IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 May 2022, accepted 28 June 2022, date of publication 4 July 2022, date of current version 13 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187991

RESEARCH ARTICLE

Score-Based Hypothesis Testing for Unnormalized Models

SUYA WU[1], ENMAO DIAO[1], (Member, IEEE), KHALIL ELKHALIL[1], JIE DING[2], (Member, IEEE), AND VAHID TAROKH[1], (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
[2]School of Statistics, University of Minnesota, Twin Cities, Minneapolis, MN 55455, USA

Corresponding author: Suya Wu (suya.wu@duke.edu)

This work was supported in part by the Air Force Research Laboratory under Grant FA-8750-20-2-0504. The work of Jie Ding was supported by the National Science Foundation under Grant ECCS-2038603.

ABSTRACT Unnormalized statistical models play an important role in machine learning, statistics, and signal processing. In this paper, we derive a new hypothesis testing procedure for unnormalized models. Our approach is motivated by the success of score matching techniques that avoid the intensive computational costs of normalization constants in many high-dimensional settings. Our proposed test statistic is the difference between Hyvärinen scores corresponding to the null and alternative hypotheses. Under some reasonable conditions, we prove that the asymptotic distribution of this statistic is Chi-squared. We outline a bootstrap approach to learn the test critical values, particularly when the distribution under the null hypothesis cannot be expressed in a closed form, and provide consistency guarantees. Finally, we conduct extensive numerical experiments and demonstrate that our proposed approach outperforms goodness-of-fit benchmarks in various settings.

INDEX TERMS Hypothesis testing, goodness-of-fit, score matching, unnormalized statistical models.

I. INTRODUCTION

Hypothesis testing is a procedure to decide whether or not to reject a hypothesis. The likelihood ratio test (LRT) is a standard method of hypothesis testing. The celebrated Neyman-Pearson lemma gives the uniformly most powerful optimality of LRT for simple hypotheses testing. However, the application of LRT requires exact likelihoods. The evaluation of exact likelihoods may be computationally cumbersome (or even intractable) for various modern statistical models including graphical models [1], energy-based models [2] and deep generative models [3].

In theory, Monte Carlo-based methods can be used to approximate the likelihoods required for evaluating LRT (e.g., [4] and the references therein). However, in practice, these approximations may suffer from computational errors. This makes it difficult to perform ideal LRT. In particular, the asymptotic optimality of LRT is not guaranteed when the class of alternatives is very large or when typical regularity conditions are violated [5, Chapter 14]. Score matching [6], [7] is a procedure for empirical minimization of the Fisher divergence from the data-generating distribution to the postulated model-based distribution. This estimation procedure has been extended to discrete [8], non-parametric [9], and directional [10] distributions. Recently, score matching-based deep generative modeling demonstrated great success and outperformed likelihood-based benchmarks in image generations [11], [12]. These prior works showed that score matching successfully selects the appropriate unnormalized model in a data-driven manner. Moreover, existing studies introduced the Hyvärinen score to Bayesian model comparison with improper or vague priors (see, e.g., [13] and the references therein). However, to the best of our knowledge, a hypothesis testing framework based on score matching has not been developed.

In this paper, we propose a new hypothesis testing procedure for unnormalized statistical models. The test statistic is defined based on the Hyvärinen score, which provides an alternative to LRT. Specifically, the difference between the Hyvärinen scores associated with the null and

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

alternative distributions is used as the test statistic, which we refer to as the Hyvärinen score test (HST). The value of HST is intimately related to Fisher divergence, which will be introduced later. We provide a summary of our contributions below.

- We propose a new test statistic consisting only of the gradient and Laplacian of the log-likelihood terms. These terms are scale-invariant and are computable for unnormalized distributions. It leads to a significant reduction in the computational complexity of the test.
- We establish the asymptotic properties of the HST under the null hypothesis. For the asymptotic distribution not given in a closed form, we propose a bootstrap procedure for estimating the critical value of the test, and we demonstrate its consistency in terms of the test power.
- We evaluate the performance of the HST test in various numerical experiments and compare it with the LRT, the Kernelized Stein Discrepancy (KSD) test [14], and the Maximum Mean Discrepancy (MMD) test [15]. The results demonstrate that our method performs competitively with LRT and outperforms other baseline methods. Our experiments further illustrate the computational advantage of our approach for unnormalized models over LRT. Additionally, the proposed approach achieves success in the OOD detection task.

The outline of our paper is described as follows. In Sections II and III, we briefly review the related work and the mathematical background for score matching and its connections to Fisher divergence. In Section IV, we formulate the composite hypothesis testing based on score matching estimates and provide theoretical justifications. In Section V, we compare our method with various baseline methods with synthetic data. We also evaluate our method for OOD detection with image and network intrusion datasets. We draw our conclusions in Section VII.

II. RELATED WORK

Our work on hypothesis testing of unnormalized models is related to KSD-based goodness-of-fit tests [14], [16] and its variations, including random feature Stein discrepancies (RΦSDs) [17], composite tests with KSD [18], and Sliced KSD (SKSD) [19]. The family of KSD tests is defined based on the Stein discrepancy, which is often computed in a reproducing kernel Hilbert space (RKHS). The test statistics for KSD tests are in the form of U-statistics or V-statistics (estimates of KSD from the data generating distribution to the null hypothesis distribution). Performing the test requires knowledge of the score function, defined as the gradient of logarithmic density function over the input, which is usually easy to compute for unnormalized statistical models. However, the test statistic is degenerate (to a single point) under the null hypothesis, requiring the wild bootstrap procedure for empirical rejection region [20]–[22]. We observe that the wild bootstrap procedure tends to overestimate the test power and has inaccurate Type I error probability when the sample size is small. Also, KSD-based tests may be limited to a particular kernel, which is commonly chosen to be the RBF kernel [23].

Recently, hypothesis testing for out-of-distribution (OOD) detection and distribution-shift detection has emerged as important areas of machine learning [24]–[28]. These emerging topics also have some intersection with the important areas of anomaly detection, adversarial attacks detection, and incremental learning. For example, in OOD detection, the in-distribution samples are assumed to follow a distribution $p(\cdot)$, while the OOD samples are distributed according to another distribution $q(\cdot)$. For an observation x, one may formulate a hypothesis testing problem as $\mathcal{H}_0 : x \sim p$ against $\mathcal{H}_1 : x \not\sim p$. In recent works, LRT-inspired empirical methods are often proposed to perform hypothesis testing for OOD detection [26]–[28]. The main assumption behind these LRT-inspired approaches is that OOD samples have a low likelihood under the in-distribution, whose probabilistic distribution has to be learned from a set of in-distribution data.

Recent works have attempted to perform OOD detection based on score matching estimates [29], [30]. The work of [29] utilizes the norm of the gradient of logarithmic likelihood at multiple noise scales for anomaly detection. The work of [30] estimates the Fisher divergence between the null and alternative distributions for feature shift detection, where they assume the data are generated from the null distribution. The test statistic we propose can be seen as the difference between two Fisher divergences, the first one from the data-generating distribution to the null distribution, and the second one from the data-generating distribution to the alternative distributions.

III. ESTIMATION OF UNNORMALIZED STATISTICAL MODELS BY SCORE MATCHING

In the rest of the paper, a data point is represented by a random variable $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. Consider the probability density functions (PDFs) $\mathbf{x} :\mapsto p(\mathbf{x})$ and $\mathbf{x} :\mapsto q(\mathbf{x})$ that represent two probability distributions on $\mathcal{X}$. The Fisher divergence from p to q is defined by

$$\mathbb{D}_{\mathrm{F}}[p||q] \triangleq \mathbb{E}_{\mathbf{x}\sim p}\left[\frac{1}{2}\left\|\nabla_{\mathbf{x}}\log p(\mathbf{x}) - \nabla_{\mathbf{x}}\log q(\mathbf{x})\right\|_2^2\right], \quad (1)$$

whenever the integral is well defined. It is worth noting that in general, one does not require the density functions to define a Fisher divergence for two distributions. We assume the existence of PDFs only for notational convenience. Under some mild regularity conditions on $p(\cdot)$ and $q(\cdot)$ [6], the Fisher divergence can be rewritten as

$$\mathbb{D}_{\mathrm{F}}[p||q] = \mathbb{E}_{\mathbf{x}\sim p}\left[\frac{1}{2}\left\|\nabla_{\mathbf{x}}\log p(\mathbf{x})\right\|_2^2 + s_{\mathrm{H}}(q,\mathbf{x})\right], \quad (2)$$

where $(q,\mathbf{x}) \mapsto s_{\mathrm{H}}(q,\mathbf{x})$ is given by

$$s_{\mathrm{H}}(q,\mathbf{x}) \triangleq \frac{1}{2}\left\|\nabla_{\mathbf{x}}\log q(\mathbf{x})\right\|_2^2 + \Delta_{\mathbf{x}}\log q(\mathbf{x}), \quad (3)$$

also known as the Hyvärinen score [6]. Here, $\Delta_{\mathbf{x}} = \sum_{i=1}^{d}\frac{\partial^2}{\partial x_i^2}$ denotes the Laplacian operator with respect to

$\mathbf{x} = (x_1, \cdots, x_d)^\top$. The above score function is a counterpart of the logarithmic score function $(q, \mathbf{x}) \mapsto -\log q(\mathbf{x})$ that corresponds to the widely used negative log-likelihood. Notably, both scores can be interpreted from an information-theoretic perspective [31] and used for statistical estimation. In particular, suppose that PDFs p and q respectively represent the (unknown) data-generating distribution and the postulated distribution, where the latter is parameterized by a density function in the form of $\mathbf{x} \mapsto q_\theta(\mathbf{x})$, $\theta \in \Theta$. One may estimate the unknown parameter θ by minimizing the Fisher divergence in Equation (2), replacing q with q_θ. The minimization problem is then reduced to the minimization of the expected Hyvärinen score

$$\theta^\star = \arg\min_{\theta\in\Theta} \mathbb{E}_{\mathbf{x}\sim p} s_{\mathrm{H}}(q_\theta, \mathbf{x}). \tag{4}$$

In the above, the subscript $\mathbf{x} \sim p$ denotes the expectation with respect to the data-generating distribution. If there exist multiple optimal solutions in Problem (4), then $\theta^\star$ can be arbitrarily selected to be any of these minimizing values. When the parametric family is well-specified in a strict sense [32], i.e. it includes the data-generating distribution, then $\mathbb{D}_{\mathrm{F}}[p||q] \geq 0$ attains its minimum $\mathbb{D}_{\mathrm{F}}[p||q_{\theta^\star}] = 0$ at $q_{\theta^\star} = p$ for $\theta^\star \in \Theta$. In this case, the $\theta^\star$ obtained from Problem (4) can be interpreted as the data-generating parameter. In the rest of the paper, we use $\mathbb{E}_\star$ to denote the expectation with respect to the (unknown) data-generating distribution of $\mathbf{x}$, namely $p_{\theta^\star}$.

Suppose that a finite sample of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed (IID) according to p. For a postulated parametric model represented by $\{q_\theta, \theta \in \Theta\}$, an estimator in parallel to Problem (4) is

$$\hat{\theta}_{\mathrm{sm}} \triangleq \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} s_{\mathrm{H}}(q_\theta, \mathbf{x}_i). \tag{5}$$

The procedure of estimating the above $\hat{\theta}_{\mathrm{sm}}$, denoted by "sm," is known as score matching estimation (see, e.g., [6], [31], [33] and the references therein), which is a surrogate for maximum likelihood estimation (MLE). It can be readily applied to unnormalized distributions, because the $s_{\mathrm{H}}(q_\theta, \mathbf{x})$ introduced in (3) is invariant when q_θ is replaced with $c_\theta q_\theta$ for any scaling factor c_θ that does not depend on the data.

