Score-based Quickest Change Detection for Unnormalized Models

Suya Wu
Vahid Tarokh's Group
Electrical and Computer Engineering
Duke University

September 6, 2023

Hyvärinen Score

Definition (Hyvärinen Score)

The Hyvärinen score is a mapping $(X, Q) \mapsto S_{\mathrm{H}}(X, Q)$ given by

$$S_{\mathrm{H}}(X, Q) \triangleq \frac{1}{2} \|\nabla_X \log q(X)\|_2^2 + \Delta_X \log q(X) \tag{1}$$

whenever it can be well defined. Here, ∇_X and $\Delta_X = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ respectively denote the gradient and the Laplacian operators acting on $X = (x_1, \cdots, x_d)^\top$.

Minimizing the Hyvärinen score is associated with score matching, which is an estimation procedure proposed by Hyvärinen and Dayan [1] to minimize the Fisher divergence:

$$\mathbb{D}_{\mathrm{F}}(P\|Q) \triangleq \mathbb{E}_{X\sim P}\left[\|\nabla_X \log p(X) - \nabla_X \log q(X)\|_2^2\right],$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Additionally, it is easy to verify that $\mathbb{D}_{\mathrm{F}}(P\|Q) > 0$ for $Q \neq P$, thus the Hyvärinen score is *strictly proper*.

Comparison between Log Score and Hyvärinen Score

	Log Score	Hyvärinen Score
Definition	$S_{\mathrm{L}}(X, Q) = -\log q(X)$	$S_{\mathrm{H}}(X, Q) = \frac{1}{2}\|\nabla_X \log q(X)\|_2^2 + \Delta_X \log q(X)$
Estimation	$\hat{\theta}_{\mathrm{MLE}} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n S_{\mathrm{L}}(X_i, Q_\theta)$ (Maximum Likelihood)	$\hat{\theta}_{\mathrm{H}} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n S_{\mathrm{H}}(X_i, Q_\theta)$ (Score Matching)
Objective	$\mathbb{D}_{\mathrm{KL}}[P\|\|Q] = \mathbb{E}_p[\log p(X) - \log q(x)]$ (KL divergence)	$\mathbb{D}_{\mathrm{F}}[P\|\|Q] = \mathbb{E}_p\|\nabla_X \log p(X) - \nabla_X \log q(X)\|^2$ (Fisher divergence)
Advantage	Classical Method	Avoid the Normalization Constant

Table: Comparison between Log Score and Hyvärinen Score

Unnormalized Statistical Models

- We consider the parametric distribution: $Q_\theta \in \mathcal{Q}_\theta$ for $\theta \in \Theta$ and its PDF q_θ
- Suppose that our knowledge of the PDF is limited: $q_\theta(X) \propto \tilde{q}_\theta(X)$, such that $q_\theta(X) = \frac{\tilde{q}_\theta(X)}{\int \tilde{q}_\theta(X) dX}$, and $\int \tilde{q}_\theta(X) dX$ is unknown.
- Obtaining the exact likelihood can be computationally challenging (or even intractable):
  - The approximations, such as Monte Carlo-based methods, may suffer from computational errors.
  - The partition function (denominator) is not easy to compute but the unnormalized form (numerator) is simple.
- Alternative solution to the analysis of the unnormalized models?
  - Avoid computing cumbersome normalizing constants
  - Consider methods that only depend on gradients of the logarithmic density function - Hyvärinen Score-based methods

Our Work

- Score-based hypothesis testing for unnormalized models
  - Collaboration with Enmao Diao, Khalil Elkhalil, Jie Ding, and Vahid Tarokh
- Score-based quickest change detection for unnormalized models
  - Collaboration with Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh
- Robust score-based quickest change detection for unnormalized models
  - Collaboration with Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh

Quickest Change Detection

Quickest Change Detection is a fundamental task to detect abrupt changes in the data stream. It is commonly assumed that the random variables are IID with a particular probability density function before the change, and are IID with another density after the change.

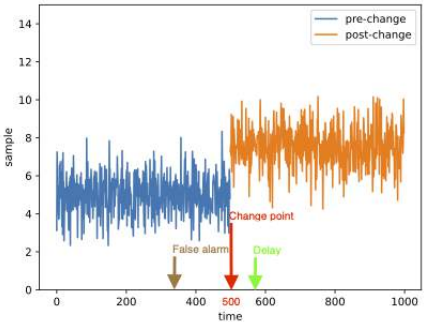


Figure: Quickest Change Detection

Quickest Change Detection

- Problem Formulation:
  - The data stream: $\{X_n\}_{n\geq 1} = (X_1, \ldots, X_n)$ defined on the probability space $(\Omega, \mathcal{F}, P_\nu)$
  - The change point: the time $\nu \geq 1$ when an abrupt change has happened
  - Under P_ν, the observations $X_1, X_2, \ldots, X_{\nu-1} \sim P_\infty$ (pre-change distribution), and $X_\nu, X_{\nu+1}, \cdots \sim P_1$ (post-change distribution)
  - The change point ν is unknown but deterministic
- A quickest change detection algorithm defines a stopping rule T w.r.t. the data stream $\{X_n\}_{n\geq 1}$.
- If $T \geq \nu$, we have made a *delayed detection*; otherwise, a *false alarm* has happened.

