IEEE Open Journal of Signal Processing

Received 21 July 2022; revised 28 October 2022; accepted 7 December 2022. Date of publication 6 January 2023; date of current version 1 February 2023. The review of this article was arranged by Associate Editor Fan Liu.

Digital Object Identifier 10.1109/OJSP.2022.3232284

Statistical Hypothesis Testing Based on Machine Learning: Large Deviations Analysis

PAOLO BRACA [1] (Senior Member, IEEE), LEONARDO M. MILLEFIORI [1] (Member, IEEE), AUGUSTO AUBRY [2] (Senior Member, IEEE), STEFANO MARANO [3] (Senior Member, IEEE), ANTONIO DE MAIO [2] (Fellow, IEEE), AND PETER WILLETT [4] (Fellow, IEEE)

[1]Research Department, Centre for Maritime Research and Experimentation, 19126 La Spezia, SP, Italy
[2]DIETI, University of Naples "Federico II", 80125 Naples, NA, Italy
[3]DIEM, University of Salerno, 84084 Fisciano, SA, Italy
[4]Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA

CORRESPONDING AUTHOR: PAOLO BRACA (e-mail: paolo.braca@cmre.nato.int)

The work of Paolo Braca and Leonardo M. Millefiori was supported by NATO Allied Command Transformation (ACT) via the project Data Knowledge Operational Effectiveness (DKOE). The work of Peter Willett was supported by NIUVT.

ABSTRACT We study the performance of Machine Learning (ML) classification techniques. Leveraging the theory of large deviations, we provide the mathematical conditions for a ML classifier to exhibit error probabilities that vanish exponentially, say $\exp(-nI)$, where n is the number of informative observations available for testing (or another relevant parameter, such as the size of the target in an image) and I is the error rate. Such conditions depend on the Fenchel-Legendre transform of the cumulant-generating function of the Data-Driven Decision Function (D3F, i.e., what is thresholded before the final binary decision is made) learned in the training phase. As such, the D3F and the related error rate I depend on the given training set. The conditions for the exponential convergence can be verified and tested numerically exploiting the available dataset or a synthetic dataset generated according to the underlying statistical model. Coherently with the large deviations theory, we can also establish the convergence of the normalized D3F statistic to a Gaussian distribution. Furthermore, approximate error probability curves $\zeta_n \exp(-nI)$ are provided, thanks to the refined asymptotic derivation, where ζ_n represents the most representative sub-exponential terms of the error probabilities. Leveraging the refined asymptotic, we are able to compute an accurate analytical approximation of the classification performance for both the regimes of small and large values of n. Theoretical findings are corroborated by extensive numerical simulations and by the use of real-world data, acquired by an X-band maritime radar system for surveillance.

INDEX TERMS Machine learning, deep learning, large deviations principle, exact asymptotics, statistical hypothesis testing, Fenchel-Legendre transform, extended target detection, radar/sonar detection, X-band maritime radar.

I. INTRODUCTION

A. RECENT ADVANCES IN AI

As we enter the Artificial Intelligence (AI) technological age, the ambition is to offer the augmentation and the potential replacement of tedious human tasks and activities within a wide range of industrial, intellectual and social applications, with an impact similar to that produced by the industrial revolution [1]. The impact of AI can be significant in many fields: finance, healthcare, manufacturing, retail, supply chain, logistics and utilities, all potentially disrupted by the onset of AI technologies [1].

New breakthroughs in algorithmic Machine Learning (ML) and autonomous decision-making are offering countless opportunities for innovation. ML methods have been providing breakthroughs in important research problems, which in some cases have been open for more than 50 years. This is the case, for instance, of the neural network-based model, AlphaFold, that predicts with high accuracy the three-dimensional

structure of a protein from its amino acid sequence [2]. Notably, AlphaFold is the first computational method that can predict protein structures even in cases in which no similar structure is known [2]. Another example is the success of AlphaZero, based on Reinforcement Learning (RL) principles [3], [4]. AlphaZero not only outperforms all chess programs, but it discovers new ways to play chess [5], [6]. In [7], ML methods are applied to predict human physical traits (e.g., face and voice) and other relevant information (e.g., height and weight) from genomic data. More recently, ML-based techniques provided critical support to combat the COVID-19 pandemic [8]; examples are the use of ML to analyze the epidemiological curve evolution [9], [10], [11] and medical images [12], [13].

