test measures empirical probability of $S = \{x : P(x) \geq Q(x)\}$ and infers the underlying hypothesis. For the above example, $S = \{0,2\}$. For this set S, $P(S) = \frac{1}{2}+\epsilon$ and $Q(S) = \frac{1}{2}-\epsilon$. Hence, the sample complexity of Scheffe test is lower bounded by the sample complexity of the best hypothesis test between $B(1/2+\epsilon)$ and $B(1/2-\epsilon)$. Therefore by (1),

$$\begin{aligned} N^S_\delta(P,Q) &= \Omega(\log(1/\delta)/\epsilon^2) \\ &= \Omega(N^*_\delta(P,Q)/\epsilon) \\ &= \Omega(K \cdot N^*_\delta(P,Q)). \end{aligned}$$

5.3 Proof of Theorem 2

Let $Q = B(0)$, $P = B(2\epsilon)$, and $R = B(\epsilon)$, where we choose ϵ later. For this choice of P, Q, and R,

$$\mathbb{E}_{X^n\sim R}[T(P,Q,X^n)] = 0.$$

We now bound the ratio of Hellinger distances,

$$\frac{\mathrm{H}^2(Q,R)}{\mathrm{H}^2(P,R)} = \frac{1-\sqrt{1-\epsilon}}{1-\sqrt{(1-2\epsilon)(1-\epsilon)}-\sqrt{2}\epsilon}.$$

Taking the right limit as $\epsilon \to 0$ and using L'Hôpital's rule yields,

$$\lim_{\epsilon\to 0^+} \frac{1-\sqrt{1-\epsilon}}{1-\sqrt{(1-2\epsilon)(1-\epsilon)}-\sqrt{2}\epsilon} = \frac{1/2}{3/2-\sqrt{2}} = \frac{1}{3-2\sqrt{2}} = \frac{1}{(\sqrt{2}-1)^2}.$$

Hence, for every $\gamma < \frac{1}{\sqrt{2}-1}$, there exists an ϵ such that $\frac{\mathrm{H}^2(Q,R)}{\mathrm{H}^2(P,R)} \geq \gamma^2$ and $\mathbb{E}_{X^n\sim R}[T(P,Q,X^n)] = 0$.

5.4 Proof of Corollary 1

We provide the proof when $R = P$. The proof for the case when $R = Q$ is similar and omitted. By the tail bounds of the Laplace random variable, there exists a constant c' such that with probability at least $1-\delta/2$,

$$\begin{aligned} T_\epsilon(P,Q,X^n) &\geq T(P,Q,X^n) - \frac{2c'\Delta(P,Q)\cdot\log\frac{2}{\delta}}{n\epsilon} \\ &\geq T(P,Q,X^n) - \frac{c'\cdot\log\frac{2}{\delta}}{n}. \end{aligned}$$

Since $R = P$, by (3),

$$\mathbb{E}[\mathrm{T}(P,Q,X^n)] = \frac{1}{2}\chi^2(P,Q) \geq \mathrm{H}^2(P,Q).$$

Similar to the proof of Theorem 1, applying the Bernstein inequality yields that with probability at least $1-\delta/2$,

$$\mathrm{T}(P,Q,X^n) \geq \mathrm{H}^2(P,Q) - \frac{c\mathrm{H}(Q,P)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}} - \frac{c\log\frac{2}{\delta}}{n}.$$

Combining the above two equations yields that with probability at least $1-\delta$,

$$\begin{aligned} T_\epsilon(P,Q,X^n) \geq \mathrm{H}^2(P,Q) &- \frac{c\mathrm{H}(Q,P)\sqrt{\log\frac{2}{\delta}}}{\sqrt{n}} \\ &- \frac{(c+c')\log\frac{2}{\delta}}{n}. \end{aligned}$$

Hence if $n \geq c''\left(\frac{\log\frac{1}{\delta}}{\mathrm{H}^2(P,Q)}\right)$ for a sufficiently large constant c'', then with probability at least $1-\delta$,

$$T_\epsilon(P,Q,X^n) > 0,$$

and hence the result.