Quickest Change Detection

- The objective is to minimize the detection delay subject to a constraint on false alarms.

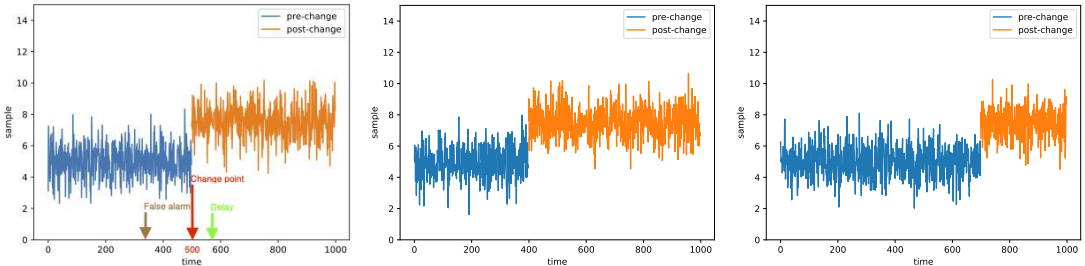


Figure: Different Change Points

Quickest Change Detection

We consider two minimax problem formulations to find the best stopping rule.

- Lorden's metric [2]: the worst-case averaged detection delay (WADD):

$$\mathcal{L}_{\mathtt{WADD}}(T) \triangleq \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_{\nu}[(T-\nu+1)^{+} | \mathcal{F}_{\nu}], \tag{2}$$

  where $(y)^{+} \triangleq \max(y, 0)$ for any $y \in \mathbb{R}$.
- Pollak's metric [3]: the worst conditional averaged detection delay (CADD):

$$\mathcal{L}_{\mathtt{CADD}}(T) \triangleq \sup_{\nu \geq 1} \mathbb{E}_{\nu}[T-\nu | T \geq \nu]. \tag{3}$$

- For false alarms, we consider the *average running length* (ARL), which is defined as:

$$\mathrm{ARL} \triangleq \mathbb{E}_{\infty}[T].$$

- The optimization problem under Pollak's metric becomes

$$\min_{T} \ \mathcal{L}_{\mathtt{CADD}}(T) \text{ subject to } \mathbb{E}_{\infty}[T] \geq \gamma. \tag{4}$$

Robust Quickest Change Detection

The pre- and post-change distributions may not be precisely known. We assume that each is known within an uncertainty class:

$$P_\infty \in \mathcal{G}_\infty, \text{ and } P_1 \in \mathcal{G}_1.$$

For simplicity, we will assume that the pre-change class is a singleton:

$$\mathcal{G}_\infty = \{P_\infty\}.$$

Our proposed method can also be extended to the case of composite $\mathcal{G}_\infty$.

Robust Quickest Change Detection

The objective is to find a stopping time T to minimize the worst-case detection delay, subject to a constraint γ on $\mathbb{E}_\infty[T]$:

$$\min_{T} \sup_{P_1 \in \mathcal{G}_1} \mathcal{L}_{\texttt{WADD}}(T) \quad \text{subject to} \quad \mathbb{E}_\infty[T] \geq \gamma, \tag{5}$$

where γ is a constraint on the ARL.
We are also interested in the version with the minimax metric introduced by Pollak [3]:

$$\min_{T} \sup_{P_1 \in \mathcal{G}_1} \mathcal{L}_{\texttt{CADD}}(T) \quad \text{subject to} \quad \mathbb{E}_\infty[T] \geq \gamma. \tag{6}$$

Least Favorable Distribution

We define the notion of least favorable distribution (LFD). This approach to defining the least favorable distribution for the quickest change detection is novel.

Definition (Least Favorable Distribution)

Assume that the family $\mathcal{G}_1 = \{G_\theta : \theta \in \Theta_1\}$ is convex and compact. We define

$$Q_1 = \arg \min_{G_\theta \in \mathcal{G}_1} \mathbb{D}_F(G_\theta \| P_\infty). \tag{7}$$

The existence of Q_1 is guaranteed by the compactness of $\mathcal{G}_1$ and the continuity of the Fisher divergence as a function of its arguments. Thus, Q_1 is the closest element of $\mathcal{G}_1$ to P_∞ in the Fisher-divergence sense.

Robust Score-based Quickest Change Detection

Given the pre-change law P_∞ (with density p_∞), we now use Q_1 and its density q_1 to design the RSCUSUM algorithm. We define the instantaneous Robust score-based CUSUM (RSCUSUM) score function $X \mapsto z_\lambda(X)$ by

$$z_\lambda(X) \stackrel{\Delta}{=} \lambda\big(S_{\mathrm{H}}(X, P_\infty) - S_{\mathrm{H}}(X, Q_1)\big), \tag{8}$$

where $\lambda > 0$ is a pre-selected multiplier, $S_{\mathrm{H}}(X, P_\infty)$ and $S_{\mathrm{H}}(X, Q_1)$ are respectively the Hyvärinen score functions of P_∞ and Q_1.

- If the post-change model is precisely known, then the Q_1 in the above equation will be replaced by the known post-change law, where the RSCUSUM is equivalent to SCUSUM.