A fundamental problem addressed by ML, which spans several research and application fields, is to discover intricate structures in large data sets, and in this context many applications make use of a class of techniques known as "deep learning" (DL) [14]. Conventional ML techniques were limited in their ability to process data in their raw form. For decades, constructing an ML system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data into suitable representations which the learning subsystem could infer patterns from [14]. Conversely, DL methods are representation-learning[1] methods with multiple levels of representation, obtained by composing simple non-linear modules, each of which transforms the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract, level. By the composition of enough such transformations, very complex functions can be learned, and in quite a few contexts DL methods represent nowadays the state of the art in terms of performance [14].

DL architectures based on Convolutional Neural Networks (CNNs) achieve unprecedented performance in skin cancer classification, with an accuracy comparable to that of dermatologists [15]. DL architectures based on Recurrent Neural Networks (RNN) are able to decode the electrocorticogram with high accuracy and at natural-speech rates [16]. These neural networks can be trained to encode sequences of neural activity into an abstract representation, and then decode this representation, word by word, into an English sentence with a 3% average word error rate [16]. DL and deep RL are also key components of new-generation autonomous driving systems, see, e.g., [17], [18], and nowadays are also widely exploited in surveillance systems, such as Synthetic Aperture Radar (SAR) imaging, see, e.g., [19], [20], [21], [22]. In space-based surveillance, DL offers the capability to accurately classify vessels from satellite sensor imaging (see, e.g., [23], [24]). In the context of maritime situational awareness and autonomous navigation (see, e.g., [25]) RNNs are able to accurately predict vessel positions several hours ahead [26], [27], [28], [29]. In video analysis and image understanding, DL methods represent the state of the art for object detection [30] and multi-object tracking [31]. DL is also used in Multiple Input Multiple Output (MIMO) communications [32], active sensing for communications [33], radar and sonar processing [34], [35], [36], [37], [38], [39].

Given the tremendous success of AI, the opportunities and challenges of merging AI and sensor data fusion are under investigation by several research groups with special focus on computational efficiency, improved decision making, security, multi-domain operations, and human-machine teaming [40]. Ethical aspects concerning the AI are also of paramount importance; in this regard, digital ethics for AI and information fusion in the context of the defense domain is discussed in [41].

B. RECENT ATTEMPTS TO MATHEMATICALLY FRAME DEEP LEARNING

Compared to the great success of DL techniques, little is known about their mathematical properties. There is still a lack of methodological and systematic approaches to analyze DL techniques [42], and no comprehensive understanding of the optimization process and internal organization of DL architectures is available; indeed, DL methods are usually regarded as "black boxes," see, e.g., [43].

In recent years, several attempts to fill this gap have been in progress, including those based on statistical and information theoretical interpretations [42], [44], [45], [46], [47]. In [44], [46], an interpretation of CNNs in terms of a cascade of filters implementing wavelet transforms and pointwise nonlinearities is provided. More recently, an asymptotic equivalence between *infinitely wide* deep neural networks and Gaussian processes is established exploiting the Central Limit Theorem (CLT) [48], [49]. In [48], as well as in the seminal work [50], the diverging parameter is the width of the hidden layers of the network.

The lack of methodological and systematic approaches to investigate the optimization process and internal organization of DL techniques also exists for the derivation of their ultimate performance limit. The authors of the present paper are not aware of any results—or even ongoing efforts—to fill this gap. In particular, it is currently unknown if DL techniques can or cannot achieve classification error probabilities close to optimal, i.e., vanishing exponentially with the data size. Classic statistical literature where the detection performance is studied in terms of large deviations include [51], [52], [53], [54], [55], [56], [57], [58].

