Sequential Hypothesis Testing Based on Machine Learning

Ryan Harvey† Paolo Braca° Leonardo M. Millefiori° Peter Willett†

***Abstract*—With the rapid proliferation of Machine-Learning (ML) and Deep Learning (DL) based decision systems, properly characterizing their often unpredictable performance is a key challenge. In this work we introduce the notion of a Sequential Data-Driven Decision Function (S-D3F), as a data-driven analogue to the Sequential Probability Ratio Test (SPRT). Key performance metrics for sequential analysis are shown suitable for use in analyzing the S-D3F's performance both in terms of error probabilities and average stopping times. The notion of rate function from large deviations theory is extended to this S-D3F test, and it is shown that with a sequential approach the S-D3F can outperform its Fixed Sample-Size (FSS) counterpart in the D3F as the average number of samples needed to make a decision diverges.**

***Index Terms*—Machine Learning, Artificial Intelligence, Sequential Inference, Sequential Analysis, Signal Detection, Large Deviations Theory.**

I. Introduction

The rapid developments in Machine Learning (ML) and Artificial Intelligence (AI) over the recent decades have led to significant paradigm changes in many fields, and aspire to revolutionize disciplines spanning industrial, academic and cultural life [12]. For example, Multi-Layer Perceptrons (MLPs) are designated as *universal function approximators*, offering the promise to approximate arbitrarily complex functions provided they are suitably smooth [19]. Deep Learning (DL) has further revolutionized artificial neural networks (ANNs) by offering a variety of architectures spanning from typical feedforward structures to more complex networks with intricate feedback loops and self-organizing structures [20].

Recently, efforts have been made to characterize statistically the performance of ML techniques when applied to hypothesis testing problems (classification) and estimation problems (regression), e.g., see [4]. In the framework of random matrix theory, the statistical performance of ML techniques is studied in an asymptotic regime where the ratio between the number of observations (features) n and the number of training samples m is fixed. While both n and m grow asymptotically, n/m is constant under these conditions, see e.g., [14], [17]. In such a regime, it is possible to analytically study effects like the "double-descent" in overparameterized cases.

As opposed to random matrix theory, in [10], techniques from large deviations theory were applied to determine the conditions necessary for a ML classifier to yield exponentially vanishing error probabilities against the number of informative observations n. In particular, the notion of the Data-Driven Decision Function (D3F) is introduced, and its corresponding error rate functions are derived such that as n grows asymptotically, they are able to well approximate the performance of the fixed-sample size test given by the D3F [10]. These expressions were then derived under a variety of hypothesis testing conditions (both simple and composite hypotheses), including those of i.i.d. observations, as well as in the more general case where conditional dependencies may be present.

To summarize, [10] derives a process for computing the expressions:

$$\alpha_n = \mathbb{P}\left[\hat{\mathcal{H}}_1|\mathcal{H}_0\right] \approx \zeta_{n,0}e^{-nI_0(\gamma)}, \tag{1}$$

$$\beta_n = \mathbb{P}\left[\hat{\mathcal{H}}_0|\mathcal{H}_1\right] \approx \zeta_{n,1}e^{-nI_1(\gamma)}, \tag{2}$$

where $\hat{\mathcal{H}}_i$ indicates that the hypothesis $\mathcal{H}_i$ was accepted.

The above give the analytical expressions for the probabilities of False Alarm (α_n), and Missed Detection (β_n) in terms of the number of informative observations, n, a sub-exponential scaling term, $\zeta_{n,0}$, and a threshold-dependent rate function $I_k(\gamma)$ for a binary hypothesis test deciding between $\mathcal{H}_0$ and $\mathcal{H}_1$ using a D3F.

These rate functions $I_k(\gamma)$ are given by the Fenchel-Legendre Transform [8]:

$$I_k(\gamma) = \sup_{t\in\mathbb{R}}(\gamma t - \varphi_k(t)),\ k = 0, 1, \tag{3}$$

where $\varphi_k(t)$ is the limit of the (hypothesis-conditioned) cumulant-generating function:

$$\varphi_k = \lim_{n\to\infty}\frac{1}{n}\varphi_{n,k}(t) = \ln\left(\mathbb{E}\left[e^{tT^{(n)}}\middle|\mathcal{H}_k\right]\right),\ k = 0, 1, \tag{4}$$

where $\mathbb{E}[X|\mathcal{H}_k]$ is the expected value of X under $\mathcal{H}_k$ and $T^{(n)}$ is the D3F decision statistic of the n informative observations.

In [9], it was experimentally verified that the analytical process in [10] is sound for D3F classifier performance predictions, however this FSS approach to data-driven classification comes with drawbacks which we examine.

A. Challenges with Fixed-Sample Size Detection

This D3F approach, as well as others (e.g., [1], [13], [26]), seek to leverage the flexibility of ML/DL approaches for detection problems (e.g., signal detection) when a classical hypothesis testing approach like the Likelihood Ratio-Test (LRT) may not be applicable/available. This can happen when there is no clearly defined log-likelihood function, or when

† Dept. Of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06029, USA

° Research Department, Centre for Maritime Research and Experimentation, 19126, La Spezia, SP, Italy

⋆ Corresponding Author (email: ryan.harvey@uconn.edu)