Robust Score-based Quickest Change Detection

Our proposed stopping rule is given by

$$T_{\texttt{RSCUSUM}} \triangleq \inf\left\{ n \geq 1 : \max_{1 \leq k \leq n} \sum_{i=k}^{n} z_\lambda(X_i) \geq \tau \right\}, \tag{9}$$

where $\tau > 0$ is a stopping threshold, which is usually pre-selected to control false alarms.

Robust Score-based Quickest Change Detection

The stopping rule of RSCUSUM can be written as

$$T_{\texttt{RSCUSUM}} = \inf\{n \geq 1 : Z(n) \geq \tau\},$$

where $Z(n)$ can be computed recursively by

$$\begin{aligned} &Z(0) = 0, \\ &Z(n) \triangleq (Z(n-1) + z_\lambda(X_n))^+, \ \forall n \geq 1. \end{aligned}$$

$Z(n)$ is referred to as the detection score of RSCUSUM at time n.

- The proposed algorithm can be applied in a recursive way, which is not too demanding in computational and memory requirements for online implementation.

Theoretical Results

Lemma

Let P_∞ be the pre-change distribution, $Q_1 \in \mathcal{G}_1$ be the least-favorable distribution, and $Q_2 \in \mathcal{G}_1$ be any other post-change distribution. Then

$$\mathbb{D}_F(Q_1 \| P_\infty) \leq \mathbb{D}_F(Q_2 \| P_\infty) - \mathbb{D}_F(Q_2 \| Q_1)\,.$$

- This lemma is the key to delay and false alarm analysis for RSCUSUM.

Theoretical Results

Theorem (Delay and False Alarm Analysis)

Subject to $\mathbb{E}_\infty[T_{RSCUSUM}] \geq \gamma > 0$, *the stopping rule* $T_{RSCUSUM}$ *satisfies*

$$\mathcal{L}_{WADD}(T_{RSCUSUM}) \sim \mathcal{L}_{CADD}(T_{RSCUSUM}) \sim \mathbb{E}_1[T_{RSCUSUM}] \sim \frac{\log \gamma}{\lambda(\mathbb{D}_F(P_1\|P_\infty) - \mathbb{D}_F(P_1\|Q_1))}$$
$$\lesssim \frac{\log \gamma}{\lambda \mathbb{D}(Q_1\|P_\infty)}, \quad \text{as } \gamma \to \infty.$$

Discussion of Least Favorable Distribution

Consider a general parametric distribution family $\mathcal{P}$ defined on $\mathcal{X}$. We use $\mathcal{P}_m$ to denote a set of a finite number of distributions belonging to $\mathcal{P}$, namely

$$\mathcal{P}_m = \{P_i,\ i = 1, \ldots, m :\ P_i \in \mathcal{P}\},\ m \in \mathbb{N}^+. \tag{10}$$

We use p_i to denote the density of each distribution $P_i,\ i = 1, \ldots, m$. Then, we define a convex set of densities

$$\mathcal{A}_m \triangleq \left\{ x \mapsto \sum_{i=1}^{m} \alpha_i p_i(x) : \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geq 0 \right\}. \tag{11}$$

We further define a set of functions

$$\mathcal{B}_m \triangleq \left\{ x \mapsto \sum_{i=1}^{m} \beta_i(x) \nabla_x \log p_i(x) : \sum_{i=1}^{m} \beta_i(x) = 1,\ \beta_i(x) \geq 0,\ p_i \in \mathcal{P}_m \right\}. \tag{12}$$

Discussion of Least Favorable Distribution

Consider the pre-change distribution P_∞ (with density p_∞) such that $P_\infty \in \mathcal{P}$ and $P_\infty \notin \mathcal{A}_m$. We use $\mathbb{E}_\infty$ to denote its corresponding expectation with p_∞.

Theorem

Assume that there exists an element $P_0 \in \mathcal{P}$ (with density p_0) such that

$$\mathbb{E}_{p_0}\left\{\|\nabla_x \log p_0(X) - \nabla_x \log p_\infty(X)\|_2^2\right\} = \min_{\phi \in \mathcal{B}_m} \mathbb{E}_\phi\left\{\|\phi(X) - \nabla_x \log p_\infty(X)\|_2^2\right\}.$$

Then, we have

$$\mathbb{E}_{p_0}\left\{\|\nabla_x \log p_0(X) - \nabla_x \log p_\infty(X)\|_2^2\right\} = \min_{p \in \mathcal{A}_m} \mathbb{E}_p\left\{\|\nabla_x \log p(X) - \nabla_x \log p_\infty(X)\|_2^2\right\}. \tag{13}$$

- The theorem provides an efficient way to identify the LFD in a convex set with only knowledge of the gradient of the log density functions.