IV. SCORE-BASED HYPOTHESIS TESTING FOR UNNORMALIZED STATISTICAL MODELS

A. HYVÄRINEN SCORE TEST

Following the notation in Section III, we denote $\mathbf{X}_n \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ as IID observations from an unknown distribution represented by a PDF p. We represent $\mathcal{Q} = \{q_\theta : \theta \in \Theta \subset \mathbb{R}^r\}$ as a parametric family of distributions, where each q_θ is a PDF. Suppose that $p = q_{\theta^\star}$ for some $\theta^\star \in \Theta$. We are interested in testing the hypothesis if $\theta^\star = \theta_0$ for a given $\theta_0 \in \Theta$. It leads to the following hypothesis testing, formulated by

$$\mathcal{H}_0 : \theta^\star = \theta_0 \quad \text{against} \quad \mathcal{H}_1 : \theta^\star \in \Theta\backslash\{\theta_0\}. \tag{6}$$

Motivated by the aforementioned connection between score matching and maximum likelihood estimation, we develop a new statistical test, referred to as the Hyvärinen score test (HST), based on the score in Equation (3). Specifically, the score-based test statistic, denoted as T_{HST}, is defined by

$$T_{\mathrm{HST}}(\mathbf{X}_n) \triangleq 2\big(\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_0) - \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathrm{sm}})\big), \tag{7}$$

where $\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta) \triangleq \sum_{i=1}^n s_{\mathrm{H}}(q_\theta, \mathbf{x}_i)$, and $\hat{\theta}_{\mathrm{sm}}$ is defined in Equation (5).

The HST rejects the null hypothesis when the test statistic T_{HST} is larger than a critical value, which can be identified using a large-sample asymptotic distribution. The following result shows the asymptotic distribution of T_{HST} under the null hypothesis, which is analogous to the classical Wilks' theorem [34].

Theorem 1 (Asymptotic distribution of T_{HST} under the null hypothesis): Assuming the regularity conditions given in the Appendix, under the null hypothesis, we have

$$T_{\mathrm{HST}}(\mathbf{X}_n) \xrightarrow{n\to\infty} \mathcal{L}\ \mathbf{z}^T\mathbf{H}\mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_r, \mathbf{H}^{-1}\mathbf{K}\mathbf{H}^{-1})$,

$$\mathbf{H} \triangleq \mathbb{E}_\star\left[\nabla_\theta^2 s_{\mathrm{H}}(q_\theta, \mathbf{x})\ |_{\theta=\theta_0}\right], \tag{8}$$

$$\mathbf{K} \triangleq \mathbb{E}_\star\left[\nabla_\theta s_{\mathrm{H}}(q_\theta, \mathbf{x})\nabla_\theta^\top s_{\mathrm{H}}(q_\theta, \mathbf{x})\ |_{\theta=\theta_0}\right], \tag{9}$$

and $\mathcal{L}$ denotes convergence in distribution.

We provide the proof in Appendix A-A.

Remark 1 (Discussion of Theorem 1): In light of Theorem 1, we can characterize the distribution of T_{HST} under the null hypothesis when the sample size goes to infinity. In practice, the above $\mathbf{H}$ and $\mathbf{K}$ can be estimated by the finite-sample average, e.g., $\hat{\mathbf{H}} \triangleq n^{-1}\sum_{i=1}^n \left[\nabla_\theta^2 s_{\mathrm{H}}(q_\theta, \mathbf{x}_i)\ |_{\theta=\theta_0}\right]$. Next, we provide an alternative way to express $\mathbf{H}$ and $\mathbf{K}$, which does not involve second derivatives and can be used to simplify the computation. Let $g_k(\theta, \mathbf{x}) \triangleq \partial_{x_k} \log q_\theta(\mathbf{x})$, and $g'_k(\theta, \mathbf{x}) \triangleq \partial^2_{x_k} \log q_\theta(\mathbf{x})$, where x_k is the k-th entry of $\mathbf{x}$. Then, under mild conditions, we have

$$\mathbf{H}_{ij} = \sum_{k=1}^{d} \mathbb{E}_\star\left[\partial_{\theta_i} g_k(\theta, \mathbf{x}) \cdot \partial_{\theta_j} g_k(\theta, \mathbf{x})\ |_{\theta=\theta_0}\right], \tag{10}$$

$$\mathbf{K}_{ij} = \sum_{k=1}^{d} \mathbb{E}_\star\Big[\big(\partial_{\theta_i} g'_k(\theta, \mathbf{x}) \cdot \partial_{\theta_j} g'_k(\theta, \mathbf{x}) - (\partial_{\theta_i} g_k(\theta, \mathbf{x}) \cdot \partial_{\theta_j} g_k(\theta, \mathbf{x})) \cdot g'_k(\theta, \mathbf{x})\big)\ |_{\theta=\theta_0}\Big], \tag{11}$$

where $\mathbf{H}_{ij}$ and $\mathbf{K}_{ij}$ are the (i,j)-th entries of $\mathbf{H}$ and $\mathbf{K}$, respectively. We provide the detailed derivations in Appendix A-B.

Evaluating $\mathbf{K}$ can be computationally infeasible when the dimension is large. To determine the rejection region for HST in this case, we propose to use the bootstrap method developed in [35]. The main idea is to approximate the critical value by the empirical α-quantile of the distribution of T_{HST} under the null hypothesis. Specifically, we generate a pool of

samples from the null distribution. Next, we resample n sub-samples from this pool with replacement in each bootstrap iteration. Then, we calculate the bootstrap test statistic on this sub-sample data. We repeat this procedure b times until we have b bootstrap samples. Subsequently, we select the HST critical value as the empirical α-quantile value of these bootstrap statistics. Finally, we decide whether or not to reject the $\mathcal{H}_0$ by comparing the test statistic with the selected critical value.

Algorithm 1 Bootstrap Hyvärinen Score Test.

Input: Test sample $\mathbf{X}_n \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, number of bootstrap samples b, bootstrap sample size m, and significance level α

Independently sample $\mathbf{Y}_m \triangleq \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$ from the null distribution

for $i = 1, \ldots, b$ **do**

Resample $\mathbf{Y}_n^{(i)}$ from $\mathbf{Y}_m$ with replacement

Compute $T_{\mathrm{HST}}^{(i)} = 2\big(\mathcal{S}_{\mathrm{H}}(\mathbf{Y}_n^{(i)}, \theta_0) - \inf_{\theta\in\Theta} \mathcal{S}_{\mathrm{H}}(\mathbf{Y}_n^{(i)}, \theta)\big)$

end for

Determine C_α = quantile($\{T_{\mathrm{HST}}^{(1)}, \ldots, T_{\mathrm{HST}}^{(b)}\}, 1-\alpha$)

Compute $T_{\mathrm{HST}} = 2\big(\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_0) - \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathrm{SM}})\big)$

Decision: Reject $\mathcal{H}_0$ if $T_{\mathrm{HST}} > C_\alpha$

Output: The test statistic T_{HST}, and the empirical p-value, which is $b^{-1}\sum_{i=1}^{b} \mathbb{1}\{T_{\mathrm{HST}}^{(i)} > T_{\mathrm{HST}}\}$

We summarize the bootstrap HST pseudo-code in Algorithm 1. We note that the computational complexity of this procedure is $O(bmn)$, where b is the bootstrap sample size, n is the test sample size, and m is the null sample size in bootstrap iterations. Using [35, Theorem 3.1], we can show that assuming $n = o(m)$, *i.e.*, $\lim_{m\to\infty} n/m \to 0$, and under some regularity conditions, the Type I error rate of the bootstrap HST goes to α, and its test power approaches one as $m \to \infty$. This is known as test consistency. We refer to [35, Theorem 3.1] for details of the regularity conditions.

B. SIMPLE HYVÄRINEN SCORE TEST

When the goal is to identify the data generating distribution from two given distributions q_{θ_0} and q_{θ_1}, where $\theta_0,\ \theta_1 \in \Theta$, one may consider the following simple hypothesis testing

$$\mathcal{H}_0 : \theta^\star = \theta_0 \quad \text{against} \quad \mathcal{H}_1 : \theta^\star = \theta_1. \tag{12}$$

We define the associated HST test statistic by

$$\tilde{T}_{\mathrm{HST}} \triangleq 2\left(\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_0) - \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_1)\right).$$

A direct application of the Central Limit Theorem gives an asymptotic distribution of $\tilde{T}_{\mathrm{HST}}$. For completeness, we provide the result below.

Proposition 1 (Asymptotic distribution of $\tilde{T}_{\mathrm{HST}}$ under the null hypothesis): Assuming the regularity conditions given in the Appendix, under the null hypothesis, we have

$$n^{-1/2} \cdot (\tilde{T}_{\mathrm{HST}} + 2\, n\mathbb{D}_{\mathrm{F}}[q_{\theta_0}||q_{\theta_1}]) \xrightarrow{n\to\infty}_{\mathcal{L}} \mathbf{z},$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_r, var_\star(s_{\mathrm{H}}(q_{\theta_0}, \mathbf{x}) - s_{\mathrm{H}}(q_{\theta_1}, \mathbf{x})))$ and $var_\star(\cdot)$ denotes the variance with respect to the null distribution.

Next, we will provide a multivariate Normal example to demonstrate the test power of HST and compare it with LRT.

Remark 2 (A multivariate Normal example): Consider a multivariate Normal distribution in $\mathbb{R}^d$ with mean $\mu^\star$ and variance Σ (known). We test the null hypothesis $\mathcal{H}_0 : \mu^\star = \mathbf{0}_d$ against the alternative hypothesis $\mathcal{H}_1 : \mu^\star = \mu_1$. We denote the test statistics of HST and LRT as T_{HST} and T_{LRT}, with test power $\beta_{\mathrm{HST},\alpha}$ and $\beta_{\mathrm{LRT},\alpha}$, respectively. Here, $\alpha \in (0,1)$ represents the significance level. Let $\tilde{\mu}_1 \triangleq \mu_1^T\Sigma^{-1}\mu_1$, and $\Phi_{1-\alpha}$ be the α upper-quantile of the standard Normal distribution. We assume that the sample size n satisfies $n > \tilde{\mu}_1/\Phi_{1-\alpha}^2$. Then, it can be derived that

$$\beta_{\mathrm{HST},\alpha} \le \beta_{\mathrm{LRT},\alpha} + \frac{1}{\sqrt{2\pi}s}\left(\exp\left\{-\frac{s^2}{2}\right\} - \exp\left\{-\frac{t^2}{2}\right\}\right),$$

where

$$s \triangleq \Phi_{1-\alpha} - \sqrt{\frac{\tilde{\mu}_1}{n}},\ \text{and}\ t \triangleq \Phi_{1-\alpha} - \frac{\sigma_{\min}}{\sigma_{\max}}\sqrt{\frac{\tilde{\mu}_1}{n}}.$$

Here, $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest eigenvalues of Σ, respectively. We provide the derivation of the above inequality in Appendix A-C.

It is known from the Neyman-Pearson Lemma that LRT is the uniformly most powerful test for the simple hypothesis testing problem. The above result shows that the test power of HST is close to that of LRT in the multivariate Normal example. In particular, HST achieves the same power as LRT, namely $\beta_{\mathrm{HST},\alpha} = \beta_{\mathrm{LRT},\alpha}$, for any significance level α when $\sigma_{\min} = \sigma_{\max}$.

V. EXPERIMENTS ON SYNTHETIC DATA

In this section, we conduct extensive numerical experiments on synthetic data drawn from multivariate Normal Distribution, Exponential family, and Gauss-Bernoulli RBM. We compare the performance of HST with those of the classical LRT, KSD-U [14], KSD-V [16], and the MMD two sample test [15] at significance level $\alpha = 0.05$.

Specifically, we evaluate two HST tests: 1) the bootstrap HST for simple hypothesis, and 2) the bootstrap HST for the composite hypothesis, respectively given in Equations (6) and (12). These are denoted by HST (simple) and HST (composite). Likewise, we perform two LRT tests: LRT (simple) and LRT (composite). For a fair comparison, we apply an identical bootstrap procedure to all HST and LRT tests considered here. In terms of KSD-U, KSD-V, and the MMD two-sample test, we apply the wild bootstrap procedure [22]. We choose the RBF kernel for all types of kernel tests and set its bandwidth as the median of pairwise data distances.

We consider two sets of experiments. First, we vary the alternative hypotheses with a fixed sample size. Second, we vary the sample sizes at a fixed perturbation level. For a given $\theta_0 \in \Theta$, the alternative parameter θ_1 is perturbed by adding a normal noise, which is distributed by $\mathcal{N}(0, \sigma_{ptb}^2)$ at different perturbation levels σ_{ptb}. For each perturbation level

or sample size, we conduct 100 random experiments. In each random experiment, we draw $m = 10^4$ samples from the null and alternative distributions. We choose bootstrap sample size $b = 1000$ in our simulations. Note that we have reused the null samples for all tests. We depict the empirical power and the Type I error rate with error bars below. Additionally, we will report the Receiver Operating Characteristics (ROC) curve and the Area Under the Curve-Receiver Operating Characteristics (AUC) to evaluate the performance of the tests.