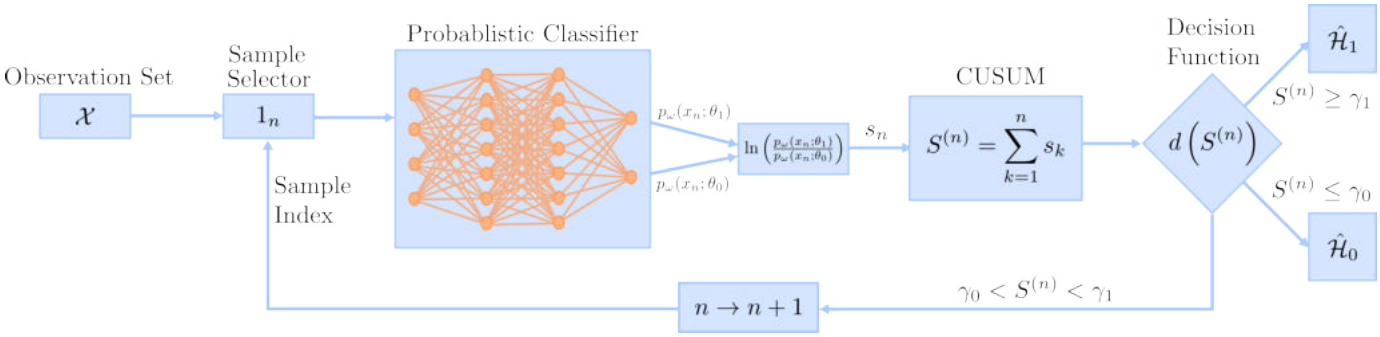


Fig. 1: Block diagram of S-D3F approach

explicit calculation of the Log-Likelihood Ratio (LLR) is too expensive to be practical.

However, as demonstrated with the D3F in [10], as well as for the Neyman-Pearson LRT in [11], there is a fundamental trade-off that must be made in these Fixed-Sample Size (FSS) tests.

In particular, given that $I_0(\gamma)$ and $I_1(\gamma)$ share a threshold, the FSS test implies a fundamental trade-off between the rates (and thus the exponential decay) of the two error probabilities α_n and β_n [6]. This trade-off is illustrated in Fig. 2 and serves as a key limitation for the FSS D3F. Specifically, it is possible to get the maximum rate of error decay under a given hypothesis at the cost of zero rate under the other hypothesis. In other words when the value of the threshold γ is fixed, we identify the pair $I_0(\gamma)$ and $I_1(\gamma)$. It is possible to show that the best achievable rate for i.i.d. observations are given by the Kullback-Liebler divergence when the decision statistic is the LLR.

In an effort to avoid this trade-off, the present work aims to extend the concept of the D3F to the sequential domain, using what we call the Sequential Data-Driven Decision Function (S-D3F) (pictured in Fig. 1). It is clear from Fig. 2 that the S-D3F is able to achieve the maximum error exponents for both hypotheses simultaneously, (as opposed to the D3F which can only achieve one at a time) indicating that it does provide a valuable alternative which does not make this compromise.

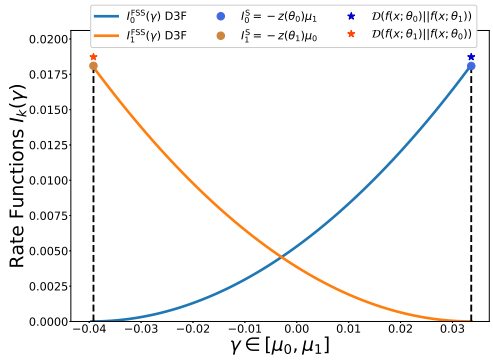


Fig. 2: Plot of FSS D3F Rate Function trade-off with shared threshold $I_k^{\text{FSS}}(\gamma)$, $k = 0, 1$ (as in [10]), the KL Divergence $\mathcal{D}(f(x;\theta_k)||f(x;\theta_{k'}))$, $k \neq k'$, and the achieved sequential rate functions I_k^{S} for the S-D3F.

The key aspect that will be explained later is that the number of observations is not fixed but is a random variable that depends on the evolution of the observations. To analyze the performance of the S-D3F, we will use concepts developed for analysis of the Sequential Probability Ratio Test (SPRT) [25], and demonstrate their extension to the S-D3F both analytically and via numerical simulation. As in [10], where the D3F was compared against the known optimal testing procedure (that of the LLR), this work will compare the performance of the S-D3F against the SPRT in a case where the likelihood ratio is available.

B. Paper Contributions

This work makes two contributions. Firstly, we extend the D3F procedure into the domain of sequential analysis and develop approximate expressions for the error probabilities, Average Stopping Number (ASN), and approximate error rates based on the equivalent metrics used for the SPRT. In this procedure, empirical approximations of the D3F's performance on the characterization set $\mathcal{C}$ will inform the parameters of these performance metrics. We make comparisons against a simple Laplace hypothesis test with the SPRT to produce numerical results for the metrics of concern. Secondly, we compare the notion of rate function used in both the FSS case and the sequential case for the D3F, demonstrating the S-D3F is able to achieve simultaneous rates that are near optimal.

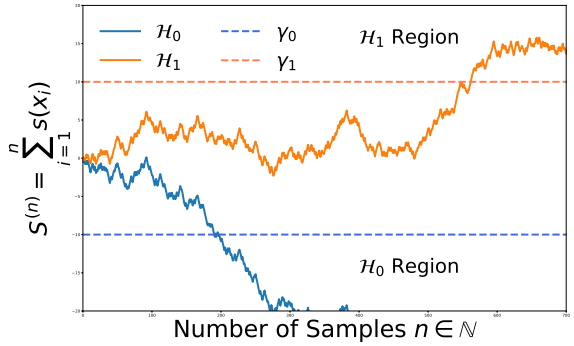
II. Problem Formulation

Let $\mathcal{X}$ be a sequence of i.i.d. (real-valued) observations $x_1, x_2, \ldots$ which originate from one of two possible hypotheses, denoted $\mathcal{H}_0$ and $\mathcal{H}_1$. We assume that the distribution of the observation set $\mathcal{X}$ remains consistent throughout all of our observations and that $\mathcal{X}$ is sufficiently large.

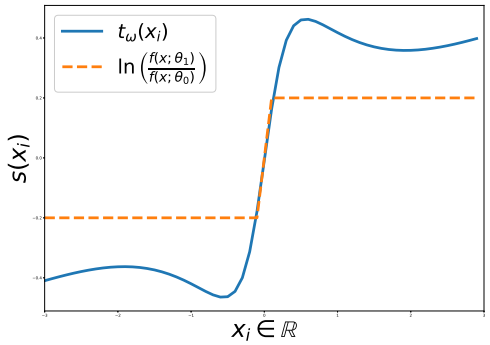
Thus, the realized observation set $\mathcal{X}$ then consists of a set:

$$\mathcal{X} = \{x_n : n \in \mathbb{N}, x_n \sim f(x;\theta_k)\}, \tag{5}$$

where $f(x;\theta_k)$ is a parameterized likelihood function corresponding to $\mathcal{H}_k$ for $k = 0, 1$. We will in general require that $f(x;\theta_k)$ be a well defined continuous density with finite moments. We define θ_k to be the parameter which distinguishes



(a) Example S-D3F trials for Laplace Distributed $\mathcal{H}_0$,$\mathcal{H}_1$.



(b) Per sample test statistics $s(x_i)$ for the $\mathcal{H}_0, \mathcal{H}_1$. The per-sample D3F outlined in Sec. IV-A is given by $t_\omega(x_i) = \ln\left(\frac{p_\omega(\mathcal{H}_1|x_i)}{p_\omega(\mathcal{H}_0|x_i)}\right)$, and $\ln\left(\frac{f(x;\theta_1)}{f(x;\theta_0}\right)$ is the LLR increment used in the SPRT. The former is derived automatically from the training data.

Fig. 3: Example S-D3F trial run under each hypothesis (left). Test statistic used in empirical trials in Sec. IV-A (right).

the two $\mathcal{H}_k$ hypotheses. This set $\mathcal{X}$ is either finite or countable, but must be of sufficient length for a decision to be made.

Define $S^{(n)}$ to be a real-valued decision statistic of the first n observations of $\mathcal{X}$ such that $S^{(n)} : \mathbb{R}^n \mapsto \mathbb{R}$. The fundamental purpose of $S^{(n)}$ is to use these n observations of $\mathcal{X}$ to decide between three distinct possibilities [25]:

1) $\hat{\mathcal{H}}_1$: Hypothesis $\mathcal{H}_1$ is accepted;
2) $\hat{\mathcal{H}}_0$: Hypothesis $\mathcal{H}_0$ is accepted;
3) $\emptyset$: More observations are needed.

This $S^{(n)}$ is known as a *sequential decision statistic* and is distinct from FSS decision statistics in one key aspect: the sample size required to come to a decision is a *random variable* rather than an *a priori* fixed quantity [2], [25]. As a result, one must specify the performance metrics of the sequential test in terms of the *distribution* of n, as opposed to the FSS regime.

While not all sequential decision statistics are of this form [18], we will assume that $S^{(n)}$ takes on the role of a cumulative sum:

$$S^{(n)} = \sum_{k=1}^{n} s_k \tag{6}$$

where s_k is the single sample test-statistic corresponding to observation $x_k \in \mathcal{X}$.

In particular, we investigate the asymptotic behavior of $S^{(n)}$ as the average number of samples required to make a binary decision diverges. This is performed by first establishing a set of real-valued thresholds $\{\gamma_0, \gamma_1\}$ such that $\gamma_0 < 0 < \gamma_1$ and defining our decision function $d(S^{(n)})$ such that:

$$d(S^{(n)}) = \begin{cases} \hat{\mathcal{H}}_1, & S^{(n)} \geq \gamma_1 \\ \hat{\mathcal{H}}_0, & S^{(n)} \leq \gamma_0 \\ n \to n+1, & \gamma_0 < S^{(n)} < \gamma_1. \end{cases} \tag{7}$$

Fig. 3a depicts two example cases of this process for a particular $\{\gamma_0, \gamma_1\}$ set with their respective $S^{(n)}$.

III. Sequential Hypothesis Testing

In the case where the individual $S^{(n)}$ test statistic is given by the LLR, this sequential test procedure is known as the SPRT [2]. The SPRT has many qualities which are attractive, most notably being that with i.i.d. increments

$$s_n = \ln\left(\frac{f(x_n;\theta_1)}{f(x_n;\theta_0)}\right), \tag{8}$$

where the samples x_n are either drawn from $\mathcal{H}_0 \sim f(x;\theta_0)$ or $\mathcal{H}_1 \sim f(x;\theta_1)$, the SPRT produces the sample optimal decision procedure for a specified set of error probabilities (α, β) [2]. Appealing to the Wald-Wolfowitz Theorem, all tests of simple hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ require an average number of observations greater than or equal to the SPRT to achieve the same error probabilities [24].

Given that the distributions $\mathcal{H}_0$ and $\mathcal{H}_1$ are often either entirely unknown or too complicated to directly model, we are focusing on the case of $S^{(n)}$ being a data-driven function, as in [10]. This decision statistic, denoted $S_\omega^{(n)}$ is given by:

$$S_\omega^{(n)} = \sum_{i=1}^{n} t_\omega(x_i), \tag{9}$$

where $t_\omega(x_i)$ is given by the per-sample D3F transformation of x_i (depicted in Fig. 3b for the test case outlined in Sec. IV-A). We assume that there is a sufficiently large training set $\mathcal{Y}$ of finite size m_y, where labeled data is available under each hypothesis and is independent of $\mathcal{X}$. Via training against this labeled data, we can produce a decision statistic parameterized on our trained classifier, denoted $S_\omega^{(n)}$, which serves as a sequential analog to the FSS D3F analyzed in [10]. We also assume that there is a sufficiently large *characterization set* $\mathcal{C}$ of finite size m_c, where labeled data is available corresponding to both $\mathcal{H}_k$, and is independent of our observation set $\mathcal{X}$. Note that $\mathcal{C}$ may be a subset of the training set $\mathcal{Y}$ [10].