Numerical Results

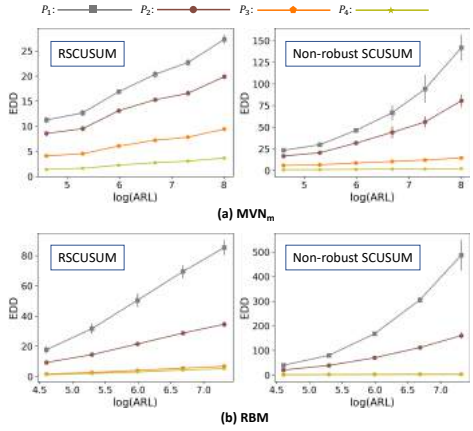


Figure: EDD versus log-scaled ARL. We respectively demonstrate the empirical EDD against log-scaled ARL for both MVN$_m$ and RBM experiments. The results demonstrate that the EDD of RSCUSUM (subplot in left rows) increases at a linear rate for all cases, while some EDD of non-robust SCUSUM (subplot in right rows) increases at an exponential rate.

Thank you!

Questions?

Reference I

[1] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching." *J. Mach. Learn. Res.*, vol. 6, no. 4, 2005.

[2] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, pp. 1897–1908, 1971.

[3] M. Pollak, "Optimal detection of a change in distribution," *Ann. Stat.*, pp. 206–227, 1985.

[4] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," in *J. Mach. Learn. Res.*, vol. 6, 2005, p. 695–709.

[5] J. Ding, R. Calderbank, and V. Tarokh, "Gradient information for representation and modeling," *Advances in Neural Information Processing Systems 32*, pp. 2396–2405, 2019.

[6] K. Elkhalil, A. Hasan, J. Ding, S. Farsiu, and V. Tarokh, "Fisher auto-encoders," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130. PMLR, 2021, pp. 352–360.

Reference II

[7] Q. Liu, J. Lee, and M. Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *International conference on machine learning*. PMLR, 2016, pp. 276–284.

[8] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. null, p. 723–773, mar 2012.

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

[10] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

Reference III

[12] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.

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

[14] M. Hutchinson, "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines," *Communications in Statistics - Simulation and Computation*, vol. 18, no. 3, pp. 1059–1076, 1989. [Online]. Available: https://doi.org/10.1080/03610918908812806

[15] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "Analysis and results of the 1999 darpa off-line intrusion detection evaluation," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2000, pp. 162–182.

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

Backup Slides

Backup Slides

Score Matching

For a random variable $X \in \mathcal{X} \subset \mathbb{R}^d$ and the probability density functions (PDFs) $X \mapsto p(X)$ and $X \mapsto q(X)$, we consider the Fisher divergence from p to q:

$$\mathbb{D}_{\mathrm{F}}[p||q] \triangleq \mathbb{E}_{X\sim p}\|\nabla_X \log p(X) - \nabla_X \log q(X)\|^2 = \mathbb{E}_{X\sim p}\left[\frac{1}{2}\|\nabla_X \log p(X)\|_2^2 + S_{\mathrm{H}}(X, Q)\right], \tag{14}$$

where $(q, X) \mapsto S_{\mathrm{H}}(X, Q)$ is given by

$$S_{\mathrm{H}}(X, Q) \triangleq \frac{1}{2}\|\nabla_X \log q(X)\|_2^2 + \Delta_X \log q(X), \tag{15}$$

also known as the Hyvärinen score [1]. Here, $\Delta_X = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$ denotes the Laplacian operator with respect to $X = (x_1, \cdots, x_d)^\top$.

- The minimum of (14) is achieved if and only if $q(X) = p(X)$ for any $X \in \mathcal{X}$
- The term $\mathbb{E}_{X\sim p}\left[\frac{1}{2}\|\nabla_X \log p(X)\|_2^2\right]$ can be seen as a constant to the distribution q

Score Matching

Consider $q \in \mathcal{Q} = \{q_\theta : \theta \in \Theta \subset \mathbb{R}^r\}$, where each q_θ is a PDF and $p = q_{\theta^\star}$ for some $\theta^\star \in \Theta$. The minimization over Fisher divergence is then reduced to the minimization of the expected Hyvärinen score:

$$\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{X \sim p} s_{\mathrm{H}}(q_\theta, X). \tag{16}$$

Suppose that a finite sample of points $X_1, \ldots, X_n$ are independent and identically distributed (IID) according to p. An estimator in parallel to Problem (16) is

$$\hat{\theta}_{\mathrm{H}} \triangleq \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} s_{\mathrm{H}}(q_\theta, X_i). \tag{17}$$

The procedure of estimating the above $\hat{\theta}_{\mathrm{H}}$ is known as score matching estimation [4, 5, 6], which is a surrogate for maximum likelihood estimation (MLE).

Hypothesis Testing

Hypothesis testing is a procedure to decide whether or not to reject a hypothesis. Given IID observations $X_n = \{X_1, \ldots, X_n\}$ according to $p_{\theta^\star}$, we are interested in testing the hypothesis if $\theta^\star = \theta_0$ for a given $\theta_0 \in \Theta$.

- Problem Formulation:

$$\mathcal{H}_0 : \theta^\star = \theta_0 \quad \text{against} \quad \mathcal{H}_1 : \theta^\star \in \Theta \backslash \{\theta_0\}. \tag{18}$$

- Generalized Likelihood Ratio Test (GLRT)
  - Take the ratio of log-likelihoods as the test statistic:

$$T_{\mathtt{LRT}}(X_n) = -2 \left[\ \ell(\theta_0, X_n) - \ell(\hat{\theta}_{\mathtt{MLE}}, X_n) \ \right], \tag{19}$$

    where $\ell(\theta, X_n) = \ln \mathcal{L}(\theta, X_n)$ and $\hat{\theta}_{\mathtt{MLE}} = \sup_{\theta \in \Theta} \mathcal{L}(\theta, X_n)$.
  - Neyman-Pearson Lemma: The uniformly most powerful optimality for simple hypothesis testing

Hyvärinen Score Test

We develop a new statistical test, referred to as the Hyvärinen score test (HST) that is applicable to unnormalized models.

- Test statistic, denoted as $T_{\mathtt{HST}}$:

$$T_{\mathtt{HST}}(X_n) \triangleq 2\big(\mathcal{S}_{\mathtt{H}}(X_n, \theta_0) - \mathcal{S}_{\mathtt{H}}(X_n, \hat{\theta}_{\mathtt{H}})\big), \tag{20}$$

$$\tilde{T}_{\mathtt{HST}}(X_n) \triangleq 2\big(\mathcal{S}_{\mathtt{H}}(X_n, \theta_0) - \mathcal{S}_{\mathtt{H}}(X_n, \theta_1)\big), \tag{21}$$

where $\mathcal{S}_{\mathtt{H}}(X_n, \theta) \triangleq \sum_{i=1}^{n} S_{\mathtt{H}}(X_i, Q_\theta)$, and $\hat{\theta}_{\mathtt{H}}$ is the score matching estimate.

The HST rejects the null hypothesis when the test statistic is larger than a critical value, which can be identified using a large-sample asymptotic distribution.

Theoretical Results

Proposition (Asymptotic distribution of $\tilde{T}_{\mathtt{HST}}$ under the null hypothesis)

Assuming some regularity conditions[a], under the null hypothesis, we have

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

where $Z \sim \mathcal{N}(\mathbf{0}_r, \mathrm{Var}_\star(S_{\mathtt{H}}(X, Q_{\theta_0}) - S_{\mathtt{H}}(X, Q_{\theta_1})))$ and $\mathrm{Var}_\star$ denotes the variance w.r.t. the null distribution.

[a] see the conditions in backup slides.

Theoretical Results

Theorem (Asymptotic distribution of $T_{\mathtt{HST}}$ under the null hypothesis)

Assuming some regularity conditions[a]*, under the null hypothesis, we have*

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

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

$$\mathbf{H} \triangleq \mathbb{E}_\star \left[\nabla_\theta^2 S_H(X, Q_\theta) \mid_{\theta=\theta_0} \right], \tag{22}$$

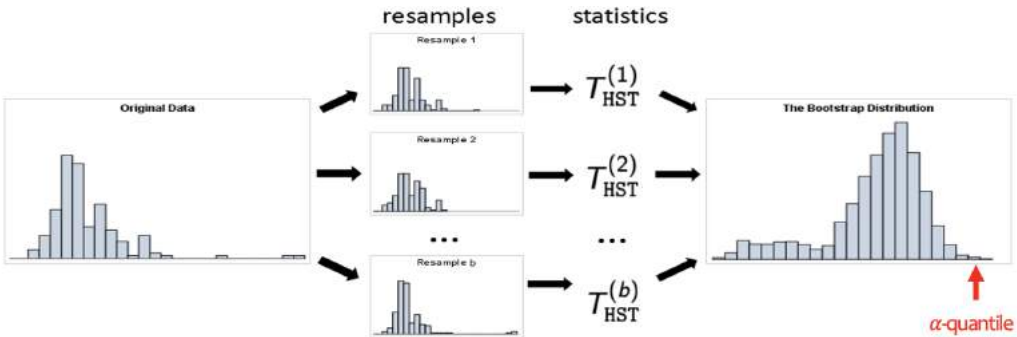
$$\mathbf{K} \triangleq \mathbb{E}_\star \left[\nabla_\theta S_H(X, Q_\theta) \nabla_\theta^\top S_H(X, Q_\theta) \mid_{\theta=\theta_0} \right], \tag{23}$$

$\xrightarrow{n\to\infty}_{\mathcal{L}}$ *denotes convergence in distribution, and* $\mathbb{E}_\star$ *denotes the expectation w.r.t. the null distribution.*

[a] see the conditions in backup slides.

Bootstrap Hyvärinen Score Test

- We propose to use a bootstrap method to empirically determine the rejection region for HST
- The main idea is to determine the critical value by the empirical $(1-\alpha)$-quantile of the distribution of $T_{\mathtt{HST}}(X_n)$ under the null hypothesis
- α is the Type I error probability that we want to control



Experimental Results

1. We consider comparing the performance of HST with
   - the GLRT, and LRT
   - the Kernelized Stein Discrepancy (KSD) test, by Liu et al. [7],
   - and the Maximum Mean Discrepancy (MMD) test, by Gretton et al. [8].
2. The results demonstrate that our method performs competitively with LRT and outperforms other baseline methods in terms of empirical Type II error rates.
3. Our experiments further illustrate the computational advantage of our approach for unnormalized models over LRT.
4. Additionally, we will show that the proposed approach achieves success in the Out-of-distribution (OOD) detection task.