A. MULTIVARIATE NORMAL DISTRIBUTION

Consider the bivariate Normal distribution $\mathcal{N}(\mu, \Sigma)$, and we perform tests for μ (with known Σ) and Σ (respectively with known μ). The null hypothesis is the mean μ_0 (respectively the covariance Σ_0) is equal to its true value $\mu^\star$ (respectively $\Sigma^\star$). The corresponding values for the alternative are assigned by perturbing μ_0 by a noise term $\mathcal{N}(0, \sigma_{ptb}^2)$ (respectively by multiplying the diagonal elements of Σ_0 by a log-Normal noise with the Normal distribution $\mathcal{N}(0, \sigma_{ptb}^2)$). In the rest of the contexts, we may denote the operator log to each entries of the matrix Σ as $\log(\Sigma)$.

We use $\mu_0 = (0, 0)^T$ and $\log(\Sigma_0) = [[0, -0.3], [-0.3, 0]]$ as the null hypotheses on testing μ and Σ respectively. When testing the hypothesis: $\mu^\star = \mu_0$, we let $\mu_1 = \mu_0 + \mathcal{N}(0, \sigma_{ptb}^2)$, and assume that $\Sigma^\star = \Sigma_0$ is known. In the first set of experiments, the perturbation level σ_{ptb} takes values $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.85, 0.9, 0.95, 1, 2\}$ with a fixed sample size $n = 100$. In the second set of experiments, the sample size takes values $\{5, 10, 20, 30, 40, 50, 80, 150, 200\}$ and the perturbation standard deviation is fixed to be $\sigma_{ptb} = 1$.

When testing on the hypothesis: $\Sigma^\star = \Sigma_0$, we let $\log(\Sigma_1) = \log(\Sigma_0) + \mathcal{N}(0, \sigma_{ptb}^2)$, and assume that $\mu^\star = \mu_0$ is known. The perturbation levels and the sample sizes are chosen in the same manner as above.

In Figure 1 and 2, we give power comparison between various tests by varying the perturbation levels. The results show that HST performs competitively with LRT and outperforms other baselines. In Figure 3 and 4, we provide power comparison between tests by varying the sample size. The results demonstrate that KSD-based tests and the MMD two-sample test have unsatisfactory Type I error and power performance, particularly for small sample sizes. Moreover, KSD-based tests and the MMD two-sample test demonstrate instability, particularly for small sample sizes. In contrast, HST shows remarkable consistency, computational efficiency, and robustness.

In Figure 5 and 6, we depict the test statistics of the above tests by varying σ_{ptb}. The results demonstrate that HST can discriminate the null and the alternative hypotheses quite well. We note that the test statistic of HST (simple) is decreasing under the null hypothesis, as the aggregate Hyvärinen score of the alternative distribution is subtracted from that of the null hypothesis. Therefore, the results show that HST (simple) provides superior performance.

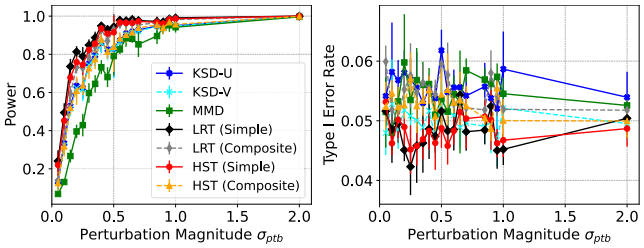


FIGURE 1. Multivariate Normal Distribution, power comparison by changing the perturbation level of μ ($n = 100$).

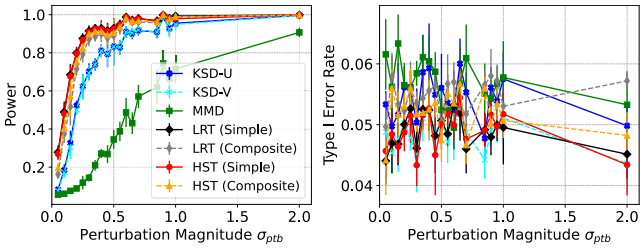


FIGURE 2. Multivariate Normal Distribution, power comparison by changing the perturbation level of $\log(\Sigma)$ ($n = 100$).

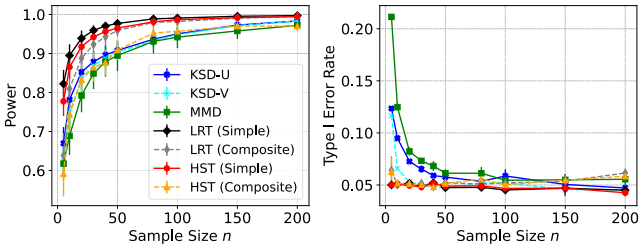


FIGURE 3. Multivariate Normal Distribution, power comparison by changing the sample size ($\sigma_{ptb} = 1$ on μ).

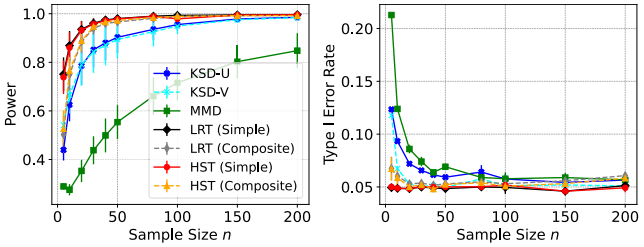


FIGURE 4. Multivariate Normal Distribution, power comparison by changing the sample size ($\sigma_{ptb} = 1$ on $\log(\Sigma)$).

B. EXPONENTIAL FAMILY

Next, we demonstrate that HST is computationally favorable over LRT on unnormalized statistical models. We assume that the random variable $\mathbf{x} \in \mathbb{R}^d$ is distributed according to

$$p_\tau(\mathbf{x}) \propto \exp\left\{-\tau\left(\sum_{i=1}^{d} x_i^4 + \sum_{1\le i\le d, i\le j\le d} x_i^2 x_j^2\right)\right\},$$

where $\tau \in \mathcal{T} \subset \mathbb{R}^+$ is the model parameter. This subfamily of exponential family belongs to pairwise interaction graphical models [36]. The calculation of the normalizing constant is difficult particularly for high dimensions. We consider the hypothesis test $\mathcal{H}_0 : \tau = \tau_0 = 1$ against $\mathcal{H}_1 :$

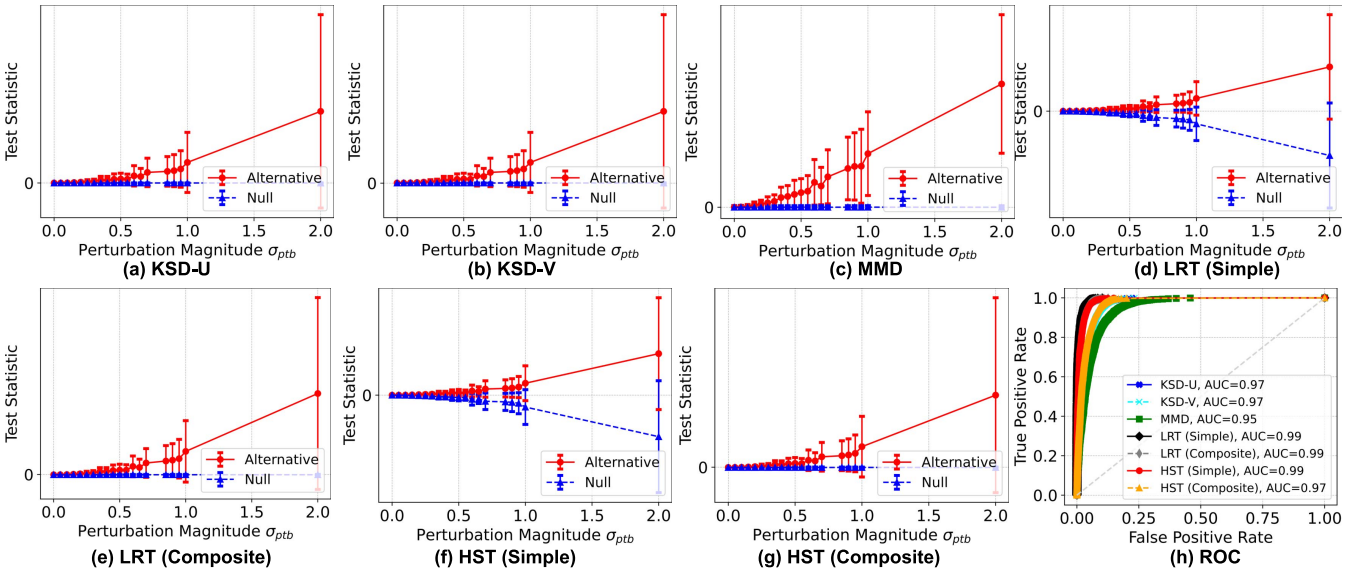


FIGURE 5. Multivariate Normal Distribution, (a-f) Values of test statistics with $n = 100$ by changing the perturbation level of μ. (g) Receiver Operating Characteristic (ROC) curves of various tests with $\sigma_{ptb} = 0.5$ and $n = 100$.

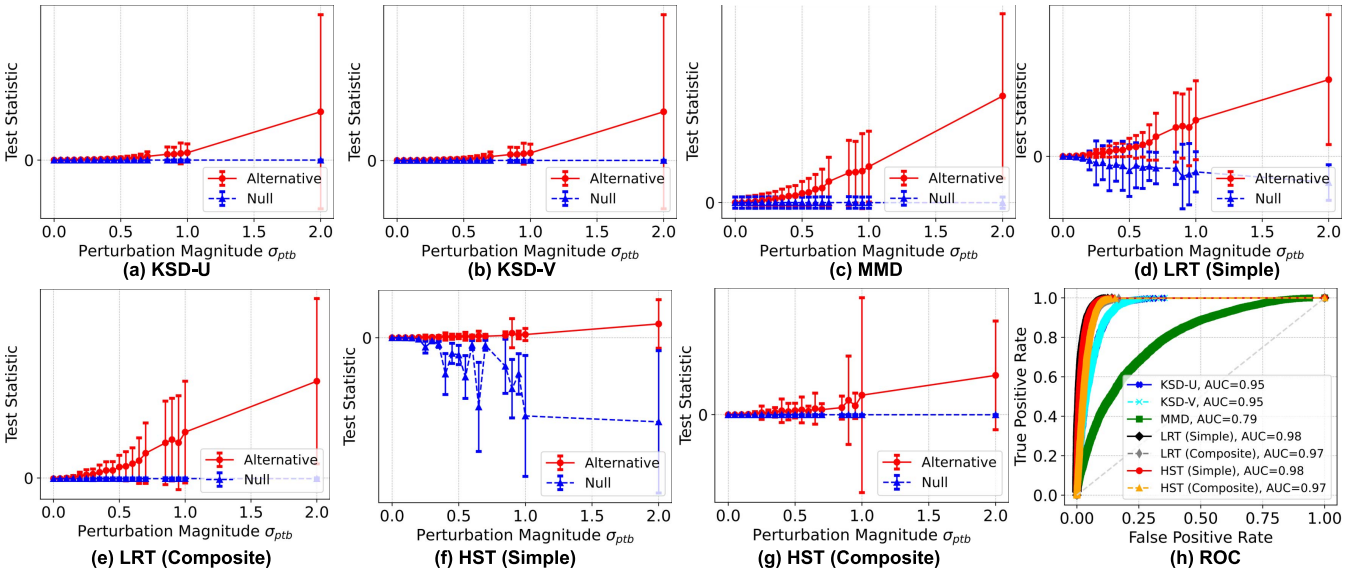


FIGURE 6. Multivariate Normal Distribution, (a-f) Values of test statistics with $n = 100$ by changing the perturbation level of log(Σ). (g) Receiver Operating Characteristic (ROC) curves of various tests with $\sigma_{ptb} = 0.5$ and $n = 100$.

$\tau = \tau_1 = \tau_0 + \tau_{ptb}$, where the perturbation τ_{ptb} varies from 0 to 2.0 with step size equal to 0.1. We use Hamiltonian Monte Carlo (HMC) in order to generate samples from the unnormalized density function. In order to perform LRT, we compute the normalizing constant by numerical integration. We conduct experiments in three different dimensions, i.e., $d = 1, 2,$ and 3. All results are reported as the average of four random experiments. We evaluate the performance of HST and LRT for this hypothesis testing scenario. Results depicted in Figure 7 demonstrate that HST performs competitively with LRT and significantly reduces the computation cost by avoiding the calculation of normalization constants.

C. GAUSS-BERNOULLI RBM

The RBM [2] is a generative graphical model defined on a bi-partite graph of hidden and visible variables. The Gauss-Bernoulli RBM has binary-valued hidden variables $\mathbf{h} \in \{0, 1\}^{d_h}$ and real-valued visible variables $\mathbf{x} \in R^{d_x}$ with joint distribution

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z} \exp\left\{ -\left(\frac{1}{2} \sum_{i=1}^{d_x} \sum_{j=1}^{d_h} \frac{x_i}{\sigma_i} W_{ij} h_j + \sum_{i=1}^{d_x} b_i x_i + \sum_{j=1}^{d_h} c_j h_j - \frac{1}{2} \sum_{i=1}^{d_x} \frac{x_i^2}{\sigma_i^2} \right) \right\},$$

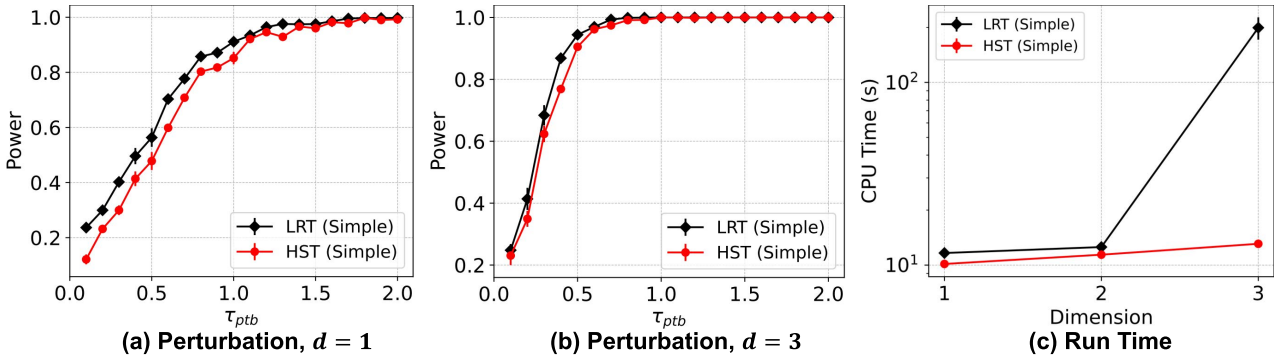


FIGURE 7. Exponential family, (a, b) power comparison between LRT and HST by changing the perturbation τ_{ptb} ($n = 100$), and (c) run time comparison between LRT and HST by varying the dimension of the distribution.

where model parameters $\theta = (\mathbf{W}, \mathbf{b}, \mathbf{c})$ and Z is the normalizing constant. We set $\sigma_i = 1$ for all $i = 1, \ldots, d_x$ in the following experiments. The probability of the visible variable $\mathbf{x}$ written as $p(\mathbf{x}) = \sum_{h\in\{0,1\}^{d_h}} p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z}\exp\{-F_\theta(\mathbf{x})\}$, where $F_\theta(\mathbf{x})$ is the free energy given by

$$F_\theta(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{d_x}(x_i - b_i)^2 - \sum_{j=1}^{d_h}\text{Softplus}\left(\sum_{i=1}^{d_x} W_{ij}x_i + b_j\right).$$

The Softplus function is defined as $\text{Softplus}(t) \triangleq \log(1 + \exp(t))$ with a default scale parameter $\beta = 1$. By Equation (3), the corresponding Hyvärinen score $S_{\text{H}}(\mathbf{X}_n, \theta)$ is given by

$$S_{\text{H}}(\mathbf{X}_n, \theta) = \sum_{n=1}^{n}\sum_{i=1}^{d_x}\left[\frac{1}{2}\left(x_{in} - b_i + \sum_{j=1}^{d_h} W_{ij}\delta_{jn}\right)^2 + \sum_{j=1}^{d_h} W_{ij}^2\delta_{jn}\left(1 - \delta_{jn}\right) - 1\right],$$

where $\delta_{jn} \triangleq \text{Sigmoid}(\sum_{i=1}^{d_x} W_{ij}x_{in} + b_j)$. The Sigmoid function is defined as $\text{Sigmoid}(t) \triangleq (1 + \exp(-t))^{-1}$.

We randomly draw the weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_h}$ from the standard Normal distribution. Here, we let the dimension of visible variables $d_x = 50$ and the dimension of hidden variables $d_h = 40$. The weight matrix of the alternative hypothesis $\mathbf{W}_1$ is constructed by adding a noise term following Normal distribution $\mathcal{N}(0, \sigma_{ptb}^2)$ with different perturbation levels σ_{ptb}. The samples are drawn using Gibbs sampling with 1000 RBM iterations to ensure convergence.

In the first set of experiments, we let σ_{ptb} take values {0, 0.005, 0.007, 0.009, 0.01, 0.011, 0.012, 0.014, 0.015, 0.016, 0.018, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.075, 0.1} with a fixed sample size $n = 100$. In Figure 8, we depict the power of HST for this scenario. Results demonstrate that both HST (simple) and HST (composite)

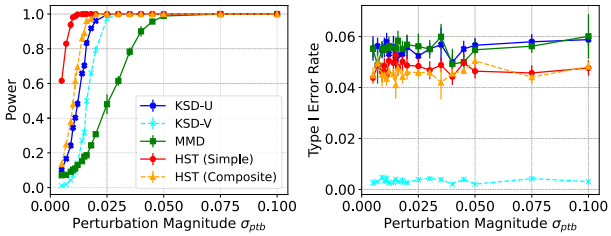


FIGURE 8. Gauss-Bernoulli RBM, power comparison of various tests varying the perturbation level σ_{ptb} ($n = 100$).

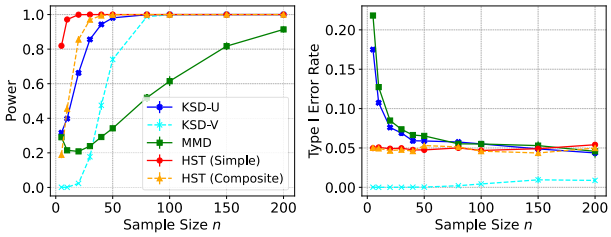


FIGURE 9. Gauss-Bernoulli RBM, Power comparison of various tests varying sample size and the perturbation level $\sigma_{ptb} = 0.03$.

outperform KSD-based tests and MMD two-sample tests for every value of σ_{ptb}.

In the second set of experiments, the sample size takes values in the set {5, 10, 20, 30, 40, 50, 80, 150, 200} but we fix $\sigma_{ptb} = 0.03$. In Figure 9, we compare the performance of various tests for different sample sizes. Results demonstrate that KSD-U, KSD-V, and MMD have unsatisfactory Type I error rates and powers, particularly for small sample sizes. Again, it is found that KSD-U, KSD-V, and the MMD two-sample test are unstable for small sample sizes, while HST shows remarkable robustness and consistency.

In Figure 10, we present the test statistics and ROC curves for the above tests. It is observed again that HST (composite) provides strong discrimination between the null and the alternative hypothesis. Note that the test statistic of HST (simple) is decreasing under the null hypothesis since the aggregate Hyvärinen score of the alternative distribution is subtracted. With the prior knowledge of the alternative hypothesis, HST (simple) performs the best. The results in the

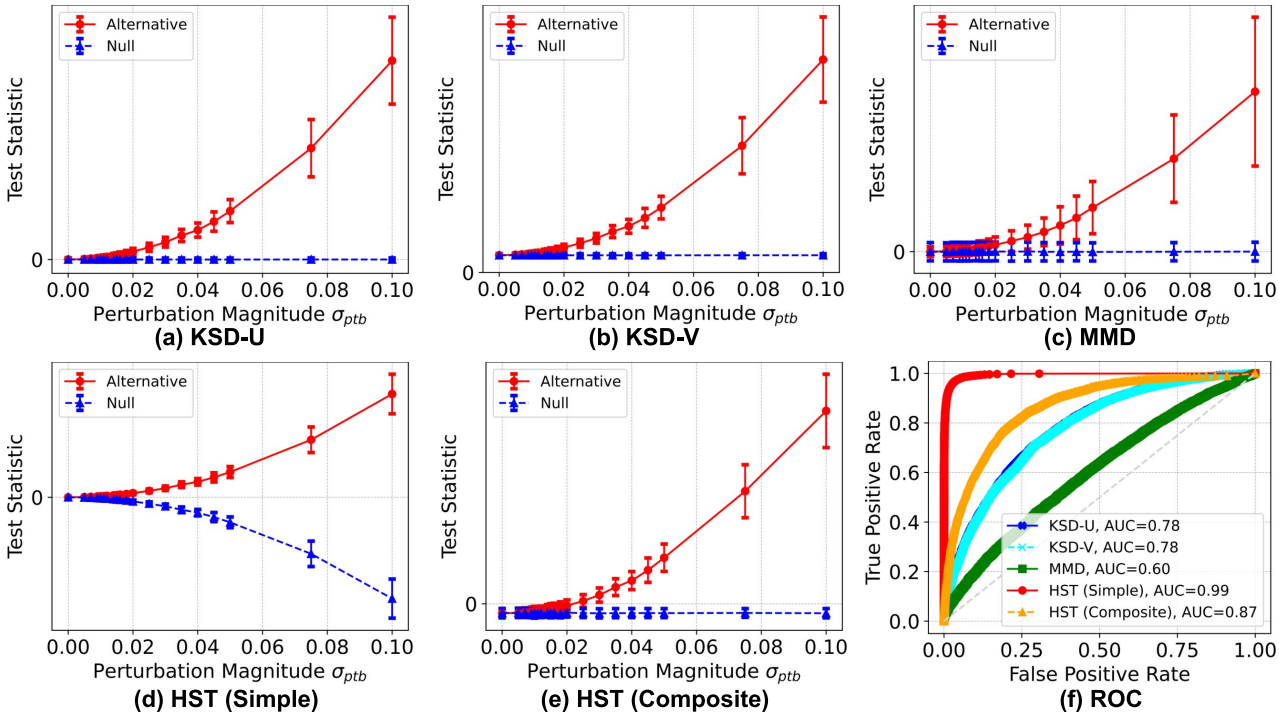


FIGURE 10. (a-e) Test statistics at $n = 100$. (f) Receiver Operating Characteristic (ROC) curves of various tests with $\sigma_{ptb} = 0.01$ and $n = 100$.

ROC curve demonstrate that HST significantly outperforms other baseline tests.

VI. OUT OF DISTRIBUTION (OOD) DETECTION

In this section, we discuss the application of HST to the out-of-distribution (OOD) detection problem. The goal of OOD detection is to determine whether a given input is from the training data distribution (in-distribution examples) or not (out-of-distribution examples). Recent works adapt the LRT heuristically to detect the OOD samples [26], [27]. Particularly, a generative model of the null distribution is learned from the in-distribution samples [3], [37]. Then the likelihood of the test sample is computed to perform LRT for OOD detection, where a low likelihood rejects the in-distribution hypothesis.

Inspired by recent works using LRT for OOD detection [26]–[28], we propose to use the aggregate Hyvärinen score $\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, q_{\theta_0}(\mathbf{x}))$ for OOD detection, where the density function q_{θ_0} is learned from the in-distribution sample and $\mathbf{X}_n$ is the out-of-distribution sample. We reject the in-distribution hypothesis when $-\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, q_{\theta_0}(\mathbf{x}))$ is larger than a threshold C. Similar to the bootstrap HST, the threshold C can be decided by repeating the tests over the in-distribution train data. However, the choice of the critical value of HST depends on the particular OOD detection problem. Below, we will use the ROC curve and AUC to evaluate the performance of the tests.

A. IMAGE DATA

We evaluate the performance of HST on the computer vision benchmark datasets: CIFAR [38], and TinyImageNet, a subset of ImageNet [39]. We treat CIFAR-10 and TinyImageNet datasets as the in-distribution and out-of-distribution data, respectively. From Equation (3), the aggregate Hyvärinen score requires calculation of the gradient of logarithmic density function $\nabla_{\mathbf{x}} \log q_\theta(\mathbf{x})$ and the Laplacian term $\Delta_{\mathbf{x}} \log q_\theta(\mathbf{x})$.

Following [11], we model the gradient of logarithmic density function $\nabla_{\mathbf{x}} \log q_\theta(\mathbf{x})$ with a pre-trained deep generative model on CIFAR10 dataset. To this end, we use the SDE-based deep generative model, Noise Conditional Score Network with variance exploding SDEs (NCSN++) [11] that is pre-trained on CIFAR-10 to perform OOD detection. The model architecture includes four BigGAN-type [40] residual blocks per image resolution. We perform OOD detection on the test data of CIFAR-10 (in-distribution samples) and the test data of TinyImageNet (out-of-distribution samples), where we randomly crop image patches of size 32×32 to match the shape of CIFAR-10.

Subsequently, we utilize the Hutchinson's trick [41] to reduce the computation complexity of the Laplacian term $\Delta_{\mathbf{x}} \log q_\theta(\mathbf{x})$ in high dimensions. The Hutchinson method obtains the unbiased estimate of the Laplacian term by Monte Carlo sampling:

$$\begin{aligned}\Delta_{\mathbf{x}} \log q_\theta(\mathbf{x}) &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^T \cdot \nabla_{\mathbf{x}} f(\mathbf{x}, \theta) \cdot \boldsymbol{\epsilon}\right] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^T \cdot \nabla_{\mathbf{x}}\left(\boldsymbol{\epsilon}^T f(\mathbf{x}, \theta)\right)\right],\end{aligned}$$

where $f(\mathbf{x}, \theta) \triangleq \nabla_{\mathbf{x}} \log q_\theta(\mathbf{x})$, and random projections $\boldsymbol{\epsilon}$ are Normally distributed. Our heuristic results indicate that

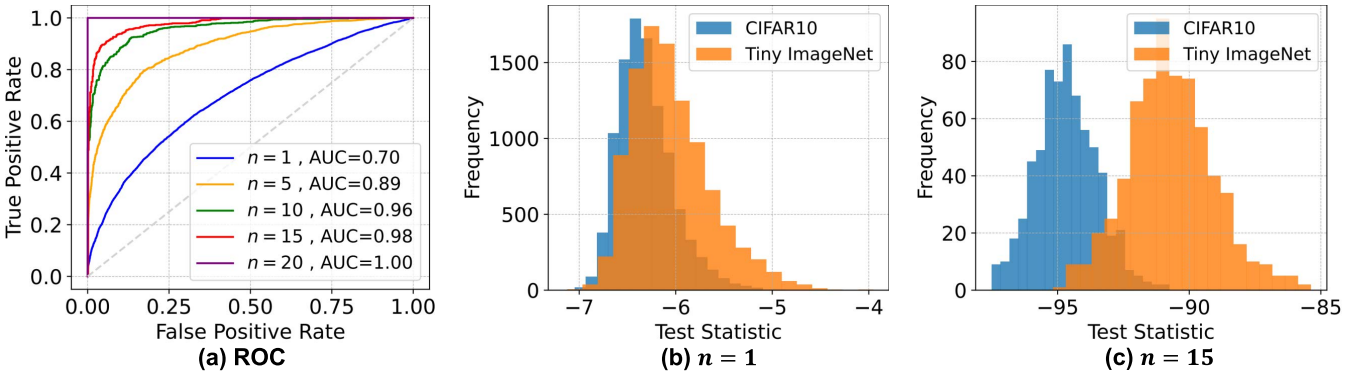


FIGURE 11. (a) ROC curves and (b, c) histograms of test statistics of HST for OOD Detection on CIFAR10 (in-distribution) and Tiny ImageNet datasets (out-distribution).

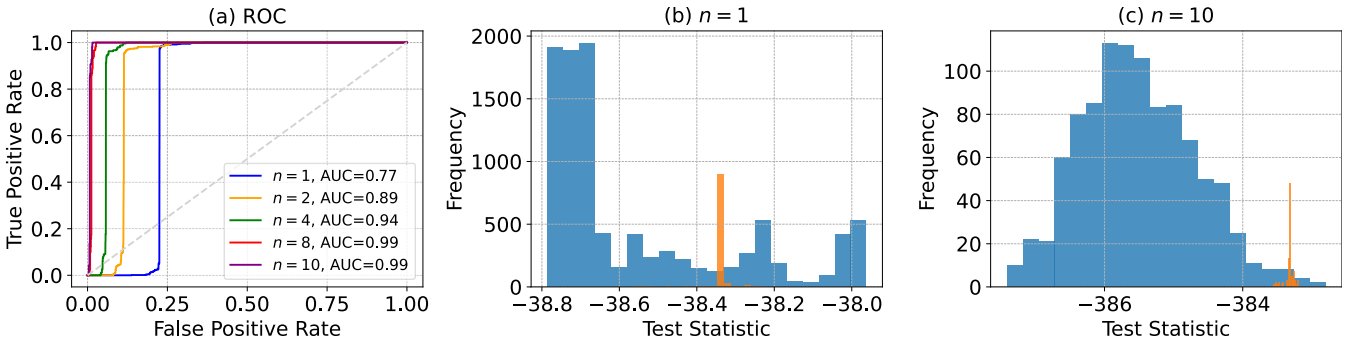


FIGURE 12. (a) ROC curves and (b, c) histograms of test statistics of the 'back' attack and 'normal' network of HST on KDD Cup'99 dataset.

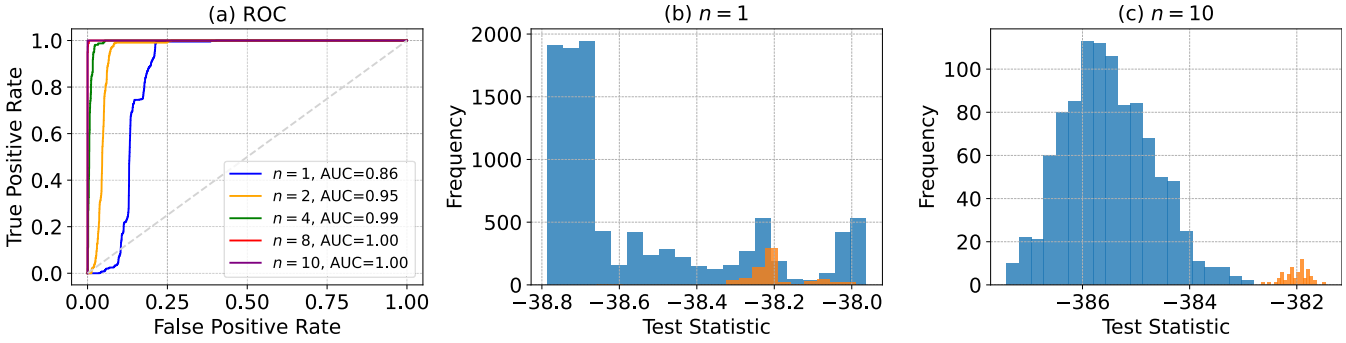


FIGURE 13. (a) ROC curves and (b, c) histograms of test statistics of the 'ipsweep' attack and 'normal' network of HST on KDD Cup'99 dataset.

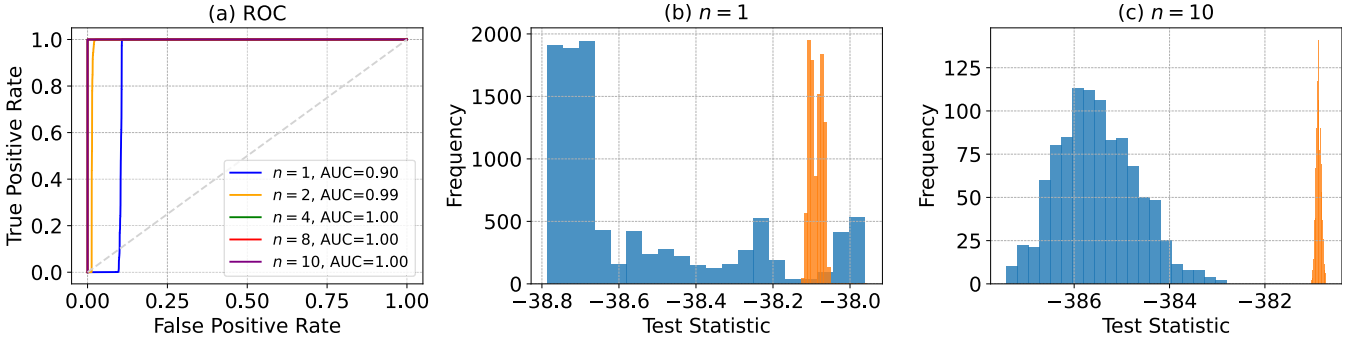


FIGURE 14. (a) ROC curves and (b, c) histograms of test statistics of the 'neptune' attack and 'normal' network of HST on KDD Cup'99 dataset.

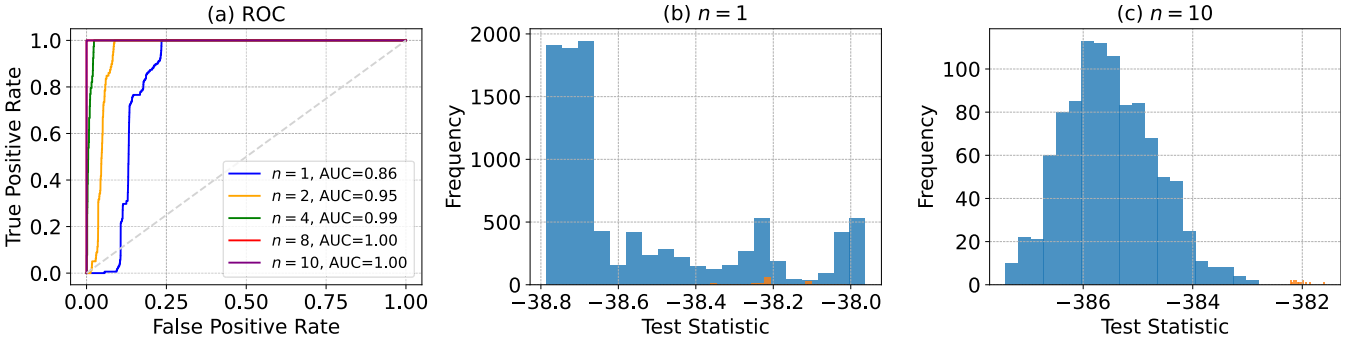


FIGURE 15. (a) ROC curves and (b, c) histograms of test statistics of the 'nmap' attack and 'normal' network of HST on KDD Cup'99 dataset.

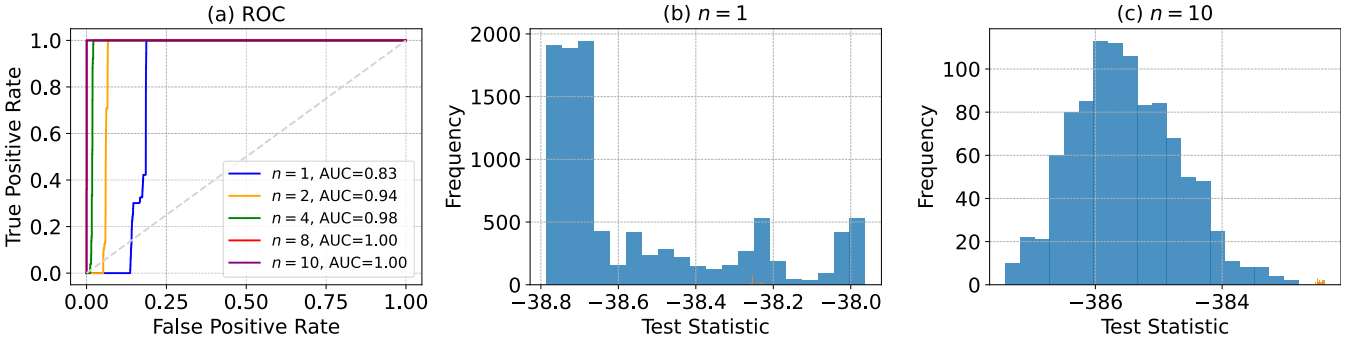


FIGURE 16. (a) ROC curves and (b, c) histograms of test statistics of the 'pod' attack and 'normal' network of HST on KDD Cup'99 dataset.