C. CONTRIBUTIONS

The derivation of ultimate performance limits for ML classifiers is addressed here using a formulation that departs from the setting adopted in the classical statistical learning theory [59]. In particular, the parameters characterizing the ML procedure such as the NN weights, the network width, and the size of the training set are fixed; the diverging parameter n quantifies the information available for the classification problem (i.e., the number of measurements in a sequence of

[1]Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification.

data). For example, if one is trying to detect a target of extent 20 pixels in a 400 by 400 image, $n = 20$. And likewise if the image is 1000 by 1000 then n remains 20—that is, the background, at least to first order, does not matter. Continuing, if the target's extent is 30 samples and it is observed in 10 consecutive frames, then $n = 300$. Note also that the notional detectability (say, the Signal-to-Noise Ratio, SNR) of the target in the image does not impact n, as it will be seen to be subsumed within the convergence rate I.[2]

Data-centric approaches exploit large amounts of data. Part of the data, the so-called training set, is used to train the network. The trained network is then fed the observation set to perform the desired classification. Additional datasets, here referred to as characterization sets,[3] may also be available and are usually exploited to infer numerically the classification error probability of the trained network. Our vision is that these characterization sets can be exploited to derive *analytical expressions* for the classification error probability, which in turn allow inferring the performance in the limit of diverging n. Under mild assumptions that can be directly checked by inspection of the characterization sets, it is shown that the classification error probability converges to zero exponentially fast with n, at a predictable rate. Thus, performance prediction beyond standard numerical estimation of the classification error probability is possible and ultimate classification limits can be established.

The main contributions of this article can be summarized as follows.

- A mathematical framework based on the Large Deviations Principle (LDP) [62], [63] is presented for performance prediction of ML techniques when n diverges; the application domain of such a performance prediction tool includes, but it is not limited to, neural network-based architectures for classification.
- It is shown that the asymptotic classification error probability depends on a suitable transformation of the Log Moment-Generating Function (LMGF) of the decision statistic; in the case of DL, this decision statistic is determined by the output layer of the network.
- The mathematical conditions for a ML classifier to exhibit exponential classification error probability $\exp(-nI + o(n))$, where I is the error rate, are provided; such conditions depend on the Fenchel-Legendre transform of the LMGF of the data-driven decision statistic learned in the training phase.
- It is shown that the error rate I approaches the optimal classification rate as the size of the training set grows; such an optimal rate describes the classification error probability of the log-likelihood ratio test and admits information-theoretical interpretations.
- Based on saddlepoint techniques, refined asymptotic expressions (often referred to as exact asymptotics) are derived in the form $\zeta_n \exp(-nI)$, where ζ_n accounts for the most representative sub-exponential term of the classification error probability.

The paper is organized as follows. In Section II, we provide a summary of the main theoretical results. In Section III, we formulate the problem in terms of statistical hypothesis testing, with simple hypotheses and assuming independent samples. In Section IV, we provide the convergence rate and the exact asymptotic formula to approximate the error probabilities. In Section V, we extend the problem to composite hypotheses and dependent data. In Section VI, we provide details on the computation of the relevant parameters, related to the asymptotic approximations of the error probabilities, and in particular the rate function. In Section VII, we discuss extensive numerical simulations and experimental results using real-world data acquired by an X-band marine radar. Finally, the conclusion is given in Section VIII, and mathematical details are provided in appendices.