6 Relationship between distances

6.1 Relationship between Hellinger distance and total variation distance

Upper bound:

$$\begin{aligned} \mathrm{H}^2(P,Q) &= \frac{1}{2}\left\|\sqrt{P}-\sqrt{Q}\right\|_2^2 \\ &\leq \frac{1}{2}\left\|(\sqrt{P}-\sqrt{Q})(\sqrt{P}+\sqrt{Q})\right\|_1 \\ &= \frac{1}{2}\|P-Q\|_1 \\ &= \mathrm{TV}(P,Q). \end{aligned}$$

Lower bound:

$$\begin{aligned} \mathrm{H}^2(P,Q) &= \frac{1}{2}\left\|\sqrt{P}-\sqrt{Q}\right\|_2^2 \\ &\overset{(a)}{\geq} \frac{1}{8}\left\|\sqrt{P}-\sqrt{Q}\right\|_2^2 \cdot \left\|\sqrt{P}+\sqrt{Q}\right\|_2^2 \\ &\overset{(b)}{\geq} \frac{1}{8}\|P-Q\|_1^2 \\ &= \frac{1}{2}\mathrm{TV}^2(P,Q), \end{aligned}$$

where (a) follows from the fact that $\left\|\sqrt{P}+\sqrt{Q}\right\|_2 \leq 2$ and (b) uses the Cauchy-Schwarz inequality.

6.2 Relationship between Hellinger distance and symmetric chi-squared statistic

$$\begin{aligned} \mathrm{H}^2(P,Q) &= \frac{1}{2}\left\|\sqrt{P}-\sqrt{Q}\right\|_2^2 \\ &= \frac{1}{2}\left\|\frac{(\sqrt{P}-\sqrt{Q})(\sqrt{P}+\sqrt{Q})}{(\sqrt{P}+\sqrt{Q})}\right\|_2^2 \\ &= \frac{1}{2}\left\|\frac{P-Q}{(\sqrt{P}+\sqrt{Q})}\right\|_2^2. \end{aligned}$$

The proof of (3) follows by observing that for every x,

$$\sqrt{P(x)+Q(x)} \leq \sqrt{P(x)}+\sqrt{Q(x)} \leq \sqrt{2(P(x)+Q(x))}.$$

7 Conclusion

We proposed a simple robust hypothesis test that has the same complexity of the optimal Neyman-Pearson test up to constants and is robust to distribution perturbations in Hellinger distance. The test is relatively parameter free and easy to use. We evaluated the test on synthetic distributions and also provided extensions with differential privacy. Bridging the $\sqrt{2}$-gap between the upper and lower bounds is an interesting future direction.

8 Acknowledgements

We thank Jaydev Acharya and Yihong Wu for helpful comments and suggestions. We thank Sivaraman Balakrishnan for helpful comments and bringing (Baraud, 2011) to our attention.

References

J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Sorting with adversarial comparators and application to density estimation. In *2014 IEEE International Symposium on Information Theory*, pages 1682–1686. IEEE, 2014.

J. Acharya, M. Falahatgar, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Maximum selection and sorting with adversarial comparators. *The Journal of Machine Learning Research*, 19(1):2427–2457, 2018.

H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Advances in Neural Information Processing Systems*, pages 3412–3421, 2018.

Z. Bar-Yossef and C. H. Papadimitriou. *The complexity of massive data set computations*. PhD thesis, University of California, Berkeley, 2002.

Y. Baraud. Estimator selection with respect to hellinger-type risks. *Probability theory and related fields*, 151(1-2):353–401, 2011.

M. Bun, G. Kamath, T. Steinke, and S. Z. Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems*, pages 156–167, 2019.

C. L. Canonne, G. Kamath, A. McMillan, A. Smith, and J. Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321. ACM, 2019.

S.-O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613, 2014.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, pages 1183–1213, 2014.

C. Daskalakis, I. Diakonikolas, and R. A. Servedio. Learning k-modal distributions via testing. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1371–1385. SIAM, 2012.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

I. Diakonikolas, D. M. Kane, and A. Stewart. Learning multivariate log-concave distributions. In *Conference on Learning Theory*, pages 711–727, 2017.

B. J. Dong. Minimum hellinger distance bsed goodness-of-fit tests in normal models: Empirical approach. *Communications for Statistical Applications and Methods*, 6(3):967–976, 1999.

C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

R. Gao, L. Xie, Y. Xie, and H. Xu. Robust hypothesis testing using wasserstein uncertainty sets. In *NeurIPS*, pages 7913–7923, 2018.

G. Gül and A. M. Zoubir. Robust hypothesis testing with squared hellinger distance. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 1083–1087. IEEE, 2014.

G. Gül and A. M. Zoubir. Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9):5572–5587, 2017.

P. J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

B. C. Levy. Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1):413–421, 2008.

J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

H. Scheffé. A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics*, 18(3):434–438, 1947.

A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014.

Y. G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pages 768–774, 1985.