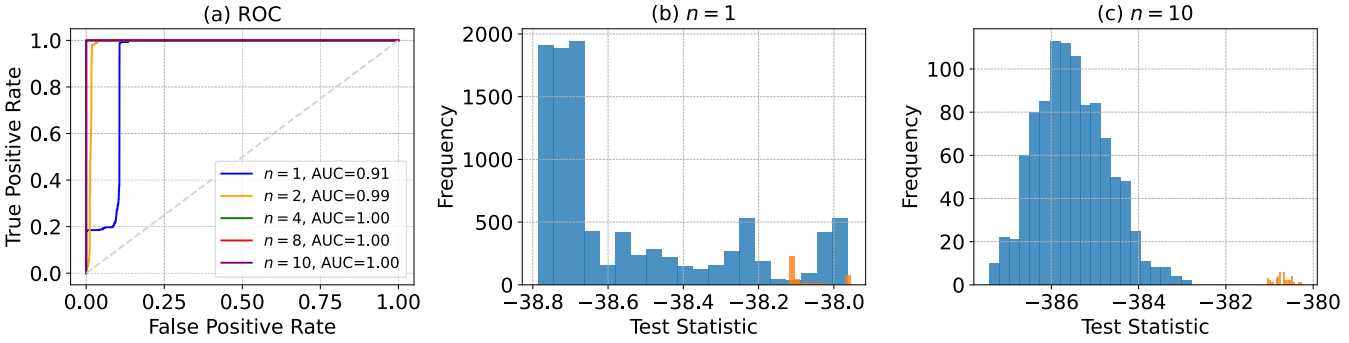


FIGURE 17. (a) ROC curves and (b, c) histograms of test statistics of the 'portsweep' attack and 'normal' network of HST on KDD Cup'99 dataset.

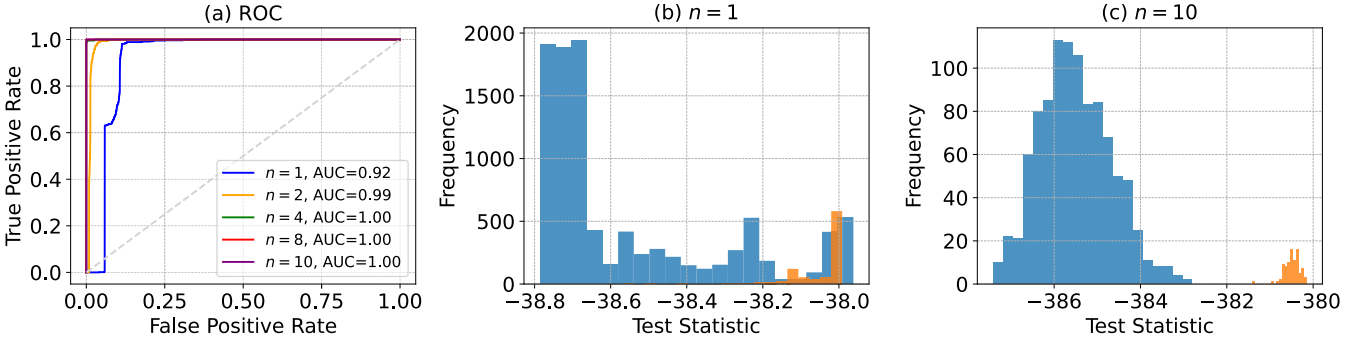


FIGURE 18. (a) ROC curves and (b, c) histograms of test statistics of the 'satan' attack and 'normal' network of HST on KDD Cup'99 dataset.

a single particle sampling may be good enough for the approximation to work in our context.

We evaluate the performance of HST for OOD detection by varying the test sample size. In Figure 11, we present the ROC curve and histograms of $-\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, q_{\theta_0}(\mathbf{x}))$ over different sample sizes. The results demonstrate that the performance of our method significantly improves when the sample size is more than $n = 10$. Note that it is difficult to detect "out-

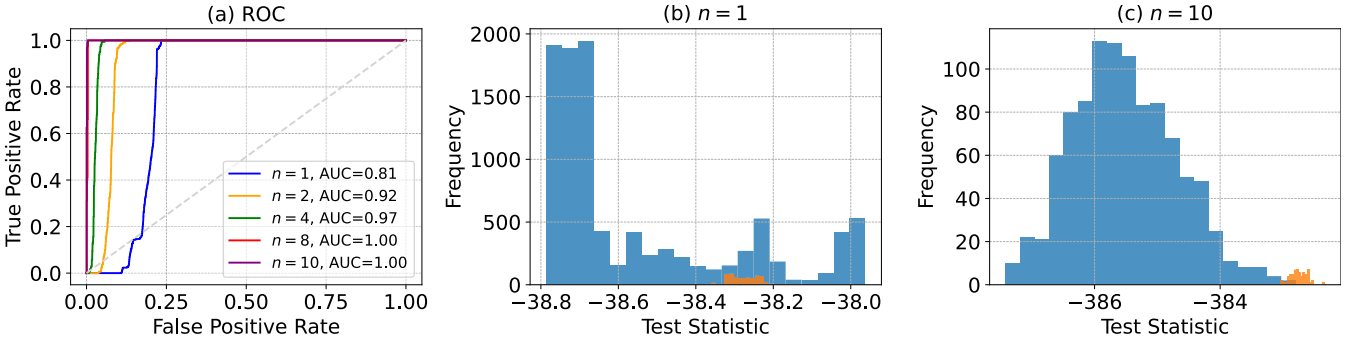


FIGURE 19. (a) ROC curves and (b, c) histograms of test statistics of the 'smurf' attack and 'normal' network of HST on KDD Cup'99 dataset.

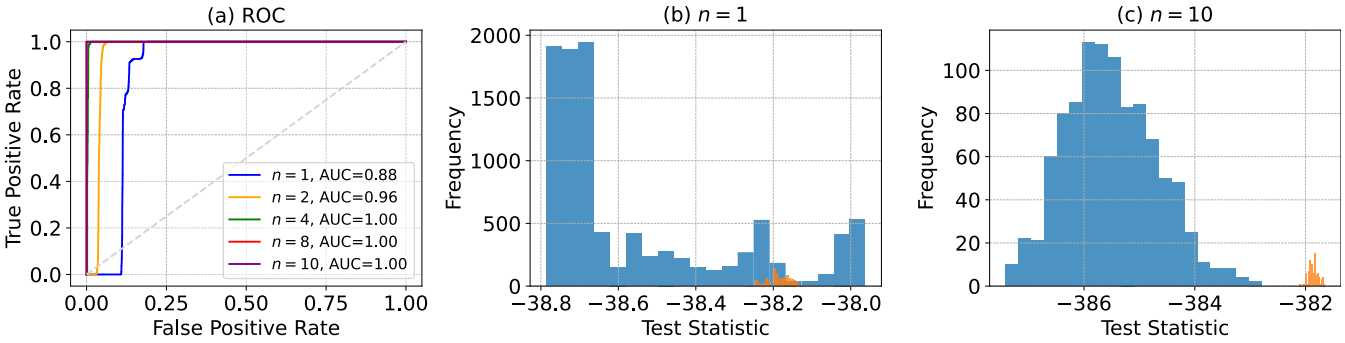


FIGURE 20. (a) ROC curves and (b, c) histograms of test statistics of the 'teardrop' attack and 'normal' network of HST on KDD Cup'99 dataset.

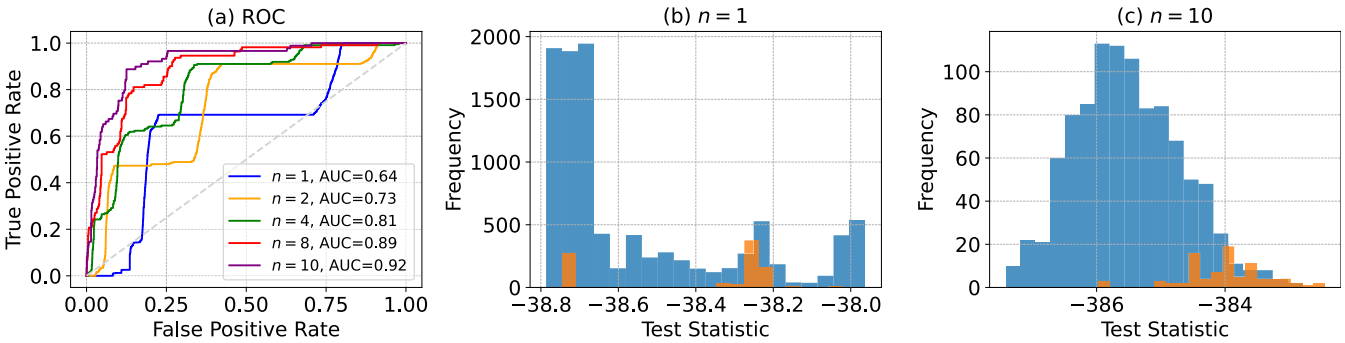


FIGURE 21. (a) ROC curves and (b, c) histograms of test statistics of the 'warezclient' attack and 'normal' network of HST on KDD Cup'99 dataset.

of-distribution" by a single data point, since a single out-of-distribution data point provides little information (as in this case, the in and out distributions are close to each other).

B. NETWORK INTRUSION DATA

We next provide experimental results on the KDD Cup'99 dataset [42]. The dataset contains various kinds of adversarial network attacks simulated for building a network intrusion detector. We treat the 'normal' network and adversarial network attacks as the in-distribution and out-of-distribution, respectively. We train a Gauss-Bernoulli RBM with in-distribution samples to detect out-of-distribution samples from adversarial network attacks.

In Table 1, we evaluate our method with AUC. The results demonstrate that our method can detect adversarial network attacks even with a single out of distribution data point.

TABLE 1. Area under the curve of receiver operating characteristics (AUC) for our test to detect malicious network attack for various values of sample size *n*.

n (size)	back	ipsweep	neptune	nmap	pod
1	0.785	0.869	0.896	0.835	0.802
2	0.895	0.961	0.986	0.946	0.933
4	0.937	0.997	1.000	0.993	0.983
8	0.991	1.000	1.000	1.000	1.000
10	0.999	1.000	1.000	1.000	1.000
n (size)	portsweep	satan	smurf	teardrop	warezclient
1	0.921	0.928	0.818	0.882	0.645
2	0.979	0.983	0.942	0.963	0.731
4	1.000	1.000	0.972	0.996	0.803
8	1.000	1.000	1.000	1.000	0.889
10	1.000	1.000	1.000	1.000	0.928

Naturally, our method performance significantly improves when more out of distribution samples are available.

From Figure 12 to 21, we depict the ROC curves and the histograms of $-\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, q_{\theta_0}(\mathbf{x}))$ for detecting the malicious network attack. The results demonstrate that our proposed method can effectively identify various kinds of malicious network intrusion with a small number of OOD samples.

VII. CONCLUSION

We extend the application of Hyvärinen score [6], [7], commonly used in parameter estimation for unnormalized parametric models, to hypothesis testing and propose a new test statistic as a surrogate of LRT for unnormalized statistical models. We analyze the asymptotic behavior of the test statistic and provide theoretical results justifying our approach. We also provide experimental results demonstrating significant performance gains, reduction in computational complexity, robustness, and stability for small sample sizes for our proposed approach. Following [2], [11], [43], we model the Hyvärinen score function with respect to data with generative models. This enables applications of our test to high dimensional settings in various scenarios of interest, such as in OOD detection.

.

APPENDIX A PROOFS

In this section, we provide details of our technical results.

A. PROOF OF THEOREM 1

Theorem 1 can be proved using the asymptotic normality of standard M-estimators. We provide detailed proof for technical completeness. We first introduce some regularity conditions.

Assumption 1: The family $q_\theta(\mathbf{x})$ is identifiable, i.e., that $\theta \neq \theta^\star \to q_\theta \neq q_{\theta^\star}$.

Assumption 2: For all $\mathbf{x} \in \mathcal{X}$, $s_{\mathrm{H}}(q_\theta, \mathbf{x})$ is continuous in $\theta \in \Theta$ for Θ compact.

Assumption 3: There exists a function ξ_1: $\mathbf{x} \to \xi_1(\mathbf{x})$ such that for any $\theta \in \Theta$, $|s_{\mathrm{H}}(q_\theta, \mathbf{x})| \leq \xi_1(\mathbf{x})$ and $\mathbb{E}_\star[\xi_1(\mathbf{x})] < \infty$.

Assumption 4: For any $\theta \in \Theta$, $q_\theta(\mathbf{x})\nabla_{\mathbf{x}} \log q_\theta(\mathbf{x}) \to 0$ as $\|\mathbf{x}\|_2 \to \infty$.

We note that the Assumption 4 may be violated in degenerate cases when the data sit on a low-dimensional manifold inside that high-dimensional space. In this case, the score function will not be well-defined in the ambient space. This problem, known as the manifold hypothesis, often occurs in deep generative modeling on image datasets [44]. Nevertheless, our work extends even to this degenerate scenario by considering sliced score matching [45] or the Riemannian score matching [46]. We will not pursue these degenerate cases here.

Lemma 1 (Consistency of score matching estimates, Theorem 2 of [6]): Under Assumptions 1-4 and assuming that $q_\theta(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}$ and $\theta \in \Theta$, the score matching estimate (defined in Equation (5)) is consistent, i.e., $\hat{\theta}_{\mathtt{sm}} \xrightarrow{n\to\infty}_{\mathcal{L}} \theta^\star$.

Assumption 5: $\theta^\star$ is an interior point of the parameter space Θ.

Assumption 6: For all $\mathbf{x} \in \mathcal{X}$, $s_{\mathrm{H}}(q_\theta, \mathbf{x})$ is twice continuously differentiable in the interior of Θ.

Assumption 7: The expected values

$$\mathbb{E}_\star\left[\nabla_\theta s_{\mathrm{H}}(q_\theta, \mathbf{x})\nabla_\theta^\top s_{\mathrm{H}}(q_\theta, \mathbf{x})\,|_{\theta=\theta_0}\right], \text{ and}$$
$$\mathbb{E}_\star\left[\nabla_\theta^2 s_{\mathrm{H}}(q_\theta, \mathbf{x})\,|_{\theta=\theta_0}\right]$$

exist and are non-singular. There exists a function $\xi_2 : \mathbf{x} \mapsto \xi_2(\mathbf{x})$ such that $\left|\frac{\partial^2 s_{\mathrm{H}}(q_\theta, \mathbf{x})}{\partial\theta_i\partial\theta_j}\right| \leq \xi_2(\mathbf{x})$ for all $1 \leq i, j \leq k$, and $\mathbb{E}_\star[\xi_2(\mathbf{x})] < \infty$.

Lemma 2 (Asymptotic Normality (Theorem 5.23 of [47])): Under the assumptions of Lemma 1 and Assumptions 5-7, we have $\sqrt{n}(\hat{\theta}_{\mathtt{sm}} - \theta^\star) \xrightarrow{n\to\infty}_{\mathcal{L}} \mathcal{N}(\mathbf{0}_r, \mathbf{V})$, where $\mathbf{V} = \mathbf{H}^{-1}\mathbf{K}\mathbf{H}^{-1}$, $\mathbf{H} \triangleq \mathbb{E}_\star\left[\nabla_\theta^2 s_{\mathrm{H}}(q_\theta, \mathbf{x})\,|_{\theta=\theta_0}\right]$ and $\mathbf{K} \triangleq \mathbb{E}_\star\left[\nabla_\theta s_{\mathrm{H}}(q_\theta, \mathbf{x})\nabla_\theta^\top s_{\mathrm{H}}(q_\theta, \mathbf{x})\,|_{\theta=\theta_0}\right]$.

Under Assumption 6, we have

$$\begin{aligned}
&\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_0) \\
&\quad= \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathtt{sm}}) + \mathcal{S}'_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathtt{sm}})(\hat{\theta}_{\mathtt{sm}} - \theta_0) \\
&\qquad+ \frac{1}{2}(\hat{\theta}_{\mathtt{sm}} - \theta_0)^T \mathcal{S}''_{\mathrm{H}}(\mathbf{X}_n, \tilde{\theta})(\hat{\theta}_{\mathtt{sm}} - \theta_0) \\
&\quad= \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathtt{sm}}) + \frac{1}{2}(\hat{\theta}_{\mathtt{sm}} - \theta_0)^T \mathcal{S}''_{\mathrm{H}}(\mathbf{X}_n, \tilde{\theta})(\hat{\theta}_{\mathtt{sm}} - \theta_0)
\end{aligned} \tag{13}$$

where $\tilde{\theta}$ is on the line segment of $\hat{\theta}_{\mathtt{sm}}$ and θ_0. Thus,

$$\begin{aligned}
&2\left(\mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \theta_0) - \mathcal{S}_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathtt{sm}})\right) \\
&\quad= \sqrt{n}(\hat{\theta}_{\mathtt{sm}} - \theta_0)^T \cdot \frac{1}{n}\mathcal{S}''_{\mathrm{H}}(\mathbf{X}_n, \tilde{\theta}) \cdot \sqrt{n}(\hat{\theta}_{\mathtt{sm}} - \theta_0).
\end{aligned}$$

Following the assumptions, we have the uniform law of large numbers $\frac{1}{n}\mathcal{S}''_{\mathrm{H}}(\mathbf{X}_n, \hat{\theta}_{\mathtt{sm}}) \xrightarrow{n\to\infty}_{\mathcal{L}} \mathbf{H}$. Under the null hypothesis, we have from Lemma 2 that $\sqrt{n}(\hat{\theta}_{\mathtt{sm}} - \theta_0) \xrightarrow{n\to\infty}_{\mathcal{L}} \mathcal{N}(\mathbf{0}_r, \mathbf{V})$. Then, we apply Slutsky's Theorem [48] to complete the proof.

B. DISCUSSION OF REMARK 1

In this section, we briefly outline the derivation of Equations (10) and (11). We will use the following technical result. For a real-valued function g and a PDF p defined on $\mathbb{R}$, if $g(x)p(x) \to 0$ as $\|x\|_2 \to \infty$, we have

$$\begin{aligned}
\int p(x)\frac{\partial \log p(x)}{\partial x}g(x)dx &= \int \frac{\partial p(x)}{\partial x}g(x)dx \\
&= -\int \frac{\partial g(x)}{\partial x}p(x)dx,
\end{aligned}$$

This holds for $\mathbf{x} \in \mathbb{R}^d$, namely

$$\int p(\mathbf{x})\frac{\partial \log p(\mathbf{x})}{\partial x_k}g(\mathbf{x})d\mathbf{x} = -\int \frac{\partial g(\mathbf{x})}{\partial x_k}p(\mathbf{x})d\mathbf{x}, \tag{14}$$

for all $k = 1, \ldots, d$.

Next, let $g_k(\theta, \mathbf{x}) \triangleq \partial_{x_k} \log q_\theta(\mathbf{x})$, and $g'_k(\theta, \mathbf{x}) \triangleq \partial^2_{x_k} \log q_\theta(\mathbf{x})$, where x_k is the kth component of $\mathbf{x}$. We write $g_k = g_k(\theta, \mathbf{x})$, and $g'_k = g'_k(\theta, \mathbf{x})$ in short. By the definition of

$s_{\mathrm{H}}(q_\theta, \mathbf{x})$, we can write

$$\mathbb{E}_\star \left(\partial_\theta s_{\mathrm{H}}(q_\theta, \mathbf{x})\right)^2 = \sum_{k=1}^{d} \mathbb{E}_\star \left((\partial_\theta g_k')^2 + (\partial_\theta g_k)^2 g_k^2 + 2(\partial_\theta g_k)(\partial_\theta g_k') g_k \right). \tag{15}$$

By Equation (14), for each $k = 1, \ldots, d$, we have

$$\int \left((\partial_\theta g_k)^2 g_k^2 \right) q_\theta(\mathbf{x}) d\mathbf{x} = -\int \left(2(\partial_\theta g_k)(\partial_\theta g_k') g_k + (\partial_\theta g_k)^2 g_k' \right) q_\theta(\mathbf{x}) d\mathbf{x}.$$

Plugging it back to Equation (15), we have

$$\mathbb{E}_\star \left(\partial_\theta s_{\mathrm{H}}(q_\theta, \mathbf{x})\right)^2 = \sum_{i=1}^{d} \int \left((\partial_\theta g_k')^2 - (\partial_\theta g_k)^2 g_k' \right) q_\theta(\mathbf{x}) d\mathbf{x}.$$

Thus, Equation (11) holds. Similarly, we calculate

$$\begin{aligned} &\mathbb{E}_\star \partial_\theta^2 s_{\mathrm{H}}(q_\theta, \mathbf{x}) \\ &= \sum_{k=1}^{d} \int \left(\partial_\theta(\partial_\theta g_k \cdot g_k) + \partial_\theta(\partial_\theta g_k') \right) q_\theta(\mathbf{x}) d\mathbf{x} \\ &= \sum_{k=1}^{d} \int \left((\partial_\theta^2 g_k) g_k + (\partial_\theta g_k)^2 + \partial_\theta^2 g_k' \right) q_\theta(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{16}$$

Again, by applying Equation (14), for each $k = 1, \ldots, d$, we have

$$\int \partial_\theta^2 g_k \cdot g_k \cdot q_\theta(\mathbf{x}) d\mathbf{x} = -\int \partial_\theta^2 g_k' \cdot q_\theta(\mathbf{x}) d\mathbf{x}.$$

Plugging it back to Equation (16), we have Equation (10) hold. The multi-dimensional case can be similarly derived by considering each component of $\mathbf{K}$ and $\mathbf{H}$.

C. DICUSSION OF REMARK 2

In this section, we provide a detailed derivation of the test power T_{LRT} and T_{HST}. By direct calculations, we have

$$T_{\mathrm{LRT}}(\mathbf{X}_n) = -\sum_{i=1}^{n} \mu_1^T \Sigma^{-1} (\mathbf{x}_i - \frac{\mu_1}{2}),$$

$$T_{\mathrm{HST}}(\mathbf{X}_n) = -\sum_{i=1}^{n} \mu_1^T \Sigma^{-2} (\mathbf{x}_i - \frac{\mu_1}{2}).$$

Recall that the HST rejects the null hypothesis when the test statistic T_{HST} is too large. For a pre-specified threshold $\eta \in \mathbb{R}$, the rejection region $R_{\mathrm{HST},\eta}$ of the HST is defined by

$$R_{\mathrm{HST},\eta} \triangleq \left\{ \mathbf{X}_n \mid \sum_{i=1}^{n} \mu_1^T \Sigma^{-2} (\mathbf{x}_i - \frac{\mu_1}{2}) > \eta \right\}. \tag{17}$$

Let $\Sigma = U \Lambda U^T$, where U is a $d \times d$ matrix whose i-th column is an eigenvector and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues in non-increasing order. Let $\mathbf{x}_i = U \Lambda^{1/2} \mathbf{z}_i + \mu$ with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and by Equation (17). Then power function of HST is given by

$$\beta(\mu) = \mathbb{P}_{\mathbf{z}} \left\{ \mu_1^T U \Lambda^{-3/2} \sum_{i=1}^{n} \mathbf{z}_i > \eta - \mu_1^T \Sigma^{-2} (\mu - \frac{\mu_1}{2}) \right\}.$$

Under the null hypothesis $\mu = \mathbf{0}_d$, the Type I error probability is given by

$$\beta_0 = \mathbb{P}_{\mathbf{z}} \left\{ \mu_1^T U \Lambda^{-3/2} \sum_{i=1}^{n} \mathbf{z}_i > \eta + \frac{\mu_1^T \Sigma^{-2} \mu_1}{2} \right\}.$$

Under the alternative $\mu = \mu_1$, we have the power

$$\beta_{\mathrm{HST}} = \mathbb{P}_{\mathbf{z}} \left\{ \mu_1^T U \Lambda^{-3/2} \sum_{i=1}^{n} \mathbf{z}_i > \eta - \frac{\mu_1^T \Sigma^{-2} \mu_1}{2} \right\}. \tag{18}$$

Since $\mathbf{z}_i$ are IID Normal, $\mu_1^T U \Lambda^{-3/2} \sum_{i=1}^{n} \mathbf{z}_i$ is distributed according to $\mathcal{N}(0, n \mu_1^T \Sigma^{-3} \mu_1)$. Hence, Type I error probability is given by

$$\beta_0 = 1 - \Phi \left(\frac{\eta}{\sqrt{n \mu_1^T \Sigma^{-3} \mu_1}} + \frac{\mu_1^T \Sigma^{-2} \mu_1}{2 \sqrt{n \mu_1^T \Sigma^{-3} \mu_1}} \right). \tag{19}$$

Given a significance level α, the threshold η can be derived as

$$\eta_\alpha = \left(q_{1-\alpha} - \frac{\mu_1^T \Sigma^{-2} \mu_1}{2 \sqrt{n \mu_1^T \Sigma^{-3} \mu_1}} \right) \cdot \sqrt{n \mu_1^T \Sigma^{-3} \mu_1}, \tag{20}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ quantile of standard Normal distribution. Combining Equation (18) and (20), we have

$$\beta_{\mathrm{HST}} = \Phi \left(q_{1-\alpha} - \frac{\mu_1^T \Sigma^{-2} \mu_1}{\sqrt{N \mu_1^T \Sigma^{-3} \mu_1}} \right). \tag{21}$$

Similarly, we can derive the power for LRT:

$$\beta_{\mathrm{LRT}} = \Phi \left(q_{1-\alpha} - \frac{\mu_1^T \Sigma^{-1} \mu_1}{\sqrt{N \mu_1^T \Sigma^{-1} \mu_1}} \right). \tag{22}$$

Next we observe that

$$\frac{\mu_1^T \Sigma^{-2} \mu_1}{\sqrt{n \mu_1^T \Sigma^{-3} \mu_1}} \geq \frac{\sigma_{\min}}{\sigma_{\max}} \frac{\mu_1^T \Sigma^{-1} \mu_1}{\sqrt{n \mu_1^T \Sigma^{-1} \mu_1}},$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are respectively the maximum and minimum eigenvalues of the covariance matrix Σ. Let $\tilde{\mu}_1 = \mu_1^T \Sigma^{-1} \mu_1$ and $\Delta_\sigma = (\sigma_{\max} - \sigma_{\min})/\sigma_{\max}$, then β_{HST} is related to β_{LRT} by

$$\begin{aligned} \beta_{\mathrm{HST}} &\leq \Phi \left(q_{1-\alpha} - \frac{\sigma_{\min}}{\sigma_{\max}} \sqrt{\frac{\tilde{\mu}_1}{N}} \right) \\ &= \Phi \left(q_{1-\alpha} - \sqrt{\frac{\tilde{\mu}_1}{n}} + \Delta_\sigma \sqrt{\frac{\tilde{\mu}_1}{n}} \right) \\ &= \beta_{\mathrm{LRT}} + \mathrm{UB(I)}. \end{aligned}$$

Let

$$s = q_{1-\alpha} - \sqrt{\frac{t}{n}}, t = q_{1-\alpha} - \sqrt{\frac{t}{n}} + \Delta_\sigma \sqrt{\frac{t}{n}}.$$

Then,

$$\begin{aligned} \text{UB(I)} &= \frac{1}{\sqrt{2\pi}} \int_s^t \frac{y}{s} \exp\left\{-\frac{y^2}{2}\right\} dy \\ &\leq \frac{1}{\sqrt{2\pi}s} \left(\exp\left\{-\frac{s^2}{2}\right\} - \exp\left\{-\frac{t^2}{2}\right\}\right), \end{aligned}$$

where the second inequality is by the convexity of the exponential function.

REFERENCES

[1] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques—Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[2] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting Structured Data*, vol. 1. 2006.

[3] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *J. Mach. Learn. Res.*, vol. 22, pp. 57:1–57:64, Nov. 2021.

[4] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002, doi: 10.1162/089976602760128018.

[5] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing Statistical Hypotheses*, vol. 3. New York, NY, USA: Springer, 2005.

[6] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, no. 4, pp. 695–709, 2005.

[7] A. Hyvärinen, "Some extensions of score matching," *Comput. Stat. Data Anal.*, vol. 51, no. 5, pp. 2499–2512, Feb. 2007, doi: 10.1016/j.csda.2006.09.003.

[8] S. Lyu, "Interpretation and generalization of score matching," in *Proc. 25th Conf. Uncertainty Artif. Intell. (UAI)*. Arlington, VA, USA: AUAI Press, 2009, pp. 359–366.

[9] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar, "Density estimation in infinite dimensional exponential families," *J. Mach. Learn. Res.*, vol. 18, no. 57, pp. 1–59, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-011.html

[10] K. V. Mardia, J. T. Kent, and A. K. Laha, "Score matching estimators for directional distributions," 2016, *arXiv:1604.08470*.

[11] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2020, *arXiv:2011.13456*.

[12] A. Vahdat, K. Kreis, and J. Kautz, "Score-based generative modeling in latent space," in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 11287–11302.

[13] S. Shao, P. E. Jacob, J. Ding, and V. Tarokh, "Bayesian model comparison with the Hyvärinen score: Computation and consistency," *J. Amer. Stat. Assoc.*, vol. 114, no. 528, pp. 1826–1837, 2019.

[14] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 276–284.

[15] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[16] K. Chwialkowski, H. Strathmann, and A. Gretton, "A kernel test of goodness of fit," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 6, Feb. 2016, pp. 3854–3867.

[17] J. H. Huggins and L. Mackey, "Random feature stein discrepancies," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2018, pp. 1903–1913.

[18] O. Key, T. Fernandez, A. Gretton, and F.-X. Briol, "Composite goodness-of-fit tests with kernels," 2021, *arXiv:2111.10275*.

[19] W. Gong, Y. Li, and J. M. Hernández-Lobato, "Sliced kernelized Stein discrepancy," 2020, *arXiv:2006.16531*.

[20] M. A. Arcones and E. Gine, "On the bootstrap of *U* and *V* statistics," *Ann. Statist.*, vol. 20, no. 2, pp. 655–674, 1992, doi: 10.1214/aos/1176348650.

[21] M. Huskova and P. Janssen, "Consistency of the generalized bootstrap for degenerate *U*-statistics," *Ann. Statist.*, vol. 21, pp. 1811–1823, Dec. 1993.

[22] K. Chwialkowski, D. Sejdinovic, and A. Gretton, "A wild bootstrap for degenerate kernel tests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, Aug. 2014, pp. 1–9.

[23] A. Gretton, B. Sriperumbudur, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, and K. Fukumizu, "Optimal kernel choice for large-scale two-sample tests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2012, pp. 1205–1213. [Online]. Available: https://www.gatsby.ucl.ac.uk/

[24] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," 2021, *arXiv:2110.11334*.

[25] M. Kirchler, S. Khorasani, M. Kloft, and C. Lippert, "Two-sample testing using deep learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 1387–1398.

[26] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[27] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood regret: An out-of-distribution detection score for variational auto-encoder," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 20685–20696.

[28] K. Kim, J. Shin, and H. Kim, "Locally most powerful Bayesian test for out-of-distribution detection using deep generative models," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 14913–14924. [Online]. Available: https://openreview.net/forum?id=-nLW4nhdkO

[29] A. Mahmood, J. Oliva, and M. Styner, "Multiscale score matching for out-of-distribution detection," 2020, *arXiv:2010.13132*.

[30] S. M. Kulinski, S. Bagchi, and D. I. Inouye, "Feature shift detection: Localizing which features have shifted via conditional distribution tests," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 19523–19533.

[31] J. Ding, R. Calderbank, and V. Tarokh, "Gradient information for representation and modeling," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2396–2405.

[32] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, Nov. 2018.

[33] K. Elkhalil, A. Hasan, J. Ding, S. Farsiu, and V. Tarokh, "Fisher auto-encoders," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, vol. 130, 2021, pp. 352–360.

[34] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, no. 1, pp. 60–62, Mar. 1938.

[35] P. J. Bickel and J.-J. Ren, "The bootstrap in hypothesis testing," *Lect. Notes-Monograph Ser.*, vol. 36, pp. 91–112, Jan. 2001. [Online]. Available: http://www.jstor.org/stable/4356107

[36] M. Yu, M. Kolar, and V. Gupta, "Statistical inference for pairwise graphical models using score matching," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[37] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, Jan. 2017, doi: 10.1080/01621459.2017.1285773.

[38] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," Toronto, ON, Canada, 2009.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*.

[41] M. F. Hutchinson, "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines," *Commun. Statist. Simul. Comput.*, vol. 18, no. 3, pp. 1059–1076, Jan. 1989, doi: 10.1080/03610918908812806.

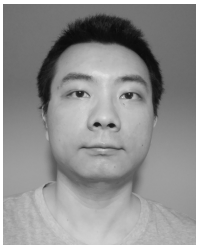
[42] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," vol. 2, Feb. 2000, pp. 12–26, doi: 10.1109/DISCEX.2000.821506.

[43] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–13.

[44] C. Fefferman, S. Mitter, and H. Narayanan, "Testing the manifold hypothesis," *J. Amer. Math. Soc.*, vol. 29, no. 4, pp. 983–1049, Feb. 2016.

[45] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Proc. Uncertainty Artif. Intell.*, 2020, pp. 574–584.
[46] V. De Bortoli, E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet, "Riemannian score-based generative modeling," 2022, *arXiv:2202.02763*.
[47] A. W. Van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[48] E. Slutsky, "Über stochastische asymptoten und grenzwerte," *Metron*, vol. 5, no. 3, pp. 3–89 1925.

SUYA WU received the B.S. degree in mathematics and statistics from Shandong University, Jinan, China, in 2017, and the M.S. degree in statistics from the University of Minnesota, Twin Cities, Minneapolis, USA, in 2019. She is currently pursuing the Ph.D. degree in electrical and computer engineering with Duke University, Durham, USA. Her research interests include statistical methods, including hypothesis testing and probabilistic generative modeling.



ENMAO DIAO (Member, IEEE) was born in Chengdu, Sichuan, China, in 1994. He received the B.S. degree in computer science and electrical engineering from the Georgia Institute of Technology, GA, USA, in 2016, and the M.S. degree in electrical engineering from Harvard University, Cambridge, USA, in 2018. He is currently pursuing the Ph.D. degree in electrical engineering with Duke University, Durham, GA, USA. His research interests include distributed machine learning, data and model compression, and audio signal processing.

KHALIL ELKHALIL received the M.S. and Ph.D. degrees in electrical engineering from the King Abdullah University of Science and Technology (KAUST), in 2015 and 2019, respectively. From September 2019 to May 2022, he was a Postdoctoral Associate with the Department of Electrical and Computer Engineering, Duke University. Since May 2022, he has been a Research Scientist at Amazon, Bellevue. His research interests include high dimensional statistics, random matrix theory, and deep generative modeling.

JIE DING (Member, IEEE) received the B.S. degree from Tsinghua University, Beijing, in 2012, and the Ph.D. degree in engineering sciences from Harvard University, Cambridge, in 2017. He joined as a Faculty Member of the University of Minnesota, Twin Cities, in 2018, where he has been an Assistant Professor with the School of Statistics and a Graduate Faculty Member appointment at the Department of Electrical and Computer Engineering. His research interests include the foundations of machine learning, statistics, and signal processing.

VAHID TAROKH (Fellow, IEEE) worked at the AT&T Laboratories-Research, until 2000. From 2000 to 2002, he was an Associate Professor with the Massachusetts Institute of Technology (MIT). In 2002, he joined Harvard University as a Hammond Vinton Hayes Senior Fellow of Electrical Engineering and a Perkins Professor in applied mathematics. He joined Duke University, in January 2018, as a Rhodes Family Professor in electrical and computer engineering, computer Science, and mathematics and a Bass Connections Endowed Professor. He was also a Gordon Moore Distinguished Research Fellow at Caltech, in 2018.