II. PREVIEW OF THE MAIN RESULTS

In this section, we briefly describe the main findings of this work. Consider a family of real-valued decision statistics $T^{(n)}$, $n \geq 1$, based on the observations $\mathcal{X}^{(n)}$. The parameter n has been exemplified earlier and will be defined later with more precision, but can be thought of as quantifying the number of informative elements per test datum. The goal is to decide between two hypotheses, i.e., $\mathcal{H}_0$ and $\mathcal{H}_1$. Since the distribution of the observations under $\mathcal{H}_0$ and $\mathcal{H}_1$ is often unknown, or too complex to derive, we focus on the case that the decision statistic is provided by a learning mechanism operating on a sufficiently large *training set* $\mathcal{Y}$ of finite size m_y available under each hypothesis, independent of $\mathcal{X}^{(n)}$. Then, the decision statistic $T^{(n)} = T^{(n)}_{\boldsymbol{\omega}}$ is referred to as the Data-Driven Decision Function (D3F), where the parameters $\boldsymbol{\omega}$ are learned at the training phase. When doable, the ideal case that $T^{(n)}$ is the Log-Likelihood Ratio (LLR), see, e.g., [64], is considered and this serves as a term of comparison with the D3F. Consider the statistical test

$$\begin{cases} T^{(n)} \geq \gamma_n & \text{decide } \mathcal{H}_1 \\ T^{(n)} < \gamma_n & \text{decide } \mathcal{H}_0, \end{cases} \tag{1}$$

where γ_n is the threshold, and the error probabilities are[4]

$$\alpha_n = \mathbb{P}\left[T^{(n)} \geq \gamma_n \,|\mathcal{H}_0\right], \quad \beta_n = \mathbb{P}\left[T^{(n)} < \gamma_n \,|\mathcal{H}_1\right]. \tag{2}$$

We study the detection performance of $T^{(n)}$ when n diverges, and propose suitable approximations for the finite sample-size regime of n. The detection performance is evaluated/characterized exploiting a distinct set of data of finite size m_z, independent of $\mathcal{Y}$ and $\mathcal{X}^{(n)}$, that we indicate with $\mathcal{Z}$ and refer to as the *characterization set*. The use of the characterization set can be seen as similar to the usage of the *validation set* (usually involved in the learning stage see, e.g. [60], [61]) to assess the ML performance. For our scope,

[2] In setting different from that presented in this work, the parameter n could also represent the step size in adaptive learning algorithms over decentralized networks [55], [56].

[3] Validation set is a more common terminology in a machine learning context [60], [61].

[4] The equality in the error events is immaterial, given that the decision statistic is assumed continuous if not otherwise stated.

an overlap between the characterization and training sets is acceptable as long as the estimators, introduced in Section VI, properly converge to their expected values, which are the quantities of interest to characterize the test performance.[5] In general, both sets $\mathcal{Y}$ and $\mathcal{Z}$ can depend on the parameter n, but this will be clarified case by case.

Let us define the LMGF of $T^{(n)}$ as

$$\varphi_{n,k}(t) = \log \mathbb{E}\left[\exp(t\, T^{(n)}) \,|\mathcal{H}_k\right], \tag{3}$$

where $\mathbb{E}[X|\mathcal{H}_k]$ is the expected value of X under $\mathcal{H}_k$, $k = 0, 1$. The mean and the variance of the decision statistic are given by the derivatives of the LMGF at zero, as follows

$$\varphi'_{n,k}(0) = \mathbb{E}\left[T^{(n)} \,|\mathcal{H}_k\right] = \mu_{n,k}, \tag{4}$$

$$\varphi''_{n,k}(0) = \mathbb{S}\left[T^{(n)} \,|\mathcal{H}_k\right]^2 = \sigma^2_{n,k}, \tag{5}$$

where $\mathbb{S}[X|\mathcal{H}_k]$ denotes the standard deviation of X under $\mathcal{H}_k$. Under mild regularity conditions, see, e.g., [65], [66], the asymptotic performance characterization depends on the limit of the scaled LMGF[6] [63]