Synthetic Experiments of Gauss-Bernoulli RBM

We consider the hypothesis testing on weight matrix $\mathbf{W}$:

$$\text{Simple Test:} \quad \mathbf{W}^\star = \mathbf{W}_0 \quad \text{versus} \quad \mathbf{W}^\star = \mathbf{W}_1 \tag{24}$$

$$\text{Composite Test:} \quad \mathbf{W}^\star = \mathbf{W}_0 \quad \text{versus} \quad \mathbf{W}^\star \neq \mathbf{W}_0 \tag{25}$$

- Randomly draw the weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_h}$ from $\mathcal{N}(0, 1)$.
- 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} to $\mathbf{W}_0$.
- The samples $X_1, \ldots, X_n \sim p_{\mathbf{W}^\star}$ are drawn using Gibbs sampling with 1000 iterations.

Synthetic Experiments of Gauss-Bernoulli RBM

In Figure below, we present the test statistics and ROC curves for the above tests.

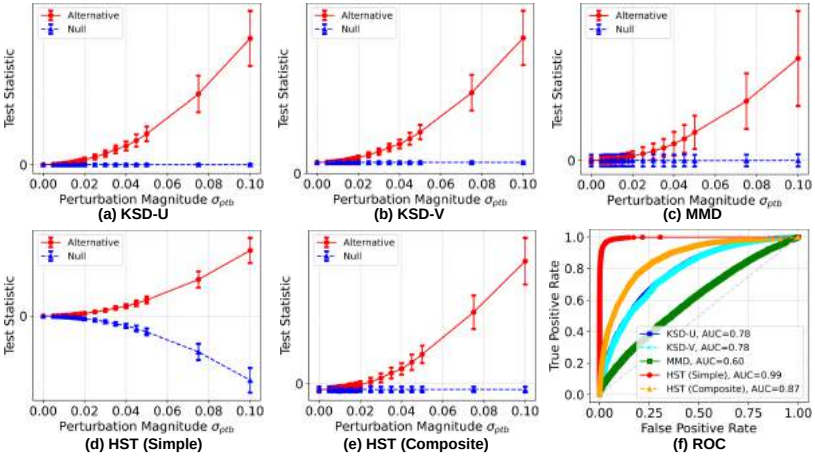


Figure: (a-e) Test statistics at $n = 100$. (f) Receiver Operating Characteristic (ROC) curves of various tests with $\sigma_{ptb} = 0.01$ and $n = 100$.

Application to Out-of-distribution Detection

- Out-of-distribution Detection (OOD):
  - Intersects with anomaly detection, adversarial attacks, and incremental learning
  - The target is to determine whether one (or few) given input is from the training data distribution (in-distribution examples) or not (out-of-distribution examples).

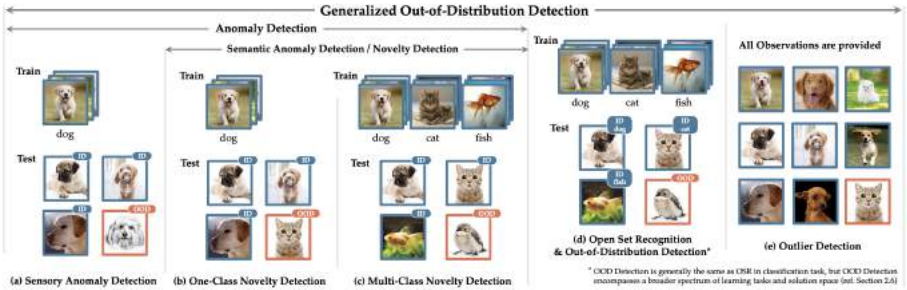


Figure: Yang et al. [9]. Illustration of sub-tasks under generalized OOD detection framework with vision tasks.

Application to Out-of-distribution Detection

HST for OOD:

- The aggregate Hyvärinen score $\mathcal{S}_{\mathrm{H}}(X_n, Q_{\theta_0})$ is used for OOD detection,

$$\mathcal{S}_{\mathrm{H}}(X_n, Q_{\theta_0}(X)) = \sum_{i=1}^{n} S_{\mathrm{H}}(X_i, Q_{\theta_0}) \triangleq \frac{1}{2} \|\nabla_{X_i} \log q_{\theta_0}(X_i)\|_2^2 + \Delta_{X_i} \log q_{\theta_0}(X_i), \tag{26}$$

  where the density function q_{θ_0} is learned from the in-distribution sample and X_n is the out-of-distribution sample.

- We reject the in-distribution hypothesis when $\mathcal{S}_{\mathrm{H}}(X_n, Q_{\theta_0})$ is larger than a threshold.
- The threshold can be decided empirically by repeating the tests over the in-distribution train data.

Image Data

We evaluate the performance of HST on the computer vision benchmark datasets:

- In-distribution: CIFAR-10 (Krizhevsky et al. [10])
- Out-of-distribution: TinyImageNet, a subset of ImageNet (Deng et al. [11]).

Figure: Left: CIFAR10 (Krizhevsky et al. [10]); Right: Tiny ImageNet(Deng et al. [11])

Image Data

Implementation Details:

- **Term 1**: The gradient of logarithmic density function $\nabla_X \log q_\theta(X)$ is modelled with a SDE-based deep generative model, Noise Conditional Score Network with variance exploding SDEs (NCSN++) [12].
  - The model architecture includes four BigGAN-type [13] residual blocks per image resolution.
  - we randomly crop image patches of size 32×32 to match the shape of CIFAR-10.
- **Term 2**: The Laplacian term $\Delta_X \log q_\theta(X)$ is not easy to be computed in high dimensions.
  - Apply Hutchinson's trick [14] to reduce its computation complexity
  - The Hutchinson method obtains the unbiased estimate of the Laplacian term by Monte Carlo sampling:

$$\Delta_X \log q_\theta(X) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\boldsymbol{\epsilon}^T \cdot \nabla_X f(X, \theta) \cdot \boldsymbol{\epsilon} \right] = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\boldsymbol{\epsilon}^T \cdot \nabla_X \left(\boldsymbol{\epsilon}^T f(X, \theta) \right) \right], \tag{27}$$

where $f(X, \theta) \triangleq \nabla_X \log q_\theta(X)$, and random projections $\boldsymbol{\epsilon}$ are Normally distributed.

Image Data

We evaluate the performance of HST for OOD detection by varying the test sample size.

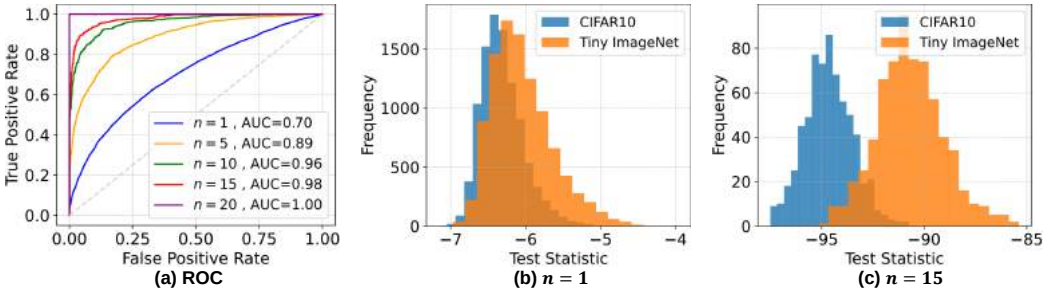


Figure: (a) ROC curves and (b, c) histograms of test statistics of HST for OOD Detection on CIFAR10 (in-distribution) and Tiny ImageNet datasets (out-distribution).

In Figure 7, we present the ROC curve and histograms of $\mathcal{S}_{\mathrm{H}}(Q_{\theta_0}, X_n)$ over different sample sizes.

Network Intrusion Detection

We next perform OOD on the KDD Cup 1999 [1] dataset. The dataset contains includes a wide variety of intrusions simulated in a military network environment [15].

- In-distribution: 'normal' network
- Out-of-distribution: attack networks.
  - A total of 24 training attack types in train data with an additional 14 types in the test data.
  - For example, unauthorized access from a remote machine attacks, e.g. guessing password.

Implementation Details:

- We train a Gauss-Bernoulli RBM with in-distribution samples to model the density function $q_\theta(X)$.
- The aggregate Hyvärinen score of Gauss-Bernoulli RBM can be computed in a closed-form (shown in backup slides).

[1] The Fifth International Conference on Knowledge Discovery and Data Mining

Network Intrusion Detection

From Figure 8, we depict the ROC curves and the histograms of $\mathcal{S}_{\mathrm{H}}(X_n, Q_{\theta_0})$ for detecting the malicious network attack.

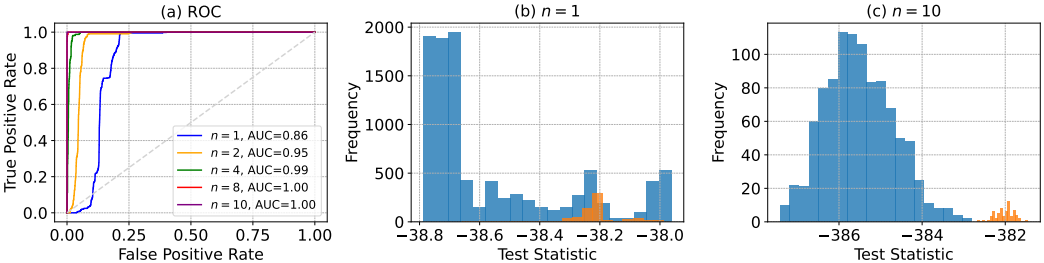


Figure: (a) ROC curves and (b, c) histograms of test statistics of the 'ipsweep' attack and 'normal' network of HST on KDD Cup'99 dataset.

The results demonstrate that our method can detect adversarial network attacks even with a single out-of-distribution data point. Naturally, our method performance significantly improves when more out-of-distribution samples are available.

Network Intrusion Detection

In Table 2, we evaluate our method with AUC.

n (size)/ Attacks	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) / Attacks	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

Table: Area Under the Curve of Receiver Operating Characteristics (AUC) for our test to detect malicious network attack for various values of sample size n.

Asymptotic Behavior of HST

We first introduce some regularity conditions.

Assumption

1. The family $q_\theta(X)$ is identifiable, i.e., that $\theta \neq \theta^\star \to q_\theta \neq q_{\theta^\star}$.
2. For all $X \in \mathcal{X}$, $s_{\mathrm{H}}(q_\theta, X)$ is continuous in $\theta \in \Theta$ for Θ compact.
3. There exists a function ξ_1: $X \to \xi_1(X)$ such that for any $\theta \in \Theta$, $|s_{\mathrm{H}}(q_\theta, X)| \leq \xi_1(X)$ and $\mathbb{E}_\star\left[\xi_1(X)\right] < \infty$.
4. For any $\theta \in \Theta$, $q_\theta(X)\nabla_X \log q_\theta(X) \to 0$ as $\|X\|_2 \to \infty$.
5. $\theta^\star$ is an interior point of the parameter space Θ.
6. For all $X \in \mathcal{X}$, $s_{\mathrm{H}}(q_\theta, X)$ is twice continuously differentiable in the interior of Θ.
7. The expected values $\mathbb{E}_\star\left[\nabla_\theta s_{\mathrm{H}}(q_\theta, X)\nabla_\theta^\top s_{\mathrm{H}}(q_\theta, X)\mid_{\theta=\theta_0}\right]$ and $\mathbb{E}_\star\left[\nabla_\theta^2 s_{\mathrm{H}}(q_\theta, X)\mid_{\theta=\theta_0}\right]$ exist and are non-singular. There exists a function $\xi_2 : X \mapsto \xi_2(X)$ such that $\left|\frac{\partial^2 s_{\mathrm{H}}(q_\theta, X)}{\partial\theta_i \partial\theta_j}\right| \leq \xi_2(X)$ for all $1 \leq i, j \leq k$, and $\mathbb{E}_\star\left[\xi_2(X)\right] < \infty$.

Asymptotic Behavior of HST

Theorem (Asymptotic distribution of T_{HST} under the null hypothesis)

Assuming the regularity conditions holds, under the null hypothesis, we have

$$T_{HST}(X_n) \xrightarrow{n\to\infty}_{\mathcal{L}} \mathbf{z}^T \mathbf{H}\mathbf{z}, \tag{28}$$

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

$$\mathbf{H} \triangleq \mathbb{E}_\star \left[\nabla_\theta^2 s_H(q_\theta, X) \mid_{\theta=\theta_0} \right], \tag{29}$$

$$\mathbf{K} \triangleq \mathbb{E}_\star \left[\nabla_\theta s_H(q_\theta, X) \nabla_\theta^\top s_H(q_\theta, X) \mid_{\theta=\theta_0} \right], \tag{30}$$

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

Gauss-Bernoulli RBM

The RBM [16] 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 $X \in R^{d_x}$ with joint distribution

$$p(X, \mathbf{h}) = \frac{1}{Z} \exp\left\{ -\left(\frac{1}{2} \sum_{i=1}^{d_x} \sum_{j=1}^{d_h} \frac{x_i}{\sigma_i} W_{ij} h_j + \sum_{i=1}^{d_x} b_i x_i + \sum_{j=1}^{d_h} c_j h_j - \frac{1}{2} \sum_{i=1}^{d_x} \frac{x_i^2}{\sigma_i^2} \right) \right\}, \tag{31}$$

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.

Gauss-Bernoulli RBM

The probability of the visible variable X written as $p(X) = \sum_{h \in \{0,1\}^{d_h}} p(X, \mathbf{h}) = \frac{1}{Z} \exp\{-F_\theta(X)\}$, where $F_\theta(X)$ is the free energy given by

$$F_\theta(X) = \frac{1}{2} \sum_{i=1}^{d_x} (x_i - b_i)^2 - \sum_{j=1}^{d_h} \text{Softplus}\left(\sum_{i=1}^{d_x} W_{ij} x_i + b_j \right). \tag{32}$$

The Softplus function is defined as $\text{Softplus}(t) \triangleq \log(1 + \exp(t))$ with a default scale parameter $\beta = 1$. By Equation (15), the corresponding Hyvärinen score $S_{\text{H}}(X_n, \theta)$ is given by

$$S_{\text{H}}(X_n, \theta) = \sum_{n=1}^{n} \sum_{i=1}^{d_x} \left[\frac{1}{2} \left(x_{in} - b_i + \sum_{j=1}^{d_h} W_{ij} \delta_{jn} \right)^2 + \sum_{j=1}^{d_h} W_{ij}^2 \delta_{jn} (1 - \delta_{jn}) - 1 \right], \tag{33}$$

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

Assumptions for Quickest Change Detection

- $P_1 \neq P_\infty$
- The same mild regularity conditions so that the Hyvärinen score is well-defined[2]:
  - The pre- and post-change PDFs, e.g. $p_1(x), p_\infty(x)$, are differentiable with respect to x.
  - The functions $\nabla_x \log p_\infty(x)$ and $\nabla_x \log p_1(x)$ are differentiable w.r.t. x.
  - The expectations $\mathbb{E}_{X\sim p_\infty}[\|\nabla_x \log p_\infty(X)\|_2^2]$ and $\mathbb{E}_{X\sim p_1}[\|\nabla_x \log p_1(X)\|_2^2]$ are finite.
  - $p(x)\nabla_x \log p_\infty(x) \to 0$ and $p(x)\nabla_x \log p_1(x) \to 0$ when $\|x\|_2 \to 0$.

[2] Hyvärinen and Dayan [1]