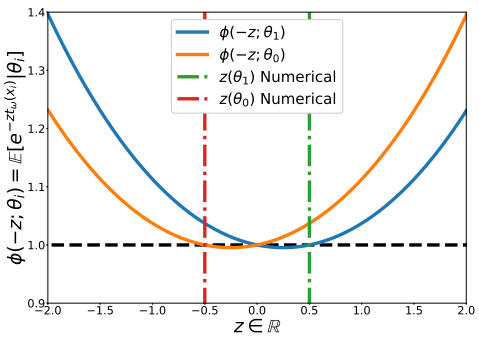


Fig. 4: Moment Generating Function of Sequential D3F in Experiment Case, $\theta_0 = -0.1, \theta_1 = 0.1$, $\mathcal{H}_k = \mathcal{L}(x;\theta_k,1)$.

The characterization set $\mathcal{C}$ will be used to derive important statistical information about the increment $t_\omega(x)$ used to produce $S_\omega^{(n)}$, including the hypothesis conditioned moment generating functions $\phi_k(z)$ (discussed in Sec. III-A) and average increments μ_k under $k = 0, 1$. This $\phi_k(z)$ will be approximated via the sample mean

$$\phi_{k,m_c}(z) = \frac{1}{m_c}\sum_{i=1}^{m_z} e^{zc_i} \tag{10}$$

which converges to $\phi_k(z)$ as $m_c \to \infty$. Thus, given a finite size $\mathcal{C}$, we approximate $\phi_k(z)$ via the sample-mean of the per-sample D3F under each hypothesis.

The performance of this S-D3F is then evaluated via two primary metrics of interest:

1) The Average Stopping Number (ASN):

$$\overline{n}_1 = \mathbb{E}\left[N\,|\mathcal{H}_1\right], \quad \overline{n}_0 = \mathbb{E}\left[N\,|\mathcal{H}_0\right] \tag{11}$$

2) The Error Probability curves:

$$\alpha_{\overline{n}_1} = \mathbb{P}[S^{(\overline{n}_1)} \geq \gamma_1 | \mathcal{H}_0], \; \beta_{\overline{n}_0} = \mathbb{P}[S^{(\overline{n}_0)} \leq \gamma_0 | \mathcal{H}_1]. \tag{12}$$

Note that in (12) we mean $\alpha_{\overline{n}_1}$ to be the probability of false-alarm given that it takes $\overline{n}_1$ samples on average to make a decision under $\mathcal{H}_1$. Similarly, $\beta_{\overline{n}_0}$ is the probability of missed detection given that it takes on average $\overline{n}_0$ samples to make a decision under $\mathcal{H}_0$.

As we will detail thoroughly in Secs. III-B and III-C, these two metrics are well approximated by functions of the threshold set (γ_0, γ_1) when certain statistical properties of the increments can be observed empirically[1].

Finally, as a comparison to the large deviations theory based FSS D3F explored in [10], we examine the *rate functions* $I_k(\gamma_1, \gamma_0)$ for our sequential test, defined as:

$$I_0(\gamma_0,\gamma_1) = -\frac{\ln(\alpha_{\overline{n}_1})}{\overline{n}_1}, \quad I_1(\gamma_0,\gamma_1) = -\frac{\ln(\beta_{\overline{n}_0})}{\overline{n}_0}, \tag{13}$$

[1]Note that even in the simple hypothesis SPRT, (i.e., the ideal case for classical sequential analysis), these metrics can only (in general) be approximated when the thresholds have finite magnitude [2].

where the k index indicates the true hypothesis $\mathcal{H}_k$.

These expressions for the S-D3F are derived from the solutions to $\phi_k(z)$ approximations, thus allowing the characterization set $\mathcal{C}$ to inform the S-D3F's expected response to the observation set $\mathcal{X}$, given that it is distributed according to one of the two hypotheses whose samples are present in $\mathcal{C}$.

We demonstrate that, using the S-D3F and under certain regularity conditions, it is possible to achieve the maximum possible error rate for *both* error probabilities as $\overline{n}_k$ diverges for $k = 0, 1$. This result is verified empirically using a large number of Monte Carlo runs and compared against the SPRT, which is guaranteed to achieve the maximum possible error exponent [7].

The S-D3F is then compared against the FSS case examined in [10], in which achieving a larger error exponent in one of the error probabilities comes at the cost of a low error exponent in the other.

Note that in the case of the comparison of the SPRT to the FSS LLR, the *asymptotic relative efficiency* (ARE) allows for the comparison of the FSS LLR and SPRT in terms of the expected number of samples required to achieve a particular set of error probabilities [3]. In the case of i.i.d. observations, [3] also demonstrates that Wald's approximations for the ASN and error exponents become exact under certain asymptotic circumstances.

A. Moment Generating Functions

Central to both metrics of concern is the moment generating function $\phi_k(z)$ defined by:

$$\phi_k(z) = \mathbb{E}[e^{\tau z}|\mathcal{H}_k], \tag{14}$$

where τ is a real valued random variable corresponding to the elementwise test-statistic used for sequential hypothesis testing. In the SPRT case, this corresponds to the single point log-likelihood ratio $\tau = s(x) = \ln\left(f(x;\theta_1)\right) - \ln\left(f(x;\theta_0)\right)$, while in the D3F case τ corresponds to $t_\omega(x)$.

We assume that $\mathbb{P}(\tau > 0) > 0$, $\mathbb{P}(\tau < 0) > 0$, $\phi_k(z)$ is real, continuous and differentiable for all z, and that $\frac{d}{dz}\phi_k(0) \neq 0$ for $k = 0, 1$. If these conditions are satisfied, then the equation,

$$\phi_k(-z) - 1 = 0, \tag{15}$$

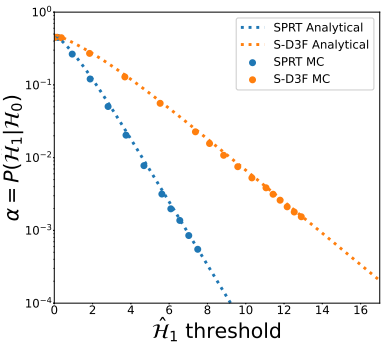
has exactly one nonzero real root $z(\theta_k)$, such that

$$\begin{aligned} z(\theta_k) > 0 \text{ if } \mathbb{E}[\tau|\mathcal{H}_k] > 0, \\ z(\theta_k) < 0 \text{ if } \mathbb{E}[\tau|\mathcal{H}_k] < 0. \end{aligned} \tag{16}$$

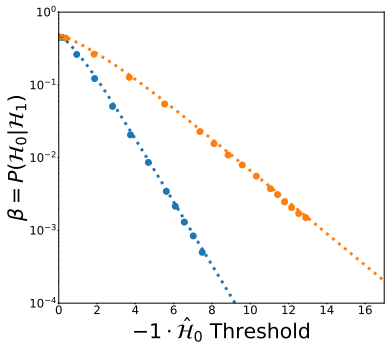
We do not need to prove this statement here, however we will utilize its results and implications. For proof of this statement see [2], [18], [25].

Fig. 4 depicts an example $\phi_k(-z)$ and numerical approximations to $z(\theta_k)$ for the S-D3F case. The $z(\theta_k)$ numerical values are key to error probability curve approximations given in Sec. III-B, which then factor into the ASN expressions (covered in Sec. III-C) and the sequential rate function approximations given in Sec. III-D.

Note that in the SPRT case, we are certain to have $z(\theta_1) = 1$ and $z(\theta_0) = -1$ if $\mathbb{E}[s|\mathcal{H}_1] > 0$ and $\mathbb{E}[s|\mathcal{H}_0] < 0$ [15].

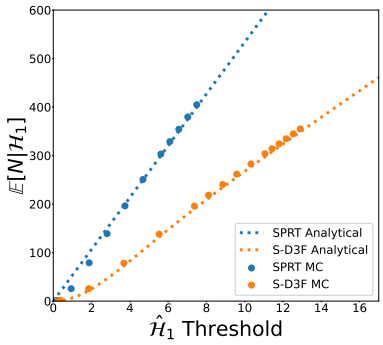


(a) Probability of false alarm $\alpha(\gamma_0, \gamma_1)$

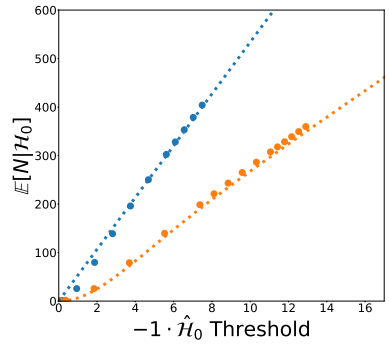


(b) Probability of missed detection $\beta(\gamma_0, \gamma_1)$

Fig. 5: Error probability curves for the S-D3F and SPRT compared against the empirical Monte Carlo error probabilities.



(a) ASN $\mathcal{H}_1$ Case: $\overline{n}_1(\gamma_0, \gamma_1)$



(b) ASN $\mathcal{H}_0$ Case: $\overline{n}_0(\gamma_0, \gamma_1)$

Fig. 6: ASN approximation curves against empirical Monte Carlo trials for SPRT and S-D3F.

B. Error Probability Curves

The first metrics of performance analysis for our S-D3F are the Error Probability curves $\alpha(\gamma_0, \gamma_1)$ and $\beta(\gamma_0, \gamma_1)$ given by:

$$\alpha(\gamma_0, \gamma_1) = \mathbb{P}\{\hat{\mathcal{H}}_1|\mathcal{H}_0\} = \mathcal{Q}(\theta_0), \tag{17}$$

$$\approx \frac{1 - e^{-\gamma_0 z(\theta_0)}}{e^{-\gamma_1 z(\theta_0)} - e^{-\gamma_0 z(\theta_0)}}, \tag{18}$$

$$\beta(\gamma_0, \gamma_1) = \mathbb{P}\{\hat{\mathcal{H}}_0|\mathcal{H}_1\} = 1 - \mathcal{Q}(\theta_1), \tag{19}$$

$$\approx \frac{1 - e^{-\gamma_1 z(\theta_1)}}{e^{-\gamma_0 z(\theta_1)} - e^{-\gamma_1 z(\theta_1)}}, \tag{20}$$

where $\mathcal{Q}(\theta)$ corresponds to the *Operating Characteristic* (OC) function:

$$\mathcal{Q}(\theta_k) = \mathbb{P}\left[S^{(n)} \leq \gamma_0 \middle| \mathcal{H}_k\right], \tag{21}$$

for $k = 0, 1$, as given in [2].

Under the assumption of "small excess"

$$\left|\mathbb{E}\left[S^{(N)} \middle| \mathcal{H}_i\right]\right| = |\gamma_i| + \epsilon, \quad \epsilon > 0, \quad \frac{\epsilon}{|\gamma_i|} \approx 0 \tag{22}$$

(i.e., at the time of making a decision our $S^{(n)}$ is approximately equal to one of the thresholds), we can use the approximation [2]:

$$\mathcal{Q}(\theta) \approx \frac{e^{-\gamma_0 z(\theta)} - 1}{e^{-\gamma_0 z(\theta)} - e^{-\gamma_1 z(\theta)}}, \tag{23}$$

which after substitution into (17) and (19), and some algebraic manipulation yields (18) and (20) respectively.

C. Average Stopping Number

Given that the conditions in Sec. III-A are satisfied and that

$$\mu_0 = \mathbb{E}[s(x)|\mathcal{H}_0] < 0 < \mu_1 = \mathbb{E}[s(x)|\mathcal{H}_1], \tag{24}$$

we can specify the ASN for our test (under the small excess assumption) to be given by

$$\overline{n}_k \approx \frac{\gamma_1 \mathcal{Q}(\theta_k) + \gamma_0 \left[1 - \mathcal{Q}(\theta_k)\right]}{\mu_k}, \tag{25}$$

meaning that (with some algebraic manipulation):

$$\overline{n}_0(\gamma_0, \gamma_1) \approx \frac{(\gamma_1 - \gamma_0) + \gamma_0 e^{-z(\theta_0)\gamma_1} - \gamma_1 e^{-z(\theta_0)\gamma_0}}{\mu_0 \left(e^{-z(\theta_0)\gamma_1} - e^{-z(\theta_0)\gamma_0}\right)}, \tag{26}$$

$$\overline{n}_1(\gamma_0, \gamma_1) \approx \frac{(\gamma_1 - \gamma_0) + \gamma_0 e^{-z(\theta_1)\gamma_1} - \gamma_1 e^{-z(\theta_1)\gamma_0}}{\mu_1 \left(e^{-z(\theta_1)\gamma_1} - e^{-z(\theta_1)\gamma_0}\right)}. \tag{27}$$

Note that in the case of the simple hypothesis SPRT, the mean increment μ_k is related exactly to Kullback-Liebler divergence between $f(x;\theta_1)$ and $f(x;\theta_0)$ given by

$$\begin{aligned}\mu_1 &= \mathbb{E}\left[\ln\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)\middle|\mathcal{H}_1\right] \triangleq \mathcal{D}\left(f(x;\theta_1)||f(x;\theta_0)\right),\\ \mu_0 &= \mathbb{E}\left[\ln\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)\middle|\mathcal{H}_0\right] \triangleq -\mathcal{D}\left(f(x;\theta_0)||f(x;\theta_1)\right).\end{aligned} \tag{28}$$

This will play a role in the following section on the sequential rate functions $I_k(\gamma_0,\gamma_1)$ we shall now develop.

D. Sequential Rate Functions

The notion of an error rate function for a binary hypothesis test has been a topic of study in both the FSS regime [6], [10], [11], as well as for the SPRT [7], [22].

We now seek to extend this notion from the SPRT to the S-D3F. In particular we specify

$$I_0(\underline{n}_1) = -\frac{\ln(\alpha_{\underline{n}_1})}{\underline{n}_1}, \quad I_1(\underline{n}_0) = -\frac{\ln(\beta_{\underline{n}_0})}{\underline{n}_0}, \tag{29}$$

to indicate the rate of the error probability decay when the truth is given by $k=0,1$ and the sequential decision process takes on average $\overline{n}_j, j=0,1, j\neq k$ to conclude.

Given that the expressions (26)-(27) do not have analytical solutions directly relating γ_k to $\overline{n}_k$, we instead specify analytical forms for I_k in terms of the pair (γ_0,γ_1).

Thus, our rate function expressions are given by

$$I_0(\gamma_0,\gamma_1) = -\frac{\ln(\alpha(\gamma_0,\gamma_1))}{\overline{n}_1(\gamma_0,\gamma_1)}, \tag{30}$$

$$I_1(\gamma_0,\gamma_1) = -\frac{\ln(\beta(\gamma_0,\gamma_1))}{\overline{n}_0(\gamma_0,\gamma_1)}, \tag{31}$$

where α and β are given by (18) and (20) respectively, and $\overline{n}_0$ and $\overline{n}_1$ are given by (26) and (27) respectively.

In the limiting case, as $|\gamma_k|\to\infty$ for $k=0,1$, these expressions converge to

$$\lim_{\gamma_1\to\infty} I_0 = -z(\theta_0)\mu_1, \quad \lim_{\gamma_0\to-\infty} I_1 \to -z(\theta_1)\mu_0, \tag{32}$$

which in the case of the D3F must be evaluated numerically via the approximated $\phi_k(z)$, and in the case of the SPRT is given by:

$$\lim_{\gamma_1\to\infty} I_0 = \mathcal{D}(f(x;\theta_1)||f(x;\theta_0)), \tag{33}$$

$$\lim_{\gamma_0\to-\infty} I_1 = \mathcal{D}(f(x;\theta_0)||f(x;\theta_1)). \tag{34}$$

IV. Experimental Results

A. Experimental Setup

For the experimental setup, a simple Laplace hypothesis test case is run where

$$\mathcal{H}_k \to f(x;\theta_k) \sim \mathcal{L}\left(x;\mu_{x,k},\sigma_x^2\right), \tag{35}$$

with $\mu_{x,0} = -0.1 = -\mu_{x,1}$ and $\sigma_x^2 = 1$. The network setup used is a single layer feedforward neural network with one 8-neuron hidden layer, and a 2-neuron output layer corresponding to the architecture used in [9]. The outputs of the neural network correspond to an approximated $p_\omega(\mathcal{H}_k|x_i)$. As in [9], [10], the final transformation $t_\omega(x_i)$ is given by:

$$t_\omega(x_i) = \ln\left(p_\omega(\mathcal{H}_1|x_i)\right) - \ln\left(p_\omega(\mathcal{H}_0|x_i)\right), \tag{36}$$

denoted as the per-sample D3F (depicted in blue in Fig. 3b).

The network used was trained on 20,000 labeled samples pulled from each hypothesis distribution with binary cross entropy as the loss function [5], and 100,000 sequences were generated under each hypothesis for testing. The thresholds (γ_0,γ_1) were symmetric such that $\gamma = |\gamma_0| = |\gamma_1|$ and were chosen such that

$$\mu' \leq \gamma \leq 350\mu',$$

where

$$\mu' = \min\{|\mu_0|,|\mu_1|\}.$$

In our test cases, $|\mu_0| \approx |\mu_1|$, so the scaling of γ against $|\mu_k|$ would have been roughly equivalent in either case, but this is not true in general.

B. Error Probability Curves

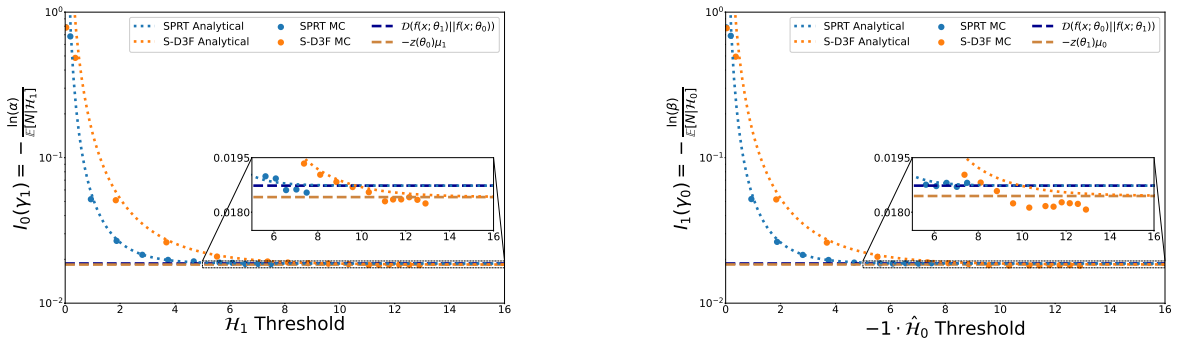
Fig. 5 shows the comparison between the error probability curves $\alpha(\gamma_0,\gamma_1)$ and $\beta(\gamma_0,\gamma_1)$ and the realized Monte Carlo results for both the SPRT and S-D3F. While it is obvious that the rate of exponential decay is higher for the SPRT, the S-D3F has well behaved exponential error probability decay and is well approximated by (18)-(20).

In this case, the moment generating functions for the S-D3F under each hypothesis are roughly symmetric, resulting in the error probabilities α and β decaying at roughly the same exponential rate. In the case of $\phi_k(z)$ being very asymmetric, there could be significant differences between α and β, with one error probability decaying much faster than the other.

The realized numbers of Missed Detections and False Alarms in the Monte Carlo trials are given in Table I (with some intermediary points omitted).

C. Average Stopping Number

Fig. 6 displays the comparison between the approximations (26)-(27) and the realized empirical stopping times. Analyzing the Monte Carlo results, it is clear that in the $\mathcal{H}_0$ case, the expression (26) is a slight underestimate of the realized average stopping time. This underestimation occurs in both the S-D3F and SPRT trials, and could be a result of either characterization error or the small excess approximation used for both analytical curves. This underestimation is minor, however, indicating that the ASN approximations (26)-(27) perform well against the empirical results.



(a) Rate Function $I_0 = -\frac{\ln(\alpha)}{\overline{n}_1}$

(b) Rate Function $I_1 = -\frac{\ln(\alpha)}{\overline{n}_0}$

Fig. 7: Comparison of analytical rate functions $I_k(\gamma_0, \gamma_1)$ against empirical Monte Carlo results.

TABLE I: Threshold selection for D3F trials (top) and SPRT trials (bottom) against realized errors where $\gamma = N_d\mu'$

N_d D3F	1	10	100	150	200	260	300	310	320	330	340	350
MD	45368	43855	12698	5446	2281	793	374	311	247	206	171	150
FA	44960	44496	12865	5581	2265	751	385	314	260	211	179	154
N_d SPRT	1	10	50	100	150	200	250	300	325	350	375	400
MD	45277	44846	26195	12220	5069	2058	860	344	215	130	84	50
FA	45161	44895	26498	12053	5045	2038	776	316	198	137	85	55

D. Sequential Rate Functions

When analyzing Figs. 7a and 7b, the approximations

$$I_0 \to -z(\theta_0)\mu_1, \quad I_1 \to -z(\theta_1)\mu_0,$$

slightly overestimate the realized rate in the Monte Carlo trials (on the order of roughly 10^{-3}). Given that the ASN curves realized in the Monte Carlo trials slightly underestimate $\overline{n}_0$ and $\overline{n}_1$ (in both the SPRT and S-D3F), it is reasonable to assume that will then produce an overestimated error exponent at a particular $\overline{n}_k$ value.

It is also possible that the relative sparsity of the errors at higher threshold values (making up only approximately 150 out of $100,000$ trials in each case, as seen in Table I) has led to an empirical estimate that overestimates the presence of errors in the true asymptotic distribution.

In either case, it is clear that the approximations for the rate functions given in (30)-(31) provide reasonable approximations of the rate of error decays when $\mathcal{H}_0$ and $\mathcal{H}_1$ are true, respectively.

E. Comparison of Rate Functions Between D3F and S-D3F

It can be seen in (30)-(31) that the rate functions achieved by the SPRT as $|\gamma_k| \to \infty$ are equivalent to the maximum possible rate functions achievable by one of the two $I_k(\gamma)$ in the FSS case. Thus, in exchange for requiring a diverging number of samples to make a decision $\mathcal{H}_k$, the sequential test offers the maximum possible error rates for both hypotheses simultaneously.

This is visually shown in Fig. 2, where to achieve the desired $I_1^{\text{FSS}}(\gamma) = I_1^{\text{S}}$ one would need to have the threshold very close to μ_0, such that $I_0^{\text{FSS}}(\gamma) \approx 0$. In this way, the S-D3F (as well as the SPRT) can achieve pairs of error rate functions which the FSS D3F cannot, at the cost of potentially having the test take longer to close in particular realizations.

V. Conclusions and Future Work

In many ways, this work with the S-D3F is fundamentally an extension of [9], [10] geared towards a different set of trade-offs. While both cases have error probability curves which decay against the number of observations (with n being a fixed number in the FSS case, and N being a random variable in the sequential case), both demonstrate the capacity for data-driven decision functions to have desirable properties similar to classical signal detection approaches.

The S-D3F allows for a near-optimal sequential test procedure with error probabilities that decay at roughly the same asymptotic rate as a classical SPRT (as shown in Fig. 7), while not requiring a clearly defined likelihood function (which in many cases isn't available). The approximations for the error probability curves and ASN, given in (18)-(20) and (26)-(27) respectively, allow for an insight into the False Alarm rate and Missed Detection rate that is often not available for data-driven methods. As a result, the S-D3F may be applied to a wide variety of signal detection tasks where a parametric likelihood function is not always clearly justifiable (e.g., extended object detection in a video stream [16]).

There are still many avenues for future work which we will pursue. Firstly, this analysis is only concerned with the simply hypothesis problem on a fairly simple hypothesis test where we had a clearly defined LLR. An extension of the S-D3F to a composite hypothesis test could allow for many of the same benefits under a much more flexible hypothesis testing paradigm.

In addition, extension of the S-D3F to cases where there are conditional dependencies between the observations would allow for the S-D3F approach to be applied to cases like video-streams where the observations may have Markovian dependencies. [23] provides extensions of the SPRT process to general non-i.i.d cases and this approach may be adaptable to an S-D3F process.

In future we also plan to pursue a D3F approach to the *quickest detection* problem. Data-driven methods could lend themselves well to problems of change-point detection where there may not be a reliable likelihood model. A D3F approach to quickest detection could allow for many of the benefits of classical changepoint detection applied to circumstances where the observations many be heterogenous and complex like contextual anomaly detection [21].

REFERENCES

[1] Z. Baird, M. K. McDonald, S. Rajan, and S. Lee, "A Neyman-Pearson criterion-based neural network detector for maritime radar," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, 2021, pp. 1–8.

[2] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[3] R. H. Berk, "Some asymptotic aspects of sequential analysis," *The Annals of Statistics*, vol. 1, no. 6, pp. 1126–1138, 11 1973.

[4] J. Berner, P. Grohs, G. Kutyniok, and P. Petersen, *The Modern Mathematics of Deep Learning*. Cambridge University Press, 2022, p. 1–111.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

[6] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 405–417, 07 1974.

[7] P. Boroumand and A. Guillén i Fàbregas, "Mismatched binary hypothesis testing: Error exponent sensitivity," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6738–6761, 2022.

[8] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge University Press, 2004.

[9] P. Braca, L. M. Millefiori, A. Aubry, S. Marano, A. De Maio, and P. Willett, "Experimental corroboration of trained classification performance predictions," in *2023 IEEE 9th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2023, pp. 1–5.

[10] ——, "Statistical hypothesis testing based on machine learning: Large deviations analysis," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 464–495, 2022.

[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, 1998.

[12] Y. K. D. et. al, "AI: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, p. 101994, 2021.

[13] G. Ford, B. J. Foster, S. A. Braun, and M. Kam, "Unknown signal detection in switching linear dynamical system noise," *IEEE Transactions on Signal Processing*, vol. 71, pp. 2220–2234, 2023.

[14] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *Annals of statistics*, vol. 50, no. 2, p. 949, 2022.

[15] C. W. Helstrom, *Elements of Signal Detection & Estimation*. Prentice Hall, 1995.

[16] K. Ismail and T. Breckon, "On the performance of extended real-time object detection and attribute estimation within urban scene understanding," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 641–646.

[17] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Communications on Pure and Applied Mathematics*, vol. 75, no. 4, pp. 667–766, 2022.

[18] N. Mukhopadhyay and B. M. de Silva, *Sequential Methods and Their Applications*, 1st ed. Chapman and Hall, 2009.

[19] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[20] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53 040–53 065, 04 2019.

[21] G. Soldi, D. Gaglione, S. Raponi, N. Forti, E. d'Afflisio, P. Kowalski, L. M. Millefiori, D. Zissis, P. Braca, P. Willett, A. Maguer, S. Carniel, G. Sembenini, and C. Warner, "Monitoring of critical undersea infrastructures: The Nord Stream and other recent case studies," *IEEE Aerospace and Electronic Systems Magazine*, vol. 38, no. 10, pp. 4–24, 2023.

[22] A. Tartakovksy, I. Nikiforov, and M. Basseville, *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, 1st ed. Chapman and Hall, 2014.

[23] A. Tartakovsky, *Sequential Change Detection and Hypothesis Testing*, 1st ed. Chapman and Hall, 2020.

[24] A. Wald and J. Wolfowitz, "Optimum Character of the Sequential Probability Ratio Test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326 – 339, 1948.

[25] A. Wald, *Sequential Analysis*. John Wiley & Sons, Inc, 1947.

[26] L. Xie, S. Zou, Y. Xie, and V. V. Veeravalli, "Sequential (quickest) change detection: Classical results and new directions," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 494–514, 2021.