$$\varphi_k(t) = \lim_{n\to\infty} \frac{1}{n}\varphi_{n,k}(n\,t). \tag{6}$$

Specifically, if the above limit exists under $\mathcal{H}_k$, usually two important properties hold. The former, referred to as *small deviations*, implies that $T^{(n)}$ converges (in some sense) for $n \to \infty$ to an asymptotic value μ_k under $\mathcal{H}_k$. The latter, referred to as *large deviations*, characterizes the convergence to zero of the probability that $T^{(n)}$ is "far away" from μ_k. The small deviations property is related to the convergence in distribution of the normalized statistic $\sqrt{n}(T^{(n)} - \mu_k)$ to a Gaussian random variable, which in our setting is $\mathcal{N}(0, \sigma_k^2)$, under $\mathcal{H}_k$. Then, we have

$$\sqrt{n}(T^{(n)} - \mu_k) \xrightarrow{d} \mathcal{N}(0, \sigma_k^2), \tag{7}$$

where $\xrightarrow{d}$ indicates the convergence in distribution, whereas μ_k and σ_k are respectively the asymptotic mean and the asymptotic standard deviation of the decision statistic. Such asymptotic values are related to (4), (5), and (6) as follows [65]

$$\varphi'_k(0) = \lim_{n\to\infty} \mu_{n,k} = \mu_k,$$
$$\varphi''_k(0) = \lim_{n\to\infty} n\,\sigma^2_{n,k} = \sigma^2_k. \tag{8}$$

The convergence in (7) is clearly related to the CLT, and this will be discussed in detail in the next sections, but a pictorial representation of this convergence is illustrated in Fig. 1. Now, let us focus on the latter property, i.e., the LDP. In view of

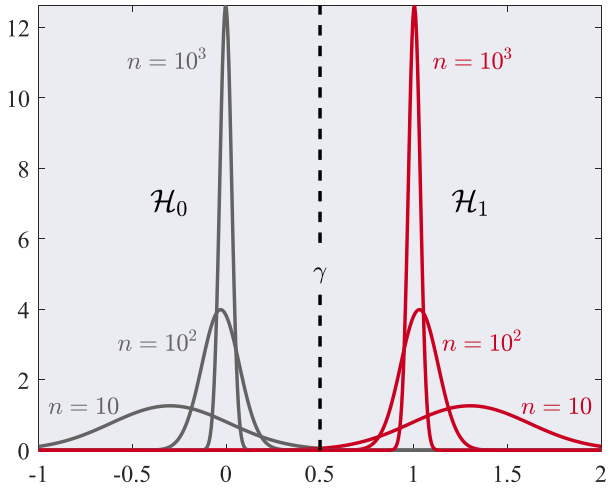


FIGURE 1. Pictorial representation of the distribution of $T^{(n)}$ for different values of n under $\mathcal{H}_k$, $k = 0, 1$. As expressed in (7), the behaviour is the same as the CLT. The distribution of $T^{(n)}$ is approximately Gaussian with mean converging to the asymptotic mean μ_k and variance $\frac{\sigma_k^2}{n}$ that decreases linearly with n. In this example we have $\mu_0 = 0$, $\mu_1 = 1$, and $\sigma_k = 1$ for $k = 0, 1$. The distribution of $T^{(n)}$ amasses around μ_k as n increases. We have represented the threshold, γ, of the test (1) in the mid point 0.5 between μ_0 and μ_1. It is easily seen that the error probabilities (2) converge to zero as n increases. However, it is worthwhile stressing that the rates of convergence of the error probabilities are not ruled by the convergence in distribution reported in this figure; instead, they are ruled by the LDP (9)-(10), involving the asymptotic behaviour of the tail probabilities.

the asymptotic convergence (7), $T^{(n)}$ gets closer and closer to μ_k under $\mathcal{H}_k$ as n increases, with a variance that vanishes, as illustrated in Fig. 1. Intuitively, in this framework, for a large enough n, we should set the threshold γ_n between μ_0 and μ_1 (otherwise we incur in the trivial situation that one of the error probabilities converges to unity). In other words, assuming that $\gamma_n \to \gamma$ and $\mu_0 < \mu_1$, it is required that $\mu_0 \le \gamma \le \mu_1$. