Diffusion-Based Hypothesis Testing and Change-Point Detection

Anonymous authors
Paper under double-blind review

Abstract

Score-based methods have recently seen increasing popularity in modeling and generation. Methods have been constructed to perform hypothesis testing and change-point detection with score functions, but these methods are in general not as powerful as their likelihood-based peers. Recent works consider generalizing the score-based Fisher divergence into a diffusion-divergence by transforming score functions via multiplication with a matrix-valued function or a weight matrix. In this paper, we extend the score-based hypothesis test and change-point detection stopping rule into their diffusion-based analogs. Additionally, we theoretically quantify the performance of these diffusion-based algorithms and study scenarios where optimal performance is achievable. We propose a method of numerically optimizing the weight matrix and present numerical simulations to illustrate the advantages of diffusion-based algorithms.

1 Introduction

In many engineering problems, one seeks to infer whether some data X was generated from a null distribution P_∞ or an alternate distribution P_1. This question underpins the detection problems of hypothesis testing and anomaly detection (Lopez-Paz & Oquab (2017); Liu et al. (2020); Angiulli et al. (2026); Rahman & Wagner (2012); Watanabe (2018) and change-point detection (Kei et al. (2025); Yamada et al. (2025); Unnikrishnan et al. (2011); Blostein (1991); Liu & Blostein (1994); Huang et al. (2021); Sun & Zou (2024)). These problems have been extensively studied under the assumption that the densities of P_∞ and P_1 are known (or can be learned from data). In this setting, the log-likelihood ratio (LLR) test (8) and cumulative sum (CUSUM) stopping rule (14) are commonly used and optimal under the metrics of (6) and (13). Further background on hypothesis testing and change-point detection are provided in Section 3.

Score-based models (Song & Ermon (2019); Song et al. (2019); Melidonis et al. (2025); Berman et al. (2025); Kollovieh et al. (2024)) directly estimate the score $s(X) = \nabla_X \log p(X) : \mathbb{R}^d \mapsto \mathbb{R}^d$ of a distribution P by training over a dataset of independent and identically distributed (i.i.d.) samples of P. Furthermore, Markov Chain Monte Carlo methods (Roberts & Rosenthal (2001); Liao et al. (2022); Aloui et al. (2025)) have been developed to sample from P using $\nabla_X \log p(X)$. Recent works have proposed *score-based* (Definition 2.1) methods of hypothesis testing (Wu et al. (2022)) and change-point detection (Wu et al. (2023)) under the assumption that only the score functions of P_∞, P_1 are known. Further background on score-based models is provided in Section 2.

Under a popular objective for simple hypothesis testing (6) and change-point detection (13), the LLR test and CUSUM stopping rule are provably optimal *for any bound on false alarm* (and in the case of hypothesis testing, for any batch size). Their score-based counterparts, however, are not in general optimal, and for specific choices of P_∞, P_1, can be shown to strictly underperform the LLR-based algorithms. Even so, the score-based methods are merely one set of algorithms that can test hypotheses and detect change-points using score functions. We investigate whether some other choice of algorithms which use the score functions can better compete with the LLR-based methods.

To do so, we first present the *diffusion divergence*, studied by Barp et al. (2022) and applied to detection tasks by Altamirano et al. (2023). The diffusion divergence generalizes the Fisher divergence by transforming scores

$\nabla_X \log p_\infty(X), \nabla_X \log p_1(X)$ by multiplication with a matrix-valued function $m(X) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times w}$ which we refer to as the *diffusion matrix function*. The score-based algorithms are one special case of the diffusion algorithms where the identity matrix $I \in \mathbb{R}^{d\times d}$ is chosen to be the diffusion matrix function $m(X)$.

The performance of the diffusion algorithms will depend upon their particular choice of $m(X)$. Using the metrics of (6), (13), a good choice of $m(X)$ will produce diffusion algorithms which perform at least as well as the score-based algorithms and no better than the (provably optimal) LLR-based algorithms. Conversely, diffusion algorithms can perform worse than even score-based algorithms for poorly chosen $m(X)$. Thus, the selection of $m(X)$ so that it is optimal in a well-defined sense is an important goal of our work.

We now summarize the main contributions of this paper:

1. We propose a hypothesis test and a change-point detection stopping rule that are based on the diffusion divergence, generalizing the score-based hypothesis test of Wu et al. (2022) and change-point detection stopping rule of Wu et al. (2023) by way of a *diffusion matrix function* $m(X)$. These algorithms are proposed in Section 4.
2. We bound (or calculate) the error exponent, expected detection delay, and mean time to false alarm of the diffusion-based algorithms asymptotically and as a function of $m(X)$. These properties are presented in Section 4.
3. For each detection task, we present an optimization objective over $m(X)$. We present this objective for change-point detection in Section 5.1 and for hypothesis testing in Section 5.2.
4. We show that for special choices of P_∞, P_1, there exists an $m(X)$ such that $Z_m(X) = Z_{\mathtt{KL}}(X)$ for all $X \in \mathbb{R}^d$. Thus, for these choices of P_∞, P_1, the diffusion-based tests are optimal and equal to the LLR-based counterparts. We show that for other choices of P_∞, P_1, there is no $m(X)$ for which $Z_m(X) = Z_{\mathtt{KL}}(X)$ with probability one under the measures of P_∞ or P_1. We argue that the best diffusion algorithm for either detection task will perform no worse than the score-based algorithm and no better than the LLR algorithm. We present this analysis in Section 5.3.

The remainder of this paper is organized as follows: in Section 2, we provide background on score-based models. In Section 3, we provide background on the detection problems of hypothesis testing and change-point detection, and discuss LLR-based and score-based approaches to these problems. In Section 5.4, we propose a differentiable loss function for the numerical optimization of $m(X)$. In Section 6, we present numerical simulations to illustrate the performance and implementability of our proposed algorithms, and we provide the implementation details of these simulations in Appendix A. Finally, in Appendix B, we provide proofs for the Theorems and Lemmas presented throughout the paper.

Throughout this paper, for any distribution P, we shall refer to the density as $p(X)$, the corresponding probability measure as $\mathbb{P}_P$, and the expectation as $\mathbb{E}_P$.

2 Score-Based Models

In this section, we provide background on score-based models, which utilize the score $\nabla_X \log p(X)$ of a distribution P. We begin by defining the Fisher divergence between two distributions:

Definition 2.1 (Fisher Divergence)**.** *The Fisher divergence between distributions P and Q is defined to be:*

$$\mathbb{D}_{\mathtt{F}}(P\|Q) = \mathbb{E}_{X\sim P}\left[\tfrac{1}{2}\left\|\nabla_X \log p(X) - \nabla_X \log q(X)\right\|_2^2\right].$$

We next define the Hyvärinen score:

Definition 2.2 (Hyvärinen Score)**.** *We define the Hyvärinen score of X with respect to distribution Q as:*

$$\mathcal{S}_{\mathtt{H}}(X, Q) = \tfrac{1}{2}\left\|\nabla \log q(X)\right\|_2^2 + \Delta_X \log q(X), \tag{1}$$

where Δ_X is the Laplacian operator: $\Delta_X f(X) = \sum_{i=1}^d \partial^2 f(X)/\partial X_i^2$.

Lemma 2.3 (Hyvärinen's Theorem)**.** *Let $s(X) : \mathbb{R}^d \mapsto \mathbb{R}^d$. Under the assumption that $\nabla_X \log p(X)$ and $s(X)$ are differentiable with a continuous derivative, that $\nabla_X \log p(X) \to 0$ and $s(X) \to 0$ as $\|X\| \to \infty$, and that both $\mathbb{E}_P[\|\nabla_X \log p(X)\|^2]$ and $\mathbb{E}_P[\|s(X)\|^2]$ are finite, we have that:*

$$\mathbb{E}_P\left[\tfrac{1}{2}\|\nabla_X \log p(X) - s(X)\|^2\right] = \mathbb{E}_P\left[\tfrac{1}{2}\|\nabla_X \log p(X)\|^2 + \tfrac{1}{2}\|s(X)\|^2 + tr(\nabla_X s(X))\right]. \tag{2}$$

If for some distribution Q with density $q(X)$ we set $s(X) = \nabla_X \log q(X)$, then we have that

$$\mathbb{D}_{\mathtt{F}}(P\|Q) = \mathbb{E}_P\left[\tfrac{1}{2}\|\nabla_X \log p(X)\|^2\right] + \mathbb{E}_P\left[\mathcal{S}_{\mathtt{H}}(X, Q)\right]. \tag{3}$$

Proof. The proof is given in Appendix B. □

If we do not know the density or score function of some distribution P, but do have a dataset $(X_i)_{i=1}^n$ of i.i.d. samples of P, we can use score matching (Song & Ermon (2019); Song et al. (2019); Vincent (2011)) to estimate $\nabla_X \log p(X)$ by $s_\theta(X) : \mathbb{R}^d \mapsto \mathbb{R}^d$, where θ parameterizes s. Score matching optimizes

$$\min_\theta \sum_{i=1}^n J_\theta(X_i) \quad \text{where} \quad J_\theta(X) = \left(\tfrac{1}{2}\|s_\theta(X)\|^2 + tr(\nabla_X s_\theta(X))\right). \tag{4}$$

We note that in the limit of large n, $\frac{1}{n}\sum_{i=1}^n J_\theta(X_i)$ converges to $\mathbb{E}_P[J_\theta(X)]$, and that if there exists a distribution Q_θ for which $s_\theta(X) = \nabla_X \log q_\theta(X)$, then $\mathbb{E}_P[J_\theta(X)] = \mathbb{E}_P[\mathcal{S}_{\mathtt{H}}(X, Q_\theta)]$. In this case, score matching produces the θ which minimizes $\mathbb{D}_{\mathtt{F}}(P\|Q_\theta)$ by (3). We note that score matching is analogous to maximum likelihood estimation when $\mathbb{D}_{\mathtt{F}}$ and $\mathcal{S}_{\mathtt{H}}$ are replaced by $\mathbb{D}_{\mathtt{KL}}$ and negative-log-likelihood, respectively.

3 Score-Based Hypothesis Testing and Change-Point Detection

We now present background on hypothesis testing and change-point detection, which we shall refer to as our *detection tasks*. Both detection tasks require us to discern between two competing hypotheses: hypothesis $\mathcal{H}_\infty$, under which the data was generated by P_∞, and hypothesis $\mathcal{H}_1$, under which the data was generated by P_1. For both tasks, we discuss LLR-based and score-based solutions.

In this section and throughout the paper, we shall make the following assumptions regarding P_∞ and P_1:

Assumption 3.1. *We assume that: (i) P_∞ and P_1 are both supported on $\mathbb{R}^d$, (ii) $P_\infty \neq P_1$, (iii) P_∞, P_1 are absolutely continuous with respect to the Lebesgue measure, and (iv) The densities p_∞, p_1 are twice differentiable with continuous derivatives.*

3.1 Simple Hypothesis Testing

Suppose n samples $(X_i)_{i=1}^n$ are drawn independently from some distribution P_*: under the null hypothesis $\mathcal{H}_\infty$, $P_* = P_\infty$ and under the alternate hypothesis $\mathcal{H}_1$, $P_* = P_1$. The task of simple hypothesis testing (hereafter referred to as hypothesis testing) is to decide whether or not to reject $\mathcal{H}_\infty$.

We define a *test* T to be a function that maps data $(X_i)_{i=1}^n$ to a decision: $T((X_i)_{i=1}^n) \in \{0, 1\}$. We reject $\mathcal{H}_\infty$ if $T((X_i)_{i=1}^n) = 1$, and fail to reject $\mathcal{H}_\infty$ otherwise. A type I (type II) error occurs when the test incorrectly rejects (incorrectly fails to reject) $\mathcal{H}_\infty$. We denote the probabilities of type I and type II error as

$$\alpha_n(T) = \mathbb{P}_\infty[T((X_i)_{i=1}^n) = 1], \qquad \beta_n(T) = \mathbb{P}_1[T((X_i)_{i=1}^n) = 0] \tag{5}$$

respectively. We refer to n as the *batch size* of test T.

The work of this paper is motivated by the Neyman-Pearson regime, in which one cannot tolerate a probability of type I error in excess of some threshold $\overline{\alpha}$. Thus, we wish to find a test T that is a solution of (or at least *approximates* a solution of):

$$\min_T \beta_n(T) \;\text{ subject to }\; \alpha_n(T) \leq \overline{\alpha}. \tag{6}$$

The *type II error exponent* describes the relationship between batch size and type II error probability for large n.

Definition 3.2 (Type II Error Exponent)**.** *For a test T, the type II error exponent E is given by*

$$E = -\limsup_{n\to\infty} \frac{\log(\beta_n(T))}{n}. \tag{7}$$

We refer to the type II error exponent of any test T as $\mathcal{B}(T)$.

We next proceed to consider two solutions to the problem of hypothesis testing.

3.1.1 Log-Likelihood Ratio Test

When one has complete knowledge of the densities of P_∞, P_1, one can use the log-likelihood ratio test:

$$T^c_{\mathtt{KL}}((X_i)_{i=1}^n) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Z_{\mathtt{KL}}(X_i) < c \\ 1 & \text{else,} \end{cases} \quad \text{where} \quad Z_{\mathtt{KL}}(X) = \log p_1(X) - \log p_\infty(X), \tag{8}$$

and where $c \in \mathbb{R}$ is some pre-determined threshold. Under the assumption that $\sum_{i=1}^n Z_{\mathtt{KL}}(X_i) \neq c$ almost surely, the Neyman-Pearson Lemma proves that the log-likelihood ratio test of (8) is optimal under the objective of (6) for all choices of $\overline{\alpha} \in (0, 1)$ and for all batch sizes $n \in \mathbb{N}$ (Neyman & Pearson (1933); Cover & Thomas (2006); Trees (2001)).

3.1.2 Hyvärinen Score Test

In Wu et al. (2022), a score-based hypothesis test which utilizes the Hyvärinen score was proposed. It is defined as

$$T^c_{\mathtt{F}}((X_i)_{i=1}^n) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Z_{\mathtt{F}}(X_i) < c \\ 1 & \text{else,} \end{cases} \quad \text{where} \quad Z_{\mathtt{F}}(X) = \mathcal{S}_{\mathtt{H}}(X, P_\infty) - \mathcal{S}_{\mathtt{H}}(X, P_1), \tag{9}$$

and where $c \in \mathbb{R}$ is some pre-determined threshold.

A composite hypothesis test was proposed in Wu et al. (2022) for the case where only the null distribution is precisely known. Furthermore, Wu et al. (2022) derived limiting distributions on $\frac{1}{n}\sum_{i=1}^n Z_{\mathtt{F}}(X_i)$ for $X_i \overset{iid}{\sim} P_\infty$ as $n \to \infty$, and provided conditions under which this limiting distribution (after scaling by $\sqrt{n}$ and shifting) is the standard normal distribution.

3.2 Online Change-Point Detection

We next consider a problem in which we must detect between $\mathcal{H}_\infty, \mathcal{H}_1$ under temporal constraints. Suppose we are given probability distributions P_∞, P_1 and a natural number $\nu \in \mathbb{N}$. A data-stream $(X_i)_{i=1}^\infty$ indexed by time $i \in \mathbb{N}$ can be generated by the following process:

$$X_i \overset{iid}{\sim} P_\infty \text{ for } i < \nu, \qquad X_i \overset{iid}{\sim} P_1 \text{ for } i \geq \nu. \tag{10}$$

We observe the data stream in real-time: at any time i, we can see past and present data, $(X_j)_{j\leq i}$, but cannot see future data, $(X_j)_{j>i}$. Our objective is to raise an alarm as soon as possible after time ν, but we wish to avoid prematurely raising the alarm before time ν.

While hypothesis testing used a *function* T, here we employ a *stopping rule* τ. A *false alarm* is the event where τ detects the change-point prematurely: $\tau < \nu$. In hypothesis testing, we considered the probability of type I error; in this setting, for any non-trivial stopping rule, the probability of a false alarm depends upon ν and can be made arbitrarily close to one (zero) for high (low) values of ν, so we instead consider the *mean time to false alarm*, a term which we shall use interchangeably with *average run length*. This quantity is given by:

$$\mathrm{ARL}(\tau) = \mathbb{E}_\infty[\tau]. \tag{11}$$

Most non-trivial stopping rules will accumulate evidence of a change for some duration of time after ν. In this setting, we consider the mean time to alarm following ν rather than the probability of instantaneous

detection. We define the worst-case average detection delay (Lorden (1971)), which we shall also refer to as expected detection delay, by

$$\mathcal{L}_{\texttt{WADD}}(\tau) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu[(\tau - \nu + 1)^+ | \mathcal{F}_{\nu-1}], \tag{12}$$

where $\mathbb{E}_\nu$ refers to the expectation over data streams following (10) for a particular ν. We seek to minimize the expected detection delay subject to a constraint on acceptable average run length $\overline{\gamma}$, as in Wu et al. (2023). Thus, we wish to find a stopping rule that is a solution of (or at least *approximates* a solution of):

$$\min_{\tau} \ \mathcal{L}_{\texttt{WADD}}(\tau) \ \text{ subject to } \ \text{ARL}(\tau) \geq \overline{\gamma}. \tag{13}$$

Next, we consider one likelihood-based and one score-based solution to change-point detection.

3.2.1 Cumulative Sum (CUSUM) Algorithm

The Cumulative Sum (CUSUM) algorithm (Page (1955); Lorden (1971); Moustakides (1986)) is a popular solution to the problem of change-point detection which is optimal under the objective of (13). The algorithm defines an *instantaneous detection score*, $Z_{\texttt{KL}}(X)$, as presented in (8). It further defines a *cumulative detection score* $Y_{\texttt{KL}}(t)$ and *stopping rule* $\tau^c_{\texttt{KL}}$ by

$$Y_{\texttt{KL}}(t) = \begin{cases} 0 & t = 0 \\ \max(0, Y_{\texttt{KL}}(t-1) + Z_{\texttt{KL}}(X_t)) & t > 0, \end{cases} \qquad \tau^c_{\texttt{KL}} = \min\{t \in \mathbb{N} : Y_{\texttt{KL}}(t) \geq c\} \tag{14}$$

for some choice of threshold $c \geq 0$.

3.2.2 Score-Based Cumulative Sum (SCUSUM) Algorithm

The Score-Based Cumulative Sum (SCUSUM) algorithm (Wu et al. (2023)) replaces log-likelihood ratios by differences of Hyvärinen scores. With instantaneous detection score $Z_{\texttt{F}}(X)$ defined as in (9), the SCUSUM algorithm defines a cumulative detection score $Y_{\texttt{F}}(t)$ and a stopping rule $\tau^c_{\texttt{F}}$ as

$$Y_{\texttt{F}}(t) = \begin{cases} 0 & t = 0 \\ \max(0, Y_{\texttt{F}}(t-1) + Z_{\texttt{F}}(X_t)) & t > 0, \end{cases} \qquad \tau^c_{\texttt{F}} = \min\{t \in \mathbb{N} : Y_{\texttt{F}}(t) \geq c\} \tag{15}$$

for some choice of stopping threshold $c > 0$.

Remark 3.3. *The presentation of SCUSUM in Wu et al. (2023) includes a scaling factor $\lambda > 0$ in the definition of $Z_{\texttt{F}}(X)$. For the limited purposes of this presentation, we can omit λ without loss of generality.*

The pre-change drift $\mathbb{E}_\infty[Z_{\texttt{F}}(X)]$ and post-change drift $\mathbb{E}_1[Z_{\texttt{F}}(X)]$ were calculated in Wu et al. (2023), where it was further demonstrated that the pre-change drift is strictly negative and the post-change drift strictly positive under certain conditions. Furthermore, Wu et al. (2023) provided theorems which bound the mean time to false alarm (11) and detection delay (12) for any given choice of P_∞, P_1.

4 Diffusion-Based Detection

We have presented the Fisher divergence as well as a score-based test and stopping rule. In this section, we first define the diffusion divergence, which generalizes the Fisher divergence. We next propose a diffusion-based test and stopping rule, generalizing the score-based algorithms introduced in Section 2. We conclude by providing theorems which bound the properties of the proposed diffusion-based test and algorithm.

Though the score-based methods are special cases of the diffusion-based ones (and though the diffusion-based methods still utilize the score $\nabla \log p(X)$), we shall refer to the Fisher divergence-based methods as *score-based* and the methods utilizing $m(X)$ as *diffusion-based* for clarity.

4.1 Diffusion Divergence

A generalization of the Fisher divergence was studied in Barp et al. (2022) and applied to change-point detection in Altamirano et al. (2023). This generalization, called the *diffusion divergence*, transforms the difference between scores in Definition 2.1 by multiplication with some matrix-valued function $m(X) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times w}$:

Definition 4.1 (Diffusion Divergence)**.** *The diffusion divergence between distributions P and Q is defined to be:*

$$\mathbb{D}_m(P\|Q) = \mathbb{E}_{X\sim P}\left[\tfrac{1}{2}\left\|m^T(X)\big(\nabla_X \log p(X) - \nabla_X \log q(X)\big)\right\|_2^2\right]$$

for some matrix-valued function $m(X) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times w}$.

We impose conditions upon $m(X)$ in Section 4. For ease of notation, we denote the *function* $m(X)$ simply as m in the subscript of $\mathbb{D}$.

We next define the *diffusion-Hyvärinen score*, which generalizes the Hyvärinen score of Definition 2.2.

Definition 4.2 (Diffusion-Hyvärinen Score)**.** *Denoting the divergence of a function $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ with Jacobian J_f as $\nabla \cdot f(X) = tr(J_f) = \sum_i \partial f_i(X)/\partial X_i$, we define*

$$\mathcal{S}_m(X, P) = \tfrac{1}{2}\left\|m^T(X)(\nabla_X \log p(X))\right\|_2^2 + \nabla \cdot m(X)m^T(X)\nabla \log p(X),$$

We conclude this section by presenting an important identity:

Lemma 4.3 (Diffusion Drifts)**.** *For distributions P, Q:*

$$\mathbb{E}_P[\mathcal{S}_m(X, P) - \mathcal{S}_m(X, Q)] = -\mathbb{D}_m(P\|Q), \qquad \mathbb{E}_Q[\mathcal{S}_m(X, P) - \mathcal{S}_m(X, Q)] = \ \mathbb{D}_m(Q\|P). \tag{16}$$

Proof. The proof is given in Appendix B. □

4.2 Diffusion-based Detection

We begin by imposing mild regularity conditions on $m(X)$, which mirror the regularity conditions presented in Hyvärinen (2005). We shall assume Assumption 4.4 everywhere in this paper.

Assumption 4.4 (Diffusion-Hyvärinen Regularity Conditions Hyvärinen (2005))**.** *For $(R, S) \in \{(P_\infty, P_1), (P_1, P_\infty)\}$, we assume that:*

1. *$m(X)$ and $\nabla \log s(X)$ are differentiable with continuous derivatives,*
2. *$\mathbb{E}_R[\|m^T(X)\nabla \log s(X)\|^2] < \infty$ and $\mathbb{E}_R[\|m^T(X)\nabla \log r(X)\|^2] < \infty$, and*
3. *$s(X)m(X)m^T(X)\nabla \log r(X) \to 0$ as $\|X\| \to \infty$.*

Lemma 4.5 (Finite Divergences)**.** *Part 2 of Assumption 4.4 implies that $\mathbb{D}_m(P_\infty\|P_1) < \infty$ and $\mathbb{D}_m(P_1\|P_\infty) < \infty$.*

Proof. The proof is given in Appendix B. □

As an expectation over a norm, the diffusion divergence is strictly nonnegative for all distributions. We next provide an assumption that the diffusion divergence is nonzero, which we shall assume everywhere this paper:

Assumption 4.6 (Positive Divergences)**.** *We assume that $m(X)$ is chosen such that $\mathbb{D}_m(P_\infty\|P_1) > 0$ and $\mathbb{D}_m(P_1\|P_\infty) > 0$.*

Remark 4.7. *The work of Barp et al. (2022) uses the diffusion divergence to generalize score matching, which requires the diffusion divergence to satisfy $\mathbb{D}_m(P\|Q) = 0 \iff P = Q$ for all P, Q. For noninvertible $m(X)$, there might exist P, Q for which $\nabla \log p(X) - \nabla \log q(X)$ lies in the kernel of $m(X)$ almost everywhere (causing $\mathbb{D}_m(P\|Q) = 0$ for $P \neq Q$), and hence Barp et al. (2022) imposed a condition of invertibility on $m(X)$ for all $X \in \mathbb{R}^d$. Throughout this paper, all diffusion divergences take arguments only from $\{P_\infty, P_1\}$. Therefore, we require only Assumption 4.6 and can therefore consider $m(X)$ which is not invertible. Hence, Assumption 4.6 permits $m(X)$ to be non-square ($w \neq d$) and the theorems of this paper allow arbitrary w.*

Next, we generalize the test statistic of Wu et al. (2022) and the instantaneous detection score of Wu et al. (2023) to include $m(X)$.

Definition 4.8 (Diffusion Test Statistic and Instantaneous Detection Score)**.** *For any choice of $m(X)$ that satisfies Assumption 4.4, and with $\mathcal{S}_m(\cdot, P)$ following Definition 4.2, we define*

$$Z_m(X) = \mathcal{S}_m(X, P_\infty) - \mathcal{S}_m(X, P_1), \tag{17}$$

Note that $Z_m(\cdot)$ is a function of its subscript unlike $Z_{\mathtt{KL}}(\cdot)$ and $Z_{\mathtt{F}}(\cdot)$. We next propose a diffusion-based hypothesis test.

Definition 4.9 (Diffusion Hypothesis Test)**.** *With some fixed choice of stopping threshold $c \in \mathbb{R}$, we define*

$$T_m^c((X_i)_{i=1}^n) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Z_m(X_i) < c \\ 1 & \text{else} \end{cases} \tag{18}$$

We further propose the diffusion change-point detection stopping rule.

Definition 4.10 (Diffusion Change-Point Detection Stopping Rule)**.** *With some fixed choice of stopping threshold $c > 0$, we define*

$$Y_m(t) = \begin{cases} 0 & t = 0 \\ \max(0, Y_m(t-1) + Z_m(X_t)) & t > 0 \end{cases} \quad \text{and} \quad \tau_m^c = \min\{t \in \mathbb{N} : Y_m(t) \geq c\} \tag{19}$$

4.3 Properties of the Proposed Hypothesis Test

We next present a theorem which bounds the performance of the diffusion hypothesis test.

Theorem 4.11 (Error Exponent)**.** *Fix P_∞, P_1, and $\overline{\alpha} \in (0, 1)$. Let $m(X) : \mathbb{R}^d \mapsto \mathbb{R}^{d \times w}$ be some diffusion matrix function which satisfies both Assumption 4.4 and $\mathbb{E}_1[\exp(-Z_m(X))] = 1$. Recall T_m^c from Definition 4.9 and $\mathcal{B}(\cdot)$ from Definition 3.2.*

Then, there exists some c such that (i) $\lim_{n\to\infty} \alpha_n(T_m^c) \leq \overline{\alpha}$ and (ii) $\mathcal{B}(T_m^c) \geq \mathbb{D}_m(P_\infty\|P_1)$.

Proof. The proof is given in Appendix B. □

4.4 Properties of the Proposed Stopping Rule

We next present theorems quantifying the mean time to false alarm and detection delay of the diffusion change-point detection stopping rule, which are generalizations of corresponding results for the Fisher divergence presented in Wu et al. (2023).

Theorem 4.12 (Average Run Length)**.** *For τ_m^c following Definition 4.10, if $m(X)$ satisfies $\mathbb{E}_\infty[\exp Z_m(X)] \leq 1$, then $\mathbb{E}_\infty[\tau_m^c] \geq e^c$.*

Proof. The proof is given in Appendix B. □

Theorem 4.13 (Detection Delay)**.** *For τ_m^c following Definition 4.10, the worst-case average detection delay is given by*

$$\mathcal{L}_{\mathtt{WADD}}(\tau_m^c) \sim \frac{c}{\mathbb{D}_m(P_1\|P_\infty)} \text{ as } c \to \infty, \tag{20}$$

where $f(c) \sim g(c)$ as $c \to \infty$ means that $\lim_{c\to\infty} f(c)/g(c) = 1$.

Proof. The proof is given in Appendix B. □

For each of our detection tasks, we have generalized one unique score-based detection algorithm into a collection of diffusion-based algorithms, each of which is different in both its choice of $m(X)$ and in its performance. We next turn our attention to making a good choice of diffusion matrix function so that our detection algorithms can perform well.

5 Optimization of the Diffusion-Matrix Function

In this section, we propose a method of finding the optimal $m(X)$.

5.1 Objective Function: Change-Point Detection

Under the objective presented in (13), we wish to minimize the detection delay (12) subject to a constraint on the average run length (11). Denoting our constraint on the average run length as $\overline{\gamma}$, we can set a stopping threshold to $c = \log \overline{\gamma}$; by Theorem 4.12 the constraint of (13) (that $\text{ARL}(\tau) \geq \overline{\gamma}$) will be satisfied so long as $\mathbb{E}_\infty[\exp Z_m(X)] \leq 1$.

The expected detection delay is a function of stopping threshold and $\mathbb{D}_m(P_1 \| P_\infty)$. Since we have already set the stopping threshold $c = \log \overline{\gamma}$, we minimize (20) by maximizing $\mathbb{D}_m(P_1 \| P_\infty)$. This produces the objective:

$$\max_m \ \mathbb{D}_m(P_1 \| P_\infty) \ \text{ s.t. } \mathbb{E}_\infty[\exp Z_m(X)] = 1. \tag{21}$$

For any $m(X)$ satisfying Assumption 4.4, and any $k > 0$, it is easy to see that $Z_{km}(X) = k^2 Z_m(X)$ for all $X \in \mathbb{R}^d$ and that $\mathbb{D}_{km}(P_1 \| P_\infty) = k^2 \mathbb{D}_m(P_1 \| P_\infty)$. Thus, if we were to optimize by simply maximizing $\mathbb{D}_m(P_1 \| P_\infty)$, then for any $m(X)$, $\mathbb{D}_{km}(P_1 \| P_\infty) > \mathbb{D}_m(P_1 \| P_\infty)$ whenever $k > 1$. Thus, some constraint is necessary for a maximum of (21) to exist. Furthermore, in change-point detection (hypothesis testing), a stopping rule (test) using instantaneous detection score (statistic) $Z_m(\cdot)$ and threshold k is identical to another stopping rule (test) using instantaneous detection score (statistic) $k^2 Z_m(\cdot)$ and threshold $k^2 c$ – adjusting the scale k does not actually change the algorithm (scale-invariance over the densities of P_∞, P_1 is an advantage of score-based methods but scale-invariance over m frustrates this optimization). Optimization over only $\mathbb{D}_m(P_1 \| P_\infty)$ may simply lead to a larger scaling on $m(X)$, and this optimization would neither improve the test nor converge to any limit. Thus, some constraint on the scale of $m(X)$ is necessary, and the condition $\mathbb{E}_\infty[\exp Z_m(X)] \leq 1$ plays this role. Though the constraint was introduced as a proof technique, it is in fact essential to the optimization process.

5.2 Objective Function: Hypothesis Testing

We wish to find a diffusion hypothesis test between P_∞, P_1 which performs reasonably well *for all batch sizes* n. The popular objective function of (6), however, is explicitly a function of a particular batch size n. We modify the objective of (6) to better account for the variable nature of the batch size:

$$\max_T \ \mathcal{B}(T) \ \text{ subject to } \ \lim_{n \to \infty} \alpha_n(T) \leq \overline{\alpha}, \tag{22}$$

where $\mathcal{B}(\cdot)$ is defined by Definition 3.2. Though the objective of (22) considers the performance of a test for large n, it does not depend upon any one particular choice of n.

For any $m(X)$ following Assumption 4.4 and satisfying $\mathbb{E}_1[\exp(-Z_m(X))] \leq 1$, Theorem 4.11 provides that $\mathbb{D}_m(P_\infty \| P_1)$ is a lower-bound on the error exponent of some particular T_m^c which obeys *any constraint* on asymptotic type I error probability. For any choice of $\overline{\alpha}$, we maximize the type II error exponent of a test by maximizing its lower bound, $\mathbb{D}_m(P_\infty \| P_1)$, while enforcing $\mathbb{E}_1[\exp(-Z_m(X))] = 1$ to satisfy the condition of the Theorem 4.11.

Thus, we propose that the following objective optimizes $m(X)$ for the problem of hypothesis testing:

$$\max_m \ \mathbb{D}_m(P_\infty \| P_1) \ \text{ s.t. } \mathbb{E}_1[\exp(-Z_m(X))] = 1. \tag{23}$$

Remark 5.1. *Recalling Definition 4.8, we observe that the objective of* (21) *takes the form of* (23) *when the roles of* P_∞, P_1 *are interchanged.*

5.3 Optimality of Diffusion-Based Hypothesis Testing and Change-Point Detection

We first demonstrate that for a special case of two Gaussian distributions with a common covariance matrix, $Z_{\mathtt{KL}}(\cdot) = Z_m(\cdot)$ for a particular choice of $m(X)$:

Theorem 5.2. *Consider vectors* $\mu_\infty, \mu_1 \in \mathbb{R}^d$ *and positive definite matrix* $V \in \mathbb{R}^{d\times d}$*. Let* $P_\infty = \mathcal{N}(\mu_\infty, V)$ *and* $P_1 = \mathcal{N}(\mu_1, V)$*, and define* $M^* = V^{\frac{1}{2}}$*. Then,* $\forall X \in \mathbb{R}^d$*,* $Z_{M^*}(X) = Z_{\mathtt{KL}}(X)$*.*

Proof. The proof is given in Appendix B. □

In general, we do not expect any particular choice of $m(X)$ to exactly recover the LLR-based tests. We demonstrate using another pair of Gaussians that there does not always exist an $m(X)$ such that $Z_m(\cdot) = Z_{\mathtt{KL}}(\cdot)$ with probability one.

Theorem 5.3. *There exists a pair of distributions* P_∞, P_1 *such that for all* $m(X)$*,* $\mathbb{P}_\infty[Z_m(X) \neq Z_{\mathtt{KL}}(X)] > 0$ *and* $\mathbb{P}_1[Z_m(X) \neq Z_{\mathtt{KL}}(X)] > 0$*.*

Proof. The proof is given in Appendix B. □

Remark 5.4. *For any* P_∞, P_1*,* $Z_m(\cdot) = Z_{\mathtt{F}}(\cdot)$ *if* $m(X) = I$*.*

It is possible to bound the performance of the *best* diffusion-based algorithm – the algorithm corresponding to the *best* choice of $m(X)$ – in terms of the performance of the score-based and LLR-based algorithms. By Remark 5.4, we observe that the best diffusion-based algorithms can perform no worse than their score-based peers (though diffusion algorithms corresponding to a poorly-chosen $m(X)$ can perform arbitrarily poorly). The diffusion-based algorithms can never outperform their LLR-based counterparts, however, as the latter are provably optimal under (6) and (13).

While we know that CUSUM (14) is optimal under the metric of (13), in general there can be stopping rules distinct from CUSUM which match the performance of CUSUM. Therefore, while for some choices of P_∞, P_1 we do not know whether LLR-like performance is attainable (such as for the P_∞, P_1 of Theorem 5.3), we do know that this optimal performance is attainable for other choices of P_∞, P_1 (such as those of Theorem 5.2).

In general, and especially for high-dimensional P_∞, P_1, the best possible $m(X)$ will not be analytically identifiable. In the general case, we simply wish to find an $m(X)$ that approximates a solution to (21), (23). In the next section, we turn to machine learning to help with this approximation.

5.4 Numerical Optimization

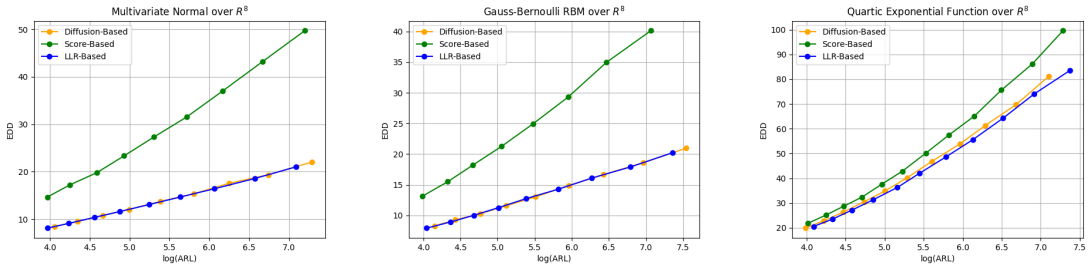
We propose a multi-layer perceptron (MLP) network $m : \mathbb{R}^d \mapsto \mathbb{R}^{d\times w}$ to approximate the best possible matrix-valued function for a given P_∞, P_1. The MLP $m(X)$ can be trained to approximate a solution of (21) or (23).

While we wish to optimize the MLP via gradient descent, the method of gradient descent requires differentiability of the loss function. The constraints ($\mathbb{E}_\infty[\exp Z_m(X)] = 1$ of (21) and $\mathbb{E}_1[\exp(-Z_m(X))] = 1$ of (23)) preclude gradient descent from being applied directly to these objectives. We turn to the use of regularization to circumvent this issue. We propose the loss functions $\mathcal{L}_{\mathtt{CPD}}$ and $\mathcal{L}_{\mathtt{HT}}$ to numerically optimize $m(X)$ for change-point detection and hypothesis testing, respectively:

$$\mathcal{L}_{\mathtt{CPD}}(m) = -\mathbb{D}_m(P_1\|P_\infty) + \alpha\big[\log \mathbb{E}_\infty(\exp(Z_m(X)))\big]^2, \tag{24}$$

$$\mathcal{L}_{\mathtt{HT}}(m) = -\mathbb{D}_m(P_\infty\|P_1) + \alpha\big[\log \mathbb{E}_1(\exp(-Z_m(X)))\big]^2, \tag{25}$$

where $\alpha > 0$ is a training hyperparameter. The condition $\mathbb{E}_\infty[\exp Z_m(X)] = 1$ of (21) ($\mathbb{E}_1[\exp(-Z_m(X))] = 1$ of (23)) is satisfied if and only if $[\log \mathbb{E}_\infty(\exp Z_m(X))]^2 = 0$ ($[\log \mathbb{E}_1(\exp(-Z_m(X)))]^2 = 0$). Therefore, though



(a) Gaussian Distributions (b) GB-RBM Distributions (c) Quartic Exponential Distributions

Figure 1: Performance of LLR-based, Fisher-based, and diffusion-based Change-Point Detection Stopping Rules. In this simulation, P_∞, P_1 were chosen to be (a) Gaussian Distributions, (b) Gauss-Bernoulli Restricted Boltzmann Machine Distributions, and (c) Quartic Exponential Distributions. ARL and EDD are averaged over 10,000 sample paths.

it does not strictly enforce the exact conditions, gradient descent over $\mathcal{L}_{\texttt{CPD}}(m)$ or $\mathcal{L}_{\texttt{HT}}(m)$ causes the neural network $m(X)$ to approximately satisfy the conditions of (21) or (23).

6 Numerical Simulations

In this section, we implement and simulate the diffusion-based methods and compare their performances against those of their LLR-based and score-based analogs for both hypothesis testing and change-point detection. We begin by introducing three model classes with analytical score functions.

6.1 Model Classes

In this section, we shall use the notation of $p(X)$ to refer to the density of some distribution P, and shall use the notation of $\tilde{p}$ to refer to the *unnormalized* density of some distribution P. Two of our three model classes provide only an unnormalized density.

For each model class below, the simulations are performed with parameters chosen in light of (Wu et al., 2023, Proposition 5) to demonstrate cases where score-based methods noticeably underperform LLR-based methods. We provide details and parameters for our model classes in Appendix A.

6.1.1 Gaussian Distribution

The multivariate Gaussian distribution is parameterized by a vector $\mu \in \mathbb{R}^d$ and a positive-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$. Its density and score function are given by:

$$p(X) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\big(-\tfrac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\big), \quad \nabla \log p(X) = -\Sigma^{-1}(X-\mu). \tag{26}$$

6.1.2 Gauss-Bernoulli Restricted Boltzmann Machine

The Gauss-Bernoulli Restricted Boltzmann Machine (GB-RBM) (Liao et al. (2022)) is simultaneously an energy-based model and a score-based model. Thus, it is a natural candidate for comparisons between LLR-based and score-based algorithms.

As presented in (Liao et al., 2022, Equation 3), the conditional density of visible variable conditioned upon latent variable is a Gaussian distribution with diagonal covariance; here, we relax this condition and require only that it be positive definite. For this formulation of the GB-RBM with latent dimension h, the model is parameterized by vectors $\mu \in \mathbb{R}^d, \phi \in \mathbb{R}^h, W \in \mathbb{R}^{d\times h}$, and positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$.

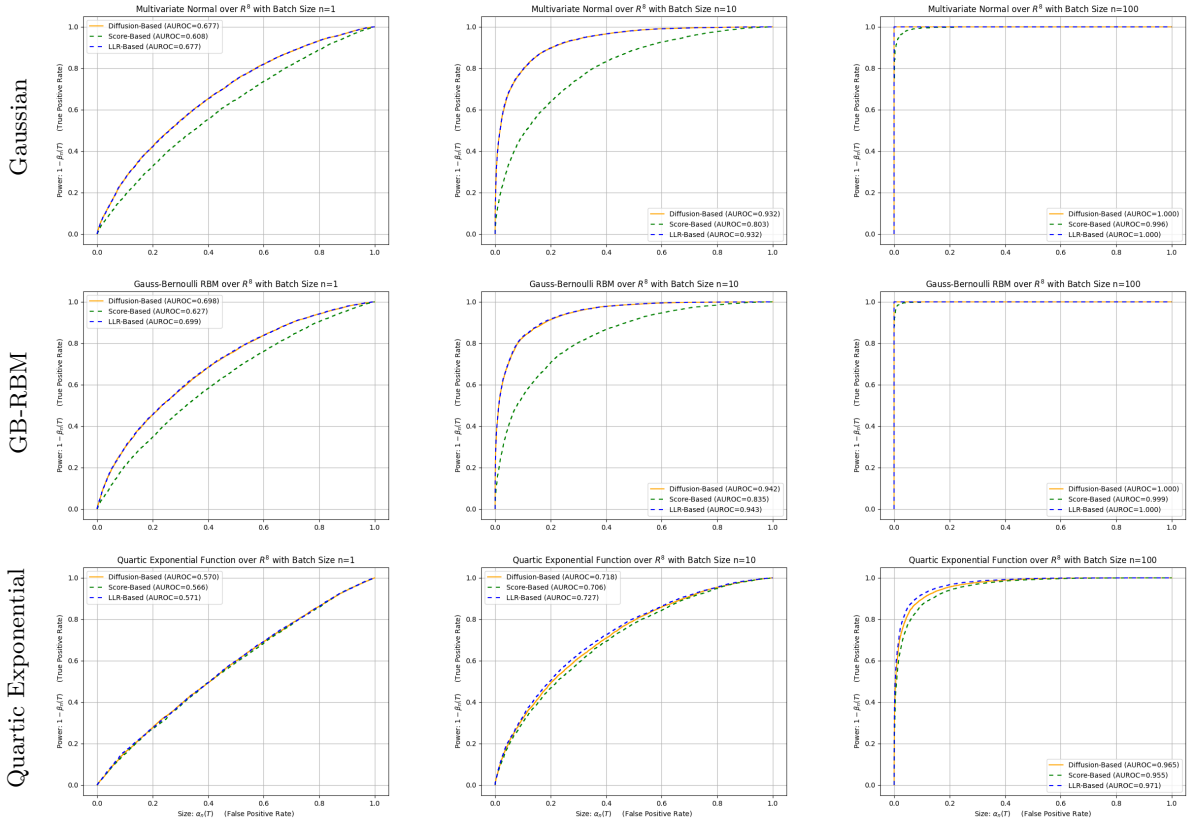


Figure 2: ROC curves for diffusion-based, score-based, and LLR-based hypothesis tests, plotted for batch sizes $n \in \{1, 10, 100\}$. P_∞, P_1 are chosen to follow Gaussian Distributions (top), Gauss-Bernoulli Restricted Boltzmann Machine Distributions (middle), and Quartic Exponential Distributions (bottom). Each ROC curve uses 10,000 batches. Similar plots for $n \in \{5, 25, 50\}$ are shown in Figure 3 in the Appendix.

We provide the free-energy F, unnormalized density $\tilde{p}$, and score $\nabla \log \tilde{p}$ of the GB-RBM:

$$\tilde{p}(X) = \exp(-F(X)), \quad \nabla \log \tilde{p}(X) = \nabla \log \tilde{p}(X) = -\Sigma^{-1}(X-\mu) + \Sigma^{-1} W \ \text{Sigmoid}(\phi + W^T \Sigma^{-1} X) \quad (27)$$
$$\text{where } F(X) = \tfrac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu) - \mathbf{1}^T \text{Softplus}(\phi + W^T \Sigma^{-1} X),$$

6.1.3 Quartic Exponential Model

We finally consider an exponential model class that is quartic (fourth-order) in X. This model is parameterized by $\mu \in \mathbb{R}^d$ and positive-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. The unnormalized density and score function of this model are:

$$\tilde{p}(X) = \exp\big(-(X^{\odot 2} - \mu)^T \Sigma^{-1} (X^{\odot 2} - \mu)\big), \quad \nabla \log \tilde{p}(X) = -4\Sigma^{-1}(X^{\odot 2} - \mu) \odot X, \quad (28)$$

where $\odot$ denotes element-wise multiplication and $X^{\odot 2}$ denotes element-wise squaring.

6.2 Implementation of Algorithms

6.2.1 LLR-Based Algorithms

For model classes that provide normalized densities, we implement the LLR-based algorithms following their description in Section 2. For model classes involving unnormalized densities, we must modify the test statistic and instantaneous detection score $Z_{\text{KL}}(\cdot)$ to account for the lack of normalization.

We let $\mathcal{R}^n_{\tilde{P}_\infty,\tilde{P}_1}$ be a Monte-Carlo estimate of $\mathbb{E}_1[\tilde{p}_\infty(x)/\tilde{p}_1(x)]$ using n samples (Section A.3). Define

$$\hat{Z}_{\mathtt{KL}}(X) = \log \frac{\tilde{p}_1(X)}{\tilde{p}_\infty(X)} + \log \mathcal{R}^n_{\tilde{P}_\infty,\tilde{P}_1}. \tag{29}$$

We let $\hat{Z}_{\mathtt{KL}}$ play the role of $Z_{\mathtt{KL}}(X)$ in our simulations. We note that the estimation of $\mathcal{R}^n_{\tilde{P}_\infty,\tilde{P}_1}$ requires a significant quantity n of data and that this estimation becomes generally intractable as the data dimension becomes large. This approximation is discussed further in Section A.3.

6.2.2 Score-Based Algorithms

Score-based algorithms follow the implementation given in Section 2. We note that for any unnormalized density $\tilde{p}$ with normalizing constant Z, the score-based algorithms are invariant to the scale of $\tilde{p}_\infty(\cdot), \tilde{p}_1(\cdot)$:

$$\nabla \log p(X) = \nabla \log(\tilde{p}(X)/Z) = \nabla \log \tilde{p}(X) - \nabla \log Z = \nabla \log \tilde{p}(X) \tag{30}$$

6.2.3 Diffusion-Based Algorithms

Like the score-based algorithms, diffusion-based algorithms are also scale invariant; hence, the diffusion-based algorithms do not need to account for scaling constants C_∞, C_1.

We simulate square matrix-valued functions. We let a feed-forward neural network play the role of $m(X)$, mapping data in $\mathbb{R}^d$ to matrices in $\mathbb{R}^{d\times d}$. We optimize this choice of $m(X)$ following the loss function of (24). Details regarding the architecture of this neural network and its training are provided in Appendix A.

6.3 Results

For both hypothesis testing and change-point detection, and for each of the three proposed model classes, we compare the performances of the diffusion-based, score-based, and LLR-based algorithms.

In Figure 2, we compare the performances of each method via Receiver Operating Characteristic (ROC) curves, considering the performance of each test under all possible choices of threshold c. We observe that for the Gaussian and GB-RBM distributions, the performance of the diffusion-based hypothesis tests nearly match the performance of their LLR-based counterparts, and that for the quartic exponential distribution, the diffusion-based hypothesis test outperforms the score-based test.

In Figure 1, we compare the performance of each change-point detection stopping rule. Each algorithm is run with many choices of threshold c. We recall the average run length (ARL) from (11), and we define the expected detection delay (EDD) of a stopping rule τ to be $\mathrm{EDD}(\tau) = \mathbb{E}_1[\tau]$. We plot the ARL and EDD for the LLR-based, score-based, and diffusion-based stopping rules. Each curve illustrates the ARL and EDD of an algorithm for several choices of stopping threshold c. Again, we observe that the diffusion-based stopping rule matches the LLR-based one in performance for the Gaussian and GB-RBM simulations and outperforms the score-based algorithm for the simulation involving the quartic exponential distribution.

Across all three model classes and for each of our two detection tasks, we observe that the diffusion-based algorithms perform no worse than the score-based methods and no better than the LLR-based methods, and that they occasionally match the LLR-based methods in performance.

7 Conclusion

In this paper, we have proposed a hypothesis test and a change-point detection stopping rule which utilize the diffusion divergence. We have studied the properties of these methods, calculating and bounding their performance metrics. We have developed an objective over the diffusion matrix function for hypothesis testing and change-point detection applications. We have demonstrated that the performances of the best-possible diffusion algorithms are no worse than the performances of the score-based algorithms and no better than the performances of the LLR algorithms. We have proposed a loss function for the training of $m(X)$ and demonstrated the stability of this process by way of simulation. Finally, we have demonstrated the implementability of the diffusion-based algorithms in numerical simulations.

References

Ahmed Aloui, Ali Hasan, Juncheng Dong, Zihao Wu, and Vahid Tarokh. Score-based metropolis-hastings algorithms, 2025. URL https://arxiv.org/abs/2501.00467.

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and scalable bayesian online changepoint detection, 2023. URL https://arxiv.org/abs/2302.04759.

Fabrizio Angiulli, Fabio Fassetti, and Luca Ferragina. Reconstruction error-based anomaly detection with few outlying examples. *Neurocomputing*, 675:133002, 2026.

Alessandro Barp, Francois-Xavier Briol, Andrew B. Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators, 2022. URL https://arxiv.org/abs/1906.08283.

Nimrod Berman, Eitan Kosman, Dotan Di Castro, and Omri Azencot. Reviving life on the edge: Joint score-based graph generation of rich edge attributes. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=pxdSm7PW5Q.

S.D. Blostein. Quickest detection of a time-varying change in distribution. *IEEE Transactions on Information Theory*, 37(4):1116–1122, 1991. doi: 10.1109/18.87003.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Joseph L Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

Yu-Chih Huang, Yu-Jui Huang, and Shih-Chun Lin. Asymptotic optimality in byzantine distributed quickest change detection. *IEEE Transactions on Information Theory*, 67(9):5942–5962, 2021. doi: 10.1109/TIT.2021.3100423.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6(4), 2005.

Yik Lun Kei, Jialiang Li, Hangjian Li, Yanzhen Chen, and OSCAR HERNAN MADRID PADILLA. Change point detection in dynamic graphs with decoder-only latent space model. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=DVeFqV56Iz.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Marcel Kollovieh, Lukas Gosch, Marten Lienen, Yan Scholten, Leo Schwinn, and Stephan Günnemann. Assessing robustness via score-based adversarial image generation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=7Oqb6zlGWl.

Tze Leung Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theory*, 44(7):2917–2929, 1998.

Renjie Liao, Simon Kornblith, Mengye Ren, David J. Fleet, and Geoffrey Hinton. Gaussian-bernoulli rbms without tears, 2022.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Yong Liu and S.D. Blostein. Quickest detection of an abrupt change in a random sequence with finite change-time. *IEEE Transactions on Information Theory*, 40(6):1985–1993, 1994. doi: 10.1109/18.340471.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Gary Lorden. On excess over the boundary. *Ann. Math. Stat.*, 41(2):520–527, 1970.

Gary Lorden. Procedures for reacting to a change in distribution. *Ann. Math. Stat.*, pp. 1897–1908, 1971.

Savvas Melidonis, Yiming Xi, Konstantinos C. Zygalakis, Yoann Altmann, and Marcelo Pereyra. Score-based denoising diffusion models for photon-starved image restoration problems. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=UYXPt7HUdl.

George V Moustakides. Optimal stopping times for detecting changes in distributions. *Ann. Stat.*, 14(4): 1379–1387, 1986.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009.

ES Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527, 1955.

Md. Saifur Rahman and Aaron B. Wagner. On the optimality of binning for distributed hypothesis testing. *IEEE Transactions on Information Theory*, 58(10):6282–6303, 2012. doi: 10.1109/TIT.2012.2206793.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351 – 367, 2001. doi: 10.1214/ss/1015346320. URL https://doi.org/10.1214/ss/1015346320.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. *CoRR*, abs/1905.07088, 2019. URL http://arxiv.org/abs/1905.07088.

Zhongchang Sun and Shaofeng Zou. Quickest change detection in autoregressive models. *IEEE Transactions on Information Theory*, 70(7):5248–5268, 2024. doi: 10.1109/TIT.2024.3384510.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.

Alexander G Tartakovsky and Venugopal V Veeravalli. General asymptotic bayesian theory of quickest change detection. *Theory of Probability & Its Applications*, 49(3):458–497, 2005.

Harry L. Van Trees. *Classical Detection and Estimation Theory*, chapter 2, pp. 19–165. John Wiley and Sons, Ltd, 2001. ISBN 9780471221081. doi: https://doi.org/10.1002/0471221082.ch2. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0471221082.ch2.

Jayakrishnan Unnikrishnan, Venugopal V Veeravalli, and Sean P Meyn. Minimax robust quickest change detection. *IEEE Transactions on Information Theory*, 57(3):1604–1614, 2011.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23 (7):1661–1674, 2011.

Shun Watanabe. Neyman–pearson test for zero-rate multiterminal hypothesis testing. *IEEE Transactions on Information Theory*, 64(7):4923–4939, 2018. doi: 10.1109/TIT.2017.2778252.

Michael Woodroofe. *Nonlinear renewal theory in sequential analysis*. SIAM, 1982.

Suya Wu, Enmao Diao, Khalil Elkhalil, Jie Ding, and Vahid Tarokh. Score-based hypothesis testing for unnormalized models. *IEEE Access*, 10:71936–71950, 2022.

Suya Wu, Enmao Diao, Taposh Banerjee, Jie Ding, and Vahid Tarokh. Score-based quickest change detection for unnormalized models. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10546–10565. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/wu23b.html.

Y Wu Y Polyanskiy. Hypothesis testing asymptotics i. In *Lecture Notes on Information Theory*. Cambridge MA, 2015. URL https://ocw.mit.edu/courses/6-441-information-theory-spring-2016/5d8f16adc3385c9ff2975b121bd620e4_MIT6_441S16_course_notes.pdf. MIT OpenCourseWare.

Akifumi Yamada, Tomohiro Shiraishi, Shuichi Nishino, Teruyuki Katsuoka, Kouichi Taji, and Ichiro Takeuchi. Change point detection in the frequency domain with statistical reliability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=FNRdaHz3qN.

Appendix

A Details of Numerical Simulations

A.1 Model Class Details

We begin by defining the model classes for which simulation results are provided. We provide V^*, μ_1^*, and μ_∞^* in (36), (37), and (38), respectively.

A.1.1 Gaussian Distribution

We let $P_\infty = \mathcal{N}(\mu_\infty^*, V^*)$ and $P_1 = \mathcal{N}(\mu_1^*, V^*)$.

A.1.2 Gauss-Bernoulli Restricted Boltzmann Machine

We first create $W_\infty \in \mathbb{R}^{8\times 6}$ where each element of W_∞ is drawn i.i.d. from a $\mathcal{N}(0,1)$ distribution, and generate $\phi_\infty \in \mathbb{R}^6$ in the same way. We then sample $W_+ \in \mathbb{R}^{8\times 6}$ and $\phi_+ \in \mathbb{R}^6$ such that each element is drawn i.i.d. from $\mathcal{N}(0, 0.1^2)$. We then let $W_1 = W_\infty + W_+$ and $\phi_1 = \phi_\infty + \phi_+$.

We let P_∞ be the GB-RBM with $\mu = \mu_\infty^*, \Sigma = V^*, W = W_\infty$ and $\phi = \phi_\infty$. We then let P_1 be the GB-RBM with $\mu = \mu_1^*, \Sigma = V^*, W = W_1$, and $\phi = \phi_1$.

A.1.3 Quartic Exponential Distribution

We let P_∞ be parameterized by $\mu = \mu_\infty^*$ and $\Sigma = V^*$. We let P_1 be parameterized by $\mu = \mu_1^*$ and $\Sigma = V^*$.

We note that if $\mu = 0$, the density of (28) can be rewritten as

$$\tilde{p}(X) = \exp\bigg(-\sum_i \sum_j (\Sigma^{-1})_{i,j} X_i^2 X_j^2 \bigg). \tag{31}$$

A.2 Sampling

We sample from P_∞, P_1 to create a training dataset of $100,000$ samples and a test dataset of $10,000$ samples. We sample from all distributions via the Metropolis-Hastings algorithm (Roberts & Rosenthal (2001)) except for the Gaussian distribution, where we perform direct Gaussian sampling. We sample new data for the creation of the curves of Figures 1 and 2.

A.3 Partition Ratio Estimation

Define

$$C_\infty = \int_{\mathbb{R}^d} \tilde{p}_\infty(X) dX \quad \text{and} \quad C_1 = \int_{\mathbb{R}^d} \tilde{p}_1(X) dX. \tag{32}$$

Then, the LLR-based test statistic and instantaneous detection score is

$$Z_{\mathtt{KL}}(X) = \log \frac{p_1(X)}{p_\infty(X)} = \log \frac{\tilde{p}_1(X)}{\tilde{p}_\infty(X)} \frac{C_\infty}{C_1} = \log \frac{\tilde{p}_1(X)}{\tilde{p}_\infty(X)} + \log \frac{C_\infty}{C_1}, \tag{33}$$

which presents a problem, as C_∞/C_1 is not known. It can, however, be calculated from an expectation:

$$\mathbb{E}_1\left[\frac{\tilde{p}_\infty(X)}{\tilde{p}_1(X)}\right] = \mathbb{E}_1\left[\frac{p_\infty(X)}{p_1(X)}\frac{C_\infty}{C_1}\right] = \int_{\mathbb{R}^d} p_1(X)\frac{p_\infty(X)}{p_1(X)}\frac{C_\infty}{C_1} dX = \frac{C_\infty}{C_1}\int p_\infty(X) dX = \frac{C_\infty}{C_1},$$

which we can approximate with a sample mean:

$$\mathcal{R}^n_{\tilde{P}_\infty, \tilde{P}_1} = \frac{1}{n}\sum_{i=1}^n \frac{\tilde{p}_\infty(X_i)}{\tilde{p}_1(X_i)} \tag{34}$$

$$V^* = \begin{bmatrix} 6.94357 & -3.41203 & -2.15460 & -0.48852 & -0.21851 & -0.39300 & -0.93257 & -0.75584 \\ -3.41203 & 3.78724 & 0.5144 & -0.30651 & 1.64793 & 0.06043 & 0.71543 & -1.44385 \\ -2.15460 & 0.5144 & 3.75500 & 2.00786 & -1.22796 & -0.94496 & -2.25916 & 0.8728 \\ -0.48852 & -0.30651 & 2.00786 & 2.93120 & -1.57410 & -1.91590 & -1.7714 & 0.02425 \\ -0.21851 & 1.64793 & -1.22796 & -1.57410 & 5.37965 & 2.21935 & -1.66047 & -2.40907 \\ -0.39300 & 0.06043 & -0.94496 & -1.91590 & 2.21935 & 6.24591 & -0.93225 & 3.02939 \\ -0.93257 & 0.71543 & -2.25916 & -1.7714 & -1.66047 & -0.93225 & 8.12932 & 0.29485 \\ -0.75584 & -1.44385 & 0.87281 & 0.02425 & -2.40907 & 3.02939 & 0.29485 & 6.82808 \end{bmatrix} \tag{36}$$

$$\mu_1^* = \begin{bmatrix} 0, 0, 0, 0, 0, 0, 0, 0 \end{bmatrix}^T \tag{37}$$

$$\mu_\infty^* = \begin{bmatrix} 0.99974, -1.11210, -0.11677, 0.1231, -0.55111, 0.29397, -0.71772, 0.93254 \end{bmatrix}^T \tag{38}$$

for a dataset $(X_i)_{i=1}^n$ of i.i.d. samples of P_1. Thus, for our LLR-based tests, we let

$$\hat{Z}_{\mathtt{KL}}(X) = \log \frac{\tilde{p}_1(X)}{\tilde{p}_\infty(X)} + \log \mathcal{R}^n_{\tilde{P}_\infty, \tilde{P}_1} \tag{35}$$

play the role of $Z_{\mathtt{KL}}(\cdot)$.

A.4 Training

We create a feed-forward neural network with 8-dimensional input, a single 36-dimensional hidden layer, and a 64-dimensional output. We place a Sigmoid activation function after all non-output layers. For each input $X \in \mathbb{R}^8$, we reshape the network's output $\overline{m}(X) \in \mathbb{R}^{64}$ into a matrix $m(X) \in \mathbb{R}^{8\times 8}$ and multiply this output by a constant 0.1, which has been found to sometimes improve the stability of the training process during the first few epochs.

Training was performed via Adam (Kingma & Ba (2015)) with learning rates 0.035 (Quartic Exponential Distribution), 0.04 (Gaussian Distribution), and 0.01 (Gauss-Bernoulli RBM). Across all distributions, training was performed with L2-regularization of $1 \cdot 10^{-5}$ and $\alpha = 10$. Hyperparameters were tuned via inspection of loss history; it should be noted that the competing models (Fisher-Based test, LLR-Based test) have no tunable parameters.

A.5 Additional Plots

We present additional ROC curves in Figure 3.

B Proofs of Theorems and Lemmas

We begin by presenting Lemmas which shall assist in the proofs of the Lemmas and Theorems of this paper.

Lemma B.1. *For completeness, we recall this Lemma and proof from Hyvärinen (2005)[Lemma 4].*

For differentiable $g, h : \mathbb{R}^d \mapsto \mathbb{R}$:

$$\int_{-\infty}^{\infty} h(X) \frac{\partial g(X)}{\partial X_1} dX_1 = \lim_{a\to\infty, b\to-\infty} \Big(g(a, X_2, \cdots, X_n)h(a, X_2, \cdots, X_n) - g(b, X_2, \cdots, X_n)h(b, X_2, \cdots, X_n) \Big) - \int_{-\infty}^{\infty} g(X) \frac{\partial h(X)}{\partial X_1} dX_1 \tag{39}$$

The same follows for all $X_i \neq X_1$.

Proof. By the product rule:

$$\frac{\partial g(X)h(X)}{\partial X_1} = g(X)\frac{\partial h(X)}{\partial X_1} + h(X)\frac{\partial g(X)}{\partial X_1} \tag{40}$$

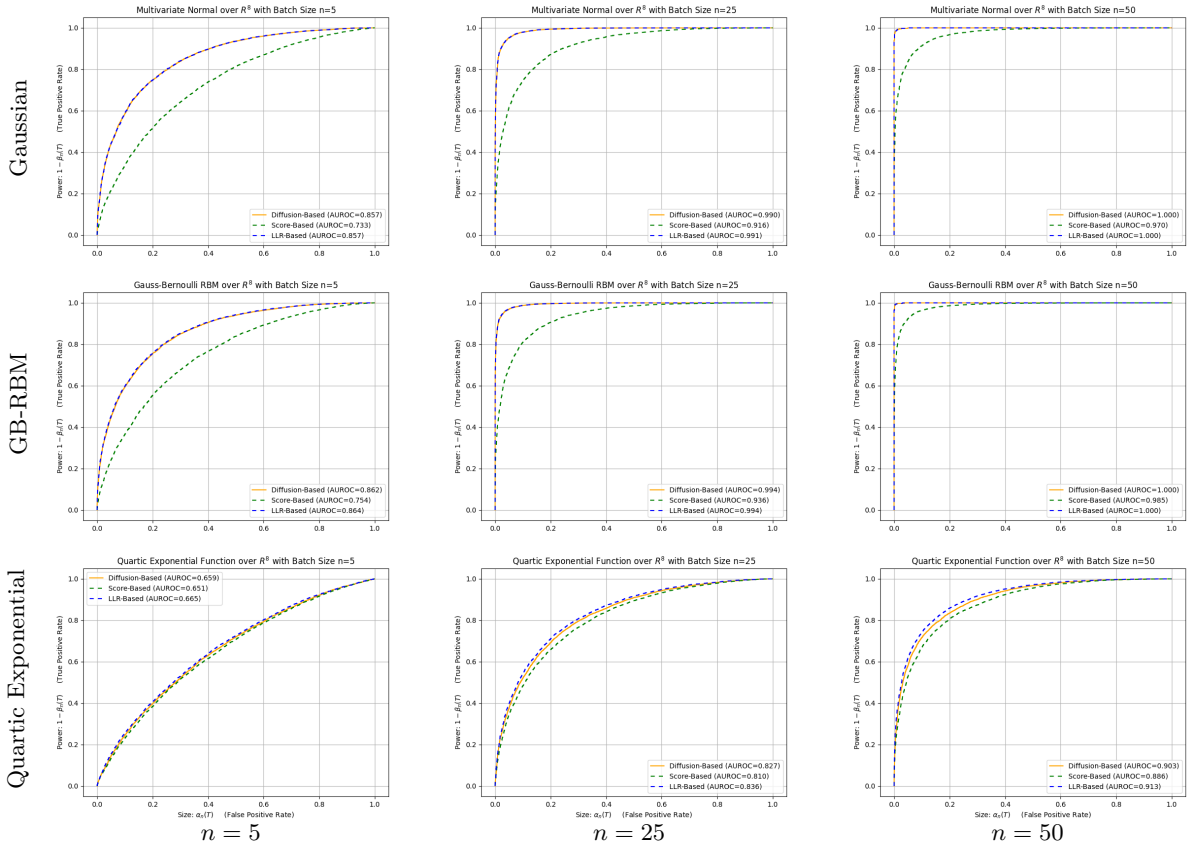


Figure 3: ROC curves for diffusion-based, score-based, and LLR-based hypothesis tests, plotted for batch sizes $n \in \{5, 25, 50\}$. P_∞, P_1 are chosen to follow Gaussian Distributions (top), Gauss-Bernoulli Restricted Boltzmann Machine Distributions (middle), and Quartic Exponential Distributions (bottom). Each ROC curve uses 10,000 batches. Note that similar plots for $n \in \{1, 10, 100\}$ are provided in Figure 2 in the main text.

Rearranging:

$$h(X)\frac{\partial g(X)}{\partial X_1} = \frac{\partial g(X)h(X)}{\partial X_1} - g(X)\frac{\partial h(X)}{\partial X_1} \tag{41}$$

The result follows from integrating both sides of (41) over $\mathbb{R}^d$. □

Lemma B.2. *Let $s(X) : \mathbb{R}^d \mapsto \mathbb{R}^d$ be a differentiable function with continuous derivatives and let P be a probability distribution with density p absolutely continuous with respect to the Lebesgue measure. Let $(v)_i$ denote the i-th element of vector v.*

If

$$\mathbb{E}_P[\|\nabla \log p(X)\|^2] < \infty \quad \textit{and} \quad \mathbb{E}_P[\|s(X)\|^2] < \infty, \tag{42}$$

then

$$\mathbb{E}_P[|(\nabla \log p(X))_i (s(X))_i|] < \infty \tag{43}$$

for all $1 \leq i \leq d$.

Proof. By the definition of the norm, we can say that:

$$\mathbb{E}_P[\|\nabla \log p(X)\|^2] = \mathbb{E}_P\left[\sum_i ((\nabla \log p(X))_i)^2\right] \tag{44}$$

so $\mathbb{E}_P[\|\nabla \log p(X)\|^2] < \infty$ implies that $\mathbb{E}_P[((\nabla \log p(X))_i)^2] < \infty$ for all $1 \leq i \leq d$. By similar reasoning, the assumptions of this Lemma imply that $\mathbb{E}_P[((s(X))_i)^2] < \infty$ for all $1 \leq i \leq d$.

By the Cauchy-Schwarz inequality, we have that:

$$\mathbb{E}_P[|(\nabla \log p(X))_i (s(X))_i|] \leq \sqrt{\mathbb{E}_P[((\nabla \log p(X))_i)^2]} \sqrt{\mathbb{E}_P[((s(X))_i)^2]} \tag{45}$$

and thus the quantity of interest is bounded from above by finite terms.

For the special choice of $s(X) = \nabla \log q(X)$, then Assumption 4.4 implies the assumptions of this Lemma, and this Lemma demonstrates that

$$\mathbb{E}_P[|(\nabla \log p(X))_i (\nabla \log q(X))_i|] < \infty \tag{46}$$

for all $1 \leq i \leq d$. □

Lemma B.3. *For $m(X), P_\infty, P_1$ satisfying Assumptions 4.4 and 3.1, we have that:*

$$\mathbb{E}_P\left[\frac{1}{2}\|m^T(X)\big(\nabla \log p(X) - \nabla \log q(X)\big)\|^2\right] = \mathbb{E}_P\left[\frac{1}{2}\|m^T(X)\nabla \log p(X)\|^2 + \mathcal{S}_m(X, Q)\right] \tag{47}$$

Proof of Lemma B.3. This proof follows the arguments of Hyvärinen (2005)[Theorem 1], with modifications for the present setting. We include the details for completeness. For generality, we prove this theorem for arbitrary $s(X)$, but note that when we let $s(X) = \nabla \log q(X)$, then this theorem calculates a new form for $\mathbb{D}_m(P\|Q)$.

Let $s(X) : \mathbb{R}^d \mapsto \mathbb{R}^d$ be differentiable with a continuous derivative. We calculate:

$$\begin{aligned}
&\mathbb{E}_P\left[\frac{1}{2}\|m^T(X)\big(\nabla \log p(X) - s(X)\big)\|^2\right] \\
&= \mathbb{E}_P\left[\frac{1}{2}\|m^T(X)\nabla \log p(X)\|^2 + \frac{1}{2}\|m^T(X)s(X)\|^2 - \underbrace{(m^T(X)\nabla \log p(X))^T m^T(X)s(X)}_{\text{Term 1}}\right].
\end{aligned} \tag{48}$$

We examine Term 1 in more detail:

$$\mathbb{E}_P[(m^T(X)\nabla \log p(X))^T m^T(X)s(X)] = \mathbb{E}_P[(\nabla \log p(X))^T m(X)m^T(X)s(X)]. \tag{49}$$

We define

$$f(X) = m(X)m^T(X)s(X). \tag{50}$$

Letting v_i and $(v)_i$ denote the i-th element of a vector v, we can simplify the expression of (49):

$$\begin{aligned}
\mathbb{E}_P[(\nabla \log p(X))^T f(X)] &= \int_{\mathbb{R}^d} p(X)(\nabla \log p(X))^T f(X) dX \\
&= \int_{\mathbb{R}^d} p(X) \sum_{i=1}^d (\nabla \log p(X))_i (f(X))_i dX \\
&= \sum_{i=1}^d \int_{\mathbb{R}^d} p(X)(\nabla \log p(X))_i (f(X))_i dX
\end{aligned} \tag{51}$$

We next choose to calculate the integrand of (51) for the case of $i = 1$:

$$\begin{aligned}
\int_{\mathbb{R}^d} p(X)(\nabla \log p(X))_1 (f(X))_1 dX &= \int_{\mathbb{R}^d} p(X) \frac{\partial \log p(X)}{\partial X_1} (f(X))_1 dX \\
&= \int_{\mathbb{R}^d} p(X) \frac{\frac{\partial p(X)}{\partial X_1}}{p(X)} (f(X))_1 dX \\
&= \int_{\mathbb{R}^d} \frac{\partial p(X)}{\partial X_1} (f(X))_1 dX
\end{aligned} \tag{52}$$

By the result of Lemma B.2, we can invoke Fubini's Theorem to expand the integral of (52) into a double integral:

$$\int_{\mathbb{R}^d} \frac{\partial p(X)}{\partial X_1}(f(X))_1 dX = \int_{\mathbb{R}^{d-1}} \underbrace{\left(\int_{-\infty}^{\infty} \frac{\partial p(X)}{\partial X_1}(f(X))_1 dX_1 \right)}_{\text{Term 2}} d(X_2, \cdots X_n). \tag{53}$$

Next, we apply Lemma B.1 to Term 2, letting $p(X)$ play the role of $g(X)$ and $(f(X))_1$ play the role of $h(X)$. The limit of Lemma B.1 evaluates to zero by Part 3 of Assumption 4.4.

$$\int_{\mathbb{R}^{d-1}} \left(0 - \int_{-\infty}^{\infty} p(X) \frac{\partial (f(X))_1}{\partial X_1} dX_1 \right) d(X_2, \cdots, X_n) = - \int_{\mathbb{R}^d} p(X) \frac{\partial (f(X))_1}{\partial X_1} dX \tag{54}$$

where in the last step we collapse the double integral into one single integral over X.

Altogether, we have that:

$$\begin{aligned}
\mathbb{E}_P[(\nabla \log p(X))^T f(X)] &= \sum_{i=1}^{d} \int_{\mathbb{R}^d} p(X)(\nabla \log p(X))_i (f(X))_i dX \\
&= - \sum_{i=1}^{d} \int_{\mathbb{R}^d} p(X) \frac{\partial (f(X))_i}{\partial X_i} dX \\
&= - \int_{\mathbb{R}^d} p(X) \sum_{i=1}^{d} \frac{\partial (f(X))_i}{\partial X_i} dX \\
&= -\mathbb{E}_P[\nabla \cdot f(X)]
\end{aligned} \tag{55}$$

Returning to (48), we substitute and arrive at:

$$\begin{aligned}
&\mathbb{E}_P \left[\frac{1}{2} \| m^T(X)(\nabla \log p(X) - s(X)) \|^2 \right] \\
&= \mathbb{E}_P \left[\frac{1}{2} \| m^T(X) \nabla \log p(X) \|^2 + \frac{1}{2} \| m^T(X) s(X) \|^2 + \nabla \cdot m(X) m^T(X) s(X) \right].
\end{aligned} \tag{56}$$

For the special case where $s(X) = \nabla \log q(X)$, we have that:

$$\begin{aligned}
\mathbb{D}_m(P \| Q) &= \mathbb{E}_P \left[\frac{1}{2} \| m^T(X)(\nabla \log p(X) - \nabla \log q(X)) \|^2 \right] \\
&= \mathbb{E}_P \left[\frac{1}{2} \| m^T(X) \nabla \log p(X) \|^2 + \frac{1}{2} \| m^T(X) \nabla \log q(X) \|^2 + \nabla \cdot m(X) m^T(X) \nabla \log q(X) \right] \\
&= \mathbb{E}_P \left[\frac{1}{2} \| m^T(X) \nabla \log p(X) \|^2 + \mathcal{S}_m(x, Q) \right]
\end{aligned} \tag{57}$$

□

Proof of Lemma 2.3. The assumptions of this Lemma are sufficient to guarantee Assumption 4.4 when $m(X) = I$, and the proof follows from the result of Lemma B.3 when $m(X) = I$. □

Proof of Lemma 4.3. This proof closely follows the arguments presented in (Wu et al., 2023, Lemma 1), with modification for the present setting. We include the details for completeness.

We know that

$$\mathbb{D}_m(P \| Q) = \mathbb{E}_P \left[\frac{1}{2} \| m^T(X) \nabla \log p(X) \|^2 + \mathcal{S}_m(X, Q) \right] \tag{58}$$

We observe that the first term inside the expectation does not depend upon Q. We denote

$$C_m(R) = \mathbb{E}_R\left[\frac{1}{2}\|m^T(X)\nabla \log r(X)\|^2\right] \tag{59}$$

for any distribution R with density r. Then:

$$\mathbb{E}_\infty[\mathcal{S}_m(X, P_\infty) - \mathcal{S}_m(X, P_1)] = \mathbb{D}_m(P_\infty\|P_\infty) - C_m(P_\infty) - \mathbb{D}_m(P_\infty\|P_1) + C_m(P_\infty) = -\mathbb{D}_m(P_\infty\|P_1)$$

and

$$\mathbb{E}_1[\mathcal{S}_m(X, P_\infty) - \mathcal{S}_m(X, P_1)] = \mathbb{D}_m(P_1\|P_\infty) - C_m(P_1) - \mathbb{D}_m(P_1\|P_1) + C_m(P_1) = \mathbb{D}_m(P_1\|P_\infty).$$

□

Proof of Lemma 4.5. Letting

$$\overline{r}(X) = m^T(X)\nabla \log r(X), \quad \overline{s}(X) = m^T(X)\nabla \log s(X),$$

then $\mathbb{D}_m(R\|S) = \frac{1}{2}\mathbb{E}_R[\|\overline{r}(X) - \overline{s}(X)\|^2]$. Then, by the triangle inequality,

$$\|\overline{r}(X) - \overline{s}(X)\| \leq \|\overline{r}(X)\| + \|\overline{s}(X)\|$$

and as the norm is nonnegative,

$$\|\overline{r}(X) - \overline{s}(X)\|^2 \leq (\|\overline{r}(X)\| + \|\overline{s}(X)\|)^2 = \|\overline{r}(X)\|^2 + \|\overline{s}(X)\|^2 + 2\|\overline{r}(X)\|\|\overline{s}(X)\|. \tag{60}$$

Finally,

$$\mathbb{E}_R[\|\overline{r}(X)\|\|\overline{s}(X)\|] \leq \sqrt{\mathbb{E}_R[\|\overline{r}(X)\|^2]\mathbb{E}_R[\|\overline{s}(X)\|^2]}$$

by the Cauchy-Schwarz Inequality. The Lemma follows from Part 2 of Assumption 4.4. □

Proof of Theorem 4.11. This proof follows the arguments of the Stein Diffusion Lemma as presented in (Y Polyanskiy, 2015, Theorem 11.1) (forward direction only).

Fix some arbitrarily small $\delta > 0$ and set $c = n(\delta - \mathbb{D}_m(P_\infty\|P_1))$. We recall Definition 4.9:

$$T_m^c((X_i)_{i=1}^n) = \begin{cases} 0 & \text{if } \sum_{i=1}^n Z_m(X_i) < c \\ 1 & \text{else.} \end{cases} \tag{61}$$

Recalling (5), the type I error probability of T_m^c is given by

$$\alpha_n(T_m^c) = \mathbb{P}_\infty\left[\sum_{i=1}^n Z_m(X_i) \geq c\right]. \tag{62}$$

By the law of large numbers, we know that

$$\frac{1}{n}\sum_{i=1}^n Z_m(X_i)\Big|_{X_i\sim P_\infty} \longrightarrow \mathbb{E}_\infty[Z_m(X)] \tag{63}$$

as $n \to \infty$. we have demonstrated in (16) that

$$\mathbb{E}_\infty[Z_m(X)] = -\mathbb{D}_m(P_\infty\|P_1). \tag{64}$$

For the $\delta > 0$ previously chosen, the definition of convergence guarantees that for all $\epsilon > 0$, there exists some N for which $n > N$ implies that

$$\mathbb{P}_\infty\left[\left|\frac{1}{n}\sum_{i=1}^n Z_m(X_i) - (-\mathbb{D}_m(P_\infty\|P_1))\right| \geq \delta\right] < \epsilon. \tag{65}$$

The inequality of (65) implies that

$$\mathbb{P}_\infty\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_m(X_i) - (-\mathbb{D}_m(P_\infty\|P_1))\right) \geq \delta\right] < \epsilon. \tag{66}$$

and hence that

$$\mathbb{P}_\infty\left[\frac{1}{n}\sum_{i=1}^{n} Z_m(X_i) \geq (\delta - \mathbb{D}_m(P_\infty\|P_1))\right] < \epsilon. \tag{67}$$

Rearranging the inequality and substituting $c = n(\delta - \mathbb{D}_m(P_\infty\|P_1))$, we can say that for all $\epsilon > 0$, there exists an N such that $n > N$ implies:

$$\mathbb{P}_\infty\left[\sum_{i=1}^{n} Z_m(X_i) \geq c\right] < \epsilon. \tag{68}$$

Letting $\epsilon = \overline{\alpha}$ establishes that $\lim_{n\to\infty} \alpha_n(T_m^c) \leq \overline{\alpha}$ for all $\overline{\alpha} \in (0, 1)$.

Next, we calculate the type II error probability. Recalling (5), the type II error probability of T_m^c is given by:

$$\beta_n(T_m^c) = \mathbb{P}_1\left[\sum_{i=1}^{n} Z_m(X_i) < c\right]. \tag{69}$$

We manipulate and apply the Chernoff Bound:

$$\begin{aligned}\mathbb{P}_1\left[\left(-\sum_{i=1}^{n} Z_m(X_i)\right) > (-c)\right] &\leq \inf_{\theta>0} e^{-\theta(-c)} \prod_{i=1}^{n} \mathbb{E}_1[\exp(-\theta Z_m(X_i))] \\ &\leq e^c \prod_{i=1}^{n} \mathbb{E}_1[\exp(-Z_m(X_i))]. \end{aligned} \tag{70}$$

We assume that $\mathbb{E}_1[\exp(-Z_m(X_i))] \leq 1$. Thus:

$$e^c \prod_{i=1}^{n} \mathbb{E}_1[\exp(-Z_m(X_i))] \leq e^c, \tag{71}$$

and

$$\beta_n(T_m^c) \leq e^c = \exp\left(n(\delta - \mathbb{D}_m(P_\infty\|P_1))\right). \tag{72}$$

We have that $\mathcal{B}(T_m^c) \geq \mathbb{D}_m(P_\infty\|P_1) - \delta$, where $\mathcal{B}(\cdot)$ is defined in Definition 3.2. As this result holds for all $\delta > 0$, it follows that $\mathcal{B}(T_m^c) \geq \mathbb{D}_m(P_\infty\|P_1)$. □

Proof of Theorem 4.12. This proof closely follows the arguments of Theorem 3 in (Wu et al., 2023, Theorem 3), with modifications for the present setting. We include the details for completeness. We prove in many steps:

Construction of a Random Walk and Martingale: We first define

$$\delta = -\log(\mathbb{E}_\infty[\exp Z_m(X)]). \tag{73}$$

From the condition of Theorem 4.12, $\delta \geq 0$. Next, define

$$\tilde{Z}_m(X) = Z_m(X) + \delta \tag{74}$$

where $Z_m(X)$ follows Definition 4.8. We further define

$$\tilde{W}_a^b = \sum_{i=a}^{b} \tilde{Z}_m(X_i)$$

$$\tilde{G}(n) = \exp \tilde{W}_1^n = \exp\left(\sum_{i=1}^{n} \tilde{Z}_m(X_i)\right)$$

We observe that $\tilde{W}$ is a random walk that can take negative values. Using Lemma 4.3 and Jensen's Inequality with respect to the convex function $-\log(\cdot)$:

$$\mathbb{D}_m(P_\infty \| P_1) = -\mathbb{E}_\infty[Z_m(X)] > -\log E_\infty[\exp Z_m(X)] = \delta. \tag{75}$$

We present Jensen's Inequality as a strict inequality (it could achieve equality only if $Z_m(X)$ is almost surely constant, but this would imply that $D_m(P_\infty \| P_1) = \mathbb{D}_m(P_1 \| P_\infty) = 0$, violating Assumption 4.6). Using (75), we observe:

$$\mathbb{E}_\infty[\tilde{Z}_m(X)] = \mathbb{E}_\infty[Z_m(X) + \delta] < -\mathbb{D}_m(P_\infty \| P_1) + \mathbb{D}_m(P_\infty \| P_1) = 0. \tag{76}$$

As $\mathbb{E}_\infty[\tilde{Z}_m(X)] < 0$, we note that $\tilde{W}$ is a random walk with a strictly negative drift.

We observe that

$$\mathbb{E}_\infty[\tilde{G}(n+1)|\mathcal{F}_n] = \tilde{G}(n)\mathbb{E}_\infty[\exp(\tilde{Z}_m(X_{n+1}))] = \tilde{G}(n)e^\delta \mathbb{E}_\infty[\exp Z_m(X_{n+1})] = \tilde{G}(n), \tag{77}$$

and that

$$\mathbb{E}_\infty[\tilde{G}(n)] = \mathbb{E}_\infty\left[\exp\left(\sum_{i=1}^{n}(Z_m(X_i) + \delta)\right)\right] = e^{n\delta}\prod_{i=1}^{n}\mathbb{E}_\infty[\exp Z_m(X_i)] = 1, \tag{78}$$

so $\tilde{G}$ is a non-negative martingale with mean one under P_∞.

Construction of Stopping Rule $\tilde{T}$: Next, we define a stopping time:

$$\tilde{T} = \inf\left\{n \geq 1 : \left(\max_{1\leq k\leq n} \tilde{W}_k^n\right) \geq c\right\} \tag{79}$$

We note that $\tilde{T}$ cannot be trivially calculated due to the strictly negative drift of the random walk given by $\tilde{W}_k^n$.

Construction of Stopping Rule M: We define a sequence of stopping times:

$$\eta_0 = 0, \qquad \eta_1 = \inf\left\{t : \tilde{W}_1^t < 0\right\}, \qquad \eta_{k+1} = \inf\left\{t > \eta_k : \tilde{W}_{\eta_k+1}^t < 0\right\},$$

and

$$M = \inf\left\{k \geq 0 : \eta_k < \infty \text{ and } \tilde{W}_{\eta_k+1}^n > c \text{ for some } n > \eta_k\right\}. \tag{80}$$

We can see that $M \leq \tilde{T}$. Since $\tilde{Z}_m(X) \geq Z_m(X)$, we know that $\tilde{T} \leq T$. Hence, $\mathbb{E}_\infty[T] \geq \mathbb{E}_\infty[M]$.

Calculation of $\mathbb{P}_\infty(M > k)$: As an intermediate step in the calculation of $\mathbb{E}_\infty[M]$, we first calculate $\mathbb{P}_\infty(M > k)$:

$$\begin{aligned}\mathbb{P}_\infty[M > k] = \mathbb{E}_\infty[\mathbb{1}_{\{M>k\}}] = \mathbb{E}_\infty[\mathbb{1}_{\{M>k\}}\mathbb{1}_{\{M\geq k\}}] &= \mathbb{E}_\infty[\mathbb{E}_\infty[\mathbb{1}_{\{M>k\}}\mathbb{1}_{\{M\geq k\}}|\mathcal{F}_{\eta_k}]] \\ &= \mathbb{E}_\infty[\mathbb{P}(M \geq k+1|\mathcal{F}_{\eta_k})\mathbb{1}_{\{M\geq k\}}].\end{aligned}$$

We next consider the probability $\mathbb{P}_\infty(M \geq k+1|\mathcal{F}_{\eta_k})$.

$$\mathbb{P}_\infty(M \geq k+1|\mathcal{F}_{\eta_k}) = 1 - \mathbb{P}_\infty(M \leq k|\mathcal{F}_{\eta_k}). \tag{81}$$

We calculate $\mathbb{P}(M \leq k|\mathcal{F}_{\eta_k})$:

$$\begin{aligned}
\mathbb{P}_\infty\left(\tilde{W}^n_{\eta_k+1} > c \text{ for some } n > \eta_k \middle| \mathcal{F}_{\eta_k}\right) &= \mathbb{P}_\infty\left(\tilde{W}^n_1 \text{ for some } n\right) \\
&= \lim_{t\to\infty} \mathbb{P}_\infty\left(\left(\max_{n:n\leq t} \tilde{W}^n_1\right) \geq c\right) \\
&= \lim_{t\to\infty} \mathbb{P}_\infty\left(\left(\max_{n:n\leq t} \exp(\tilde{W}^n_1)\right) > e^c\right) \\
&= \lim_{t\to\infty} \mathbb{P}_\infty\left(\left(\max_{n:n\leq t} \tilde{G}(n)\right) > e^c\right) \\
&\leq \frac{\mathbb{E}_\infty[\tilde{G}(t)]}{e^c} = e^{-c},
\end{aligned}$$

where in the first step we rely upon the i.i.d. (and hence stationary) nature of X_i and in the last step we invoke Doob's submartingale inequality (Doob (1953)), noting that $\tilde{G}(t)$ is a nonnegative martingale with mean one under P_∞.

From (81), we know that $\mathbb{P}(M \geq k+1|\mathcal{F}_{\eta_k}) \geq 1 - e^{-c}$ and hence that

$$\begin{aligned}
\mathbb{P}_\infty[M > k] &= \mathbb{E}_\infty[\mathbb{P}_\infty[M \geq k+1|\mathcal{F}_{\eta_k}]\mathbb{1}_{\{M\geq k\}}] \\
&\geq (1-e^{-c})\mathbb{P}_\infty[M > k-1] \\
&\geq (1-e^{-c})^2\mathbb{P}_\infty[M > k-2] \\
&\geq \cdots \geq (1-e^{-c})^k.
\end{aligned}$$

We next sum a geometric series:

$$\mathbb{E}_\infty[M] = \sum_{k=0}^{\infty} \mathbb{P}(M > k) \geq \sum_{k=0}^{\infty} (1-e^{-c})^k = \frac{1}{1-(1-e^{-c})} = e^c \tag{82}$$

and we conclude that

$$\mathbb{E}_\infty[T] \geq \mathbb{E}_\infty[\tilde{T}] \geq \mathbb{E}_\infty[M] \geq e^c. \tag{83}$$

□

Proof of Theorem 4.13. This proof closely follows the arguments presented in (Wu et al., 2023, Theorem 4), with modifications for the present setting. We include the details for completeness.

Define a random walk

$$\hat{Y}(n) = \sum_{i=1}^{n} Z_m(X_i),\ n \geq 1 \tag{84}$$

and a stopping time

$$\hat{R} = \inf\{n \geq 1 : \hat{Y}(n) \geq c\}. \tag{85}$$

We define the overshoot of the random walk over threshold c by:

$$Q_c = \hat{Y}(\hat{R}) - c. \tag{86}$$

Define $\mu_m = \mathbb{E}_1[Z_m(X)]$ and $\sigma_m^2 = \mathrm{Var}_1[Z_m(X)]$. We define

$$\mu_m = \mathbb{E}_1[Z_m(X)] = \mathbb{D}_m(P_1\|P_\infty), \tag{87}$$

where we use the result of Lemma 4.3. We further define

$$\sigma_m^2 = \mathrm{Var}_1[Z_m(X)] = \mathbb{E}_1[Z_m(X)^2] - \mu_m^2. \tag{88}$$

Under the mild assumption that

$$\mathbb{E}_1[\mathcal{S}_m(X, P_\infty)^2],\ \mathbb{E}_1[\mathcal{S}_m(X, P_1)^2] < \infty, \tag{89}$$

we can use Theorem 1 of Lorden (1970) to conclude that:

$$\sup_{c\geq 0} Q_c \leq \frac{\mathbb{E}_1[(Z_m(X)^+)^2]}{\mathbb{E}_1[Z_m(X)]} \leq \frac{\mathbb{E}_1[Z_m(X)^2]}{\mathbb{E}_1[Z_m(X)]} = \frac{\mu_m^2 + \sigma_m^2}{\mu_m^2}, \tag{90}$$

where $(\cdot)^+ = \max(0, \cdot)$. We next use Wald's Lemma Woodroofe (1982) to show that for all $c > 0$:

$$\mathbb{E}_1[\hat{R}] = \frac{c}{\mu_m} + \frac{\mathbb{E}_1[Q_c]}{\mu_m} \leq \frac{c}{\mu_m} + \frac{\mu_m^2 + \sigma_m^2}{\mu_m^2}. \tag{91}$$

For any $n \geq 0$, $\hat{Y}(n) \leq Y_m(n)$ and so $\hat{R} \geq \tau_m^c$ for the $Y_m(\cdot), \tau_m^c$ of Definition 4.10. Then:

$$\mathbb{E}_1[\tau_m^c] \leq \mathbb{E}_1[\hat{R}] \leq \frac{c}{\mu_m} + \frac{\mu_m^2 + \sigma_m^2}{\mu_m^2}, \tag{92}$$

but as $c \to \infty$, the effect of the term $(\mu_m^2 + \sigma_m^2)/\mu_m^2$ approaches zero. Thus,

$$\mathcal{L}_{\texttt{WADD}}(\tau_m^c) \sim \frac{c}{\mathbb{D}_m(P_1 \| P_\infty)}. \tag{93}$$

where $f(c) \sim g(c)$ as $c \to \infty$ means that $\lim_{c\to\infty} f(c)/g(c) = 1$.

□

Proof of Theorem 5.2. Direct calculation gives:

$$Z_{\texttt{KL}}(X) = \log p_1(X) - \log p_\infty(X) = -\frac{1}{2}(X - \mu_1)^T V^{-1}(X - \mu_1) + \frac{1}{2}(X - \mu_\infty)^T V^{-1}(X - \mu_\infty).$$

Without loss of generality, we can substitute 0 in place of μ_∞ and substitute $\mu = \mu_1 - \mu_\infty$ in place of μ_1. With this substitution:

$$Z_{\texttt{KL}}(X) = X^T V^{-1}\mu - \frac{1}{2}\mu^T V^{-1}\mu. \tag{94}$$

In the same way, we calculate $Z_M(\cdot)$ for any fixed matrix M:

$$Z_M(X) = X^T V^{-1} M M^T V^{-1}\mu - \frac{1}{2}\mu^T V^{-1} M M^T V^{-1}\mu. \tag{95}$$

In the special case where $M = I$ (Fisher divergence-based setting), the formula for $Z_M(X)$ is the formula for $Z_{\texttt{KL}}(X)$ where all instances of V^{-1} are replaced by instances of V^{-2}. In our test, setting $M = V^{\frac{1}{2}}$ lets $MM^T = V^1$, which makes the diffusion-based test equivalent to its LLR-based analog. □

Proof of Theorem 5.3. Consider $P_\infty = \mathcal{N}(0, 1/4)$ and $P_1 = \mathcal{N}(0, 1)$. Here, $X \in \mathbb{R}^d$ for $d = 1$. There does not exist any matrix-valued function $m(X) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times w}$ such that $Z_m(X) = Z_{\texttt{KL}}(X)$ across all $X \in \mathbb{R}$. Define

$$P_\infty = \mathcal{N}(0, \sigma^2),\ P_1 = \mathcal{N}(0, 1). \tag{96}$$

These distributions are supported on $\mathbb{R}$. Suppose for contradiction that there exists some $m(X) : \mathbb{R} \mapsto \mathbb{R}$ such that:

$$Z_{\texttt{KL}}(X) = Z_m(X)\ \ \forall X \in \mathbb{R}. \tag{97}$$

where $Z_{\texttt{KL}}(X)$ is given in (8). Defining

$$u(X) = m^2(X), \tag{98}$$

we can see that $u(X) \geq 0$ for all $X \in \mathbb{R}$. We shall demonstrate a contradiction by showing that $u(X)$ cannot be positive for all $X \in \mathbb{R}$ when $\sigma = 1/2$.

To simplify notation, we denote

$$s_\infty(X) = \nabla \log p_\infty(X), \tag{99}$$
$$s_1(X) = \nabla \log p_1(X). \tag{100}$$

When $X, m(X)$ are scalars, then $s_\infty(X), s_1(X)$ are also scalars, and we can re-write the diffusion-based instantaneous detection score of Definition 4.8 in simpler terms. We denote the derivatives of $u(X), s_\infty(X)$, and $s_1(X)$ as $u'(X), s'_\infty(X)$, and $s'_1(X)$, respectively.

$$\begin{aligned} Z_{\sqrt{u}}(X) =& \frac{1}{2}u(X)s_\infty^2(X) + \frac{\partial}{\partial X}\Big(u(X)s_\infty(X)\Big) - \frac{1}{2}u(X)s_1^2(X) - \frac{\partial}{\partial X}\Big(u(X)s_1(X)\Big) \\ =& \frac{1}{2}u(X)\big(s_\infty^2(X) - s_1^2(X)\big) + u'(X)\big(s_\infty(X) - s_1(X)\big) + u(X)\big(s'_\infty(X) - s'_1(X)\big) \\ =& u(X)\bigg(\frac{1}{2}(s_\infty^2(X) - s_1^2(X)) + (s'_\infty(X) - s'_1(X)\bigg) + u'(X)\bigg(s_\infty(X) - s_1(X)\bigg). \end{aligned} \tag{101}$$

We next calculate $Z_{\mathtt{KL}}(X)$. We note that for a scalar Gaussian distribution $P = \mathcal{N}(\mu, \sigma)$:

$$\log p(X) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2, \tag{102}$$

and so

$$\log p_1(X) - \log p_\infty(X) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}X^2 + \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}X^2 = \log(\sigma) + \frac{X^2}{2}\bigg(\frac{1}{\sigma^2} - 1\bigg). \tag{103}$$

Setting $Z_{\mathtt{KL}}(X) = Z_m(X)$, we arrive at the first-order ordinary differential equation (ODE):

$$u(X)\bigg(\frac{1}{2}(s_\infty^2(X) - s_1^2(X)) + (s'_\infty(X) - s'_1(X))\bigg) + u'(X)\bigg(s_\infty(X) - s_1(X)\bigg) = \log(\sigma) + \frac{X^2}{2}\bigg(\frac{1}{\sigma^2} - 1\bigg). \tag{104}$$

We calculate:

$$s_\infty(X) = -\frac{1}{\sigma^2}X,\ s_1(X) = -X, \tag{105}$$

and further observe that $s_\infty(X) - s_1(X) = (1 - \sigma^{-2})X$, that $s'_\infty(X) - s'_1(X) = (1 - \sigma^{-2})$, and that $s_\infty^2(X) - s_1^2(X) = (\sigma^{-4} - 1)X^2$. Plugging into (104):

$$u(X)\bigg(\frac{1}{2}X^2\bigg(\frac{1}{\sigma^4} - 1\bigg) + \bigg(1 - \frac{1}{\sigma^2}\bigg)\bigg) + u'(X)\bigg(X\bigg(1 - \frac{1}{\sigma^2}\bigg)\bigg) = \log(\sigma) + \frac{X^2}{2}\bigg(\frac{1}{\sigma^2} - 1\bigg). \tag{106}$$

We next divide through by $(1 - \sigma^{-2})X$, noting that this term has a zero at $X = 0$. We shall take care not to integrate the resulting ODE through $X = 0$. Furthermore, this also introduces a zero if $\sigma = 1$, though in this case $p_\infty(X) = p_1(X)$ for all $X \in \mathbb{R}$, rendering detection impossible. Performing this division, we get that

$$u(X)k(X) + u'(X) = r(X), \tag{107}$$

where

$$k(X) = \frac{\frac{1}{2}X^2\left(\frac{1}{\sigma^4} - 1\right) + \left(1 - \frac{1}{\sigma^2}\right)}{X\left(1 - \frac{1}{\sigma^2}\right)} = -\gamma X + \frac{1}{X} \quad \text{where} \quad \gamma = \frac{1}{2}\bigg(\frac{1 - \frac{1}{\sigma^4}}{1 - \frac{1}{\sigma^2}}\bigg), \tag{108}$$

and where

$$r(X) = \frac{\log(\sigma) + \frac{X^2}{2}\left(\frac{1}{\sigma^2} - 1\right)}{X\left(1 - \frac{1}{\sigma^2}\right)} = \frac{\delta}{X} - \frac{X}{2} \quad \text{where} \quad \delta = \left(\frac{\log \sigma}{1 - \frac{1}{\sigma^2}}\right). \tag{109}$$

We note that γ, δ are independent of X and that they are positive for all $\sigma > 0$.

We shall attempt to solve this ODE via the integrating factor method. We calculate that $\int k(X)dX = \int -\gamma X + X^{-1}dX = -\frac{\gamma}{2}X^2 + \log|X|$ (note that it is not necessary to include a constant of integration C, as any such constant would cancel in the step of (111)). We define

$$D(X) = \exp\left(\int k(X)dX\right) = \exp\left(-\frac{\gamma}{2}X^2\right)|X|. \tag{110}$$

We recall the ODE of (106) and multiply through by $D(X)$:

$$D(X)u(X)k(X) + D(X)u'(X) = D(X)r(X). \tag{111}$$

We observe that $\frac{\partial}{\partial X}D(X) = D(X)k(X)$ and that $\frac{\partial}{\partial X}D(X)u(X) = D(X)k(X)u(X) + D(X)u'(X)$, which is the left-hand-side of (111). We make this substitution to the left-hand-side and apply the Fundamental Theorems of Calculus:

$$D(X)u(X) - D(a)u(a) = \int_a^X D(y)r(y)dy. \tag{112}$$

We know by Lemma 4.5 that the diffusion divergence is bounded. Recalling (105), we can express $\mathbb{D}_m(P_1\|P_\infty)$ as:

$$\frac{1}{2}\mathbb{E}_1\left[\left\|m^T(X)(\nabla \log p_1(X) - \nabla \log p_\infty(X))\right\|^2\right] = \frac{1}{2\sqrt{2\pi}}\left(1 - \frac{1}{\sigma^2}\right)^2 \int_{-\infty}^{\infty} e^{-\frac{1}{2}X^2}u(X)X^2dX < \infty. \tag{113}$$

Consider $\sigma = \frac{1}{2}$. For this choice of σ, we have that $\gamma = 2.5 > 1$. Recalling that $u(X) \geq 0$ for all X and noting that $e^{h(X)}, X^2, |X| \geq 0$ for all $X, h(\cdot)$, we have that:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}X^2}u(X)X^2dX < \infty \implies \int_{-\infty}^{\infty} e^{-\frac{\gamma}{2}X^2}u(X)X^2dX < \infty \tag{114}$$

as $\frac{\gamma}{2} > \frac{1}{2}$. Furthermore, for all $X \notin [-1, 1]$, we know that $X^2 > |X|$. Thus, the convergence of (114) implies:

$$\implies \int_{-\infty}^{\infty} e^{-\frac{\gamma}{2}X^2}u(X)|X|dX < \infty. \tag{115}$$

We note that the integral of (115) is equivalent to $e^{-C}\int_{-\infty}^{\infty} D(X)u(X)dX$. As all terms of the integrand are non-negative for all X, the convergence is absolute and implies that

$$\lim_{x\to\infty} D(X)u(X) = \lim_{X\to-\infty} D(X)u(X) = 0. \tag{116}$$

Plugging (116) into (112), we get that for all $X < 0$:

$$D(X)u(X) = \int_{-\infty}^{X} D(y)r(y)dy. \tag{117}$$

We now expand the integrand:

$$\int_{-\infty}^{X} D(y)r(y)dy = \underbrace{\int_{-\infty}^{X} \exp\left(-\frac{\gamma}{2}y^2\right)\frac{|y|}{y}\delta\, dy}_{\text{term 1}} - \underbrace{\int_{-\infty}^{X} \exp\left(-\frac{\gamma}{2}y^2\right)|y|\frac{y}{2}\, dy}_{\text{term 2}}.$$

If $X < 0$, then $y < 0$ and $\frac{|y|}{y} = -1$. Thus:

$$\text{term } 1 = -\delta\sqrt{\frac{2\pi}{\gamma}}\Phi(X\sqrt{\gamma}), \tag{118}$$

where Φ denotes the CDF of a standard scalar Gaussian with mean zero and standard deviation one. We next integrate term 2 using integration by parts, keeping in mind that $X < 0 \implies y < 0$, and hence $y|y| = -y^2$:

$$\begin{aligned}
\text{term } 2 &= \int_{-\infty}^{X} e^{-\frac{\gamma}{2}y^2}\frac{(-y^2)}{2}dy \\
&= \frac{1}{2\gamma}\int_{-\infty}^{X}(-\gamma y e^{-\frac{\gamma}{2}y^2})(y)dy \\
&= \frac{1}{2\gamma}\left([\gamma y e^{-\frac{\gamma}{2}y^2}]_{-\infty}^{X} - \int_{-\infty}^{X} e^{-\frac{\gamma}{2}y^2}dy\right) \\
&= \frac{1}{2\gamma}\left(Xe^{-\frac{\gamma}{2}X^2} - \sqrt{\frac{2\pi}{\gamma}}\Phi(X\sqrt{\gamma})\right).
\end{aligned} \tag{119}$$

We now know that

$$u(X) = \frac{1}{D(X)}\underbrace{\left(\zeta\Phi(X\sqrt{\gamma}) - \frac{1}{2\gamma}Xe^{-\frac{\gamma}{2}X^2}\right)}_{\tilde{u}(X)}, \tag{120}$$

where $\zeta = (-\delta - (2\gamma)^{-1})\sqrt{2\pi\gamma^{-1}}$. We observe that $D(X) \geq 0$ for all X, and the sign of $u(X)$ is equal to the sign of $\tilde{u}(X)$. Evaluating, we observe that $\tilde{u}(-1) \approx 0.054 > 0$ but $\tilde{u}(-0.05) \approx -0.013 < 0$. As such, the $u(X) < 0$ for some $X \in \mathbb{R}$, a contradiction. No matrix-valued function $m(X) : \mathbb{R} \mapsto \mathbb{R}$ can enforce that $Z_{\texttt{KL}}(X) = Z_m(X)$ for all $X \in \mathbb{R}$.

Consider a more general case, where $m(X) : \mathbb{R} \mapsto \mathbb{R}^{1\times w}$. Letting $s_P(X) = \nabla \log p(X)$, the diffusion-Hyvärinen score becomes:

$$\mathcal{S}_m(X, P) = \frac{1}{2}\|m^T(X)s_P(X)\|^2 + \nabla \cdot m(X)m^T(X)s_P(X). \tag{121}$$

Expanding the first term:

$$\begin{aligned}
\frac{1}{2}\|m^T(X)s_P(X)\|^2 =& \frac{1}{2}\|(m_1(X)s_P(X), \cdots m_w(X)s_P(X))^T\|^2 \\
=& \frac{1}{2}(m_1^2(X)s_P^2(X) + \cdots + m_w^2(X)s_P^2(X)) \\
=& \frac{1}{2}\left(\sum_{i=1}^{w} m_i^2(X)\right)s_P^2(X),
\end{aligned} \tag{122}$$

as $m_i(X)$ is scalar for any $1 \leq i \leq w$. Expanding the second term:

$$\nabla\cdot m(X)m^T(X)s_P(X) = \nabla \cdot \|m^T(X)\|^2 s_P(X) = \nabla \cdot \left(\sum_{i=1}^{w} m_i^2(X)\right)s_P(X). \tag{123}$$

In this case, scaling by $m(X) : \mathbb{R} \mapsto \mathbb{R}^{1\times w}$ is equivalent to scaling by $\tilde{m}(X) : \mathbb{R} \mapsto \mathbb{R}$ where

$$\tilde{m}(X) = \sqrt{\sum_{i=1}^{w} m_i^2(X)}. \tag{124}$$

We know that there cannot exist an $\tilde{m}(X)$ such that the diffusion test statistic matches the log-likelihood ratio. Thus, for this choice of P_∞, P_1, for all matrix-valued functions $m(X) : \mathbb{R} \mapsto \mathbb{R}^{1\times w}$ where $w \in \mathbb{N}$, $Z_m(\cdot)$ is not equal to $Z_{\mathtt{KL}}(\cdot)$.

Define $h(X) = |Z_{\mathtt{KL}}(X) - Z_m(X)|$. We know that for all $m(X)$ there exists at least one value $X^* \in \mathbb{R}$ for which $h(X^*) \neq 0$. By Assumptions 3.1 4.4, $Z_m(X)$, $Z_{\mathtt{KL}}(X)$, and $h(X)$ are all continuous in X. Letting $\epsilon = h(X^*)/2$, we have from the definition of continuity that there exists a $\delta > 0$ such that $h(X) > \epsilon$ for all $X \in (X^* - \delta, X^* + \delta)$. We know from Assumption 3.1 that P_∞, P_1 are supported on $\mathbb{R}^d$; hence $\mathbb{P}_\infty[X \in (X^* - \delta, X^* + \delta)] > 0$ and $\mathbb{P}_1[X \in (X^* - \delta, X^* + \delta)] > 0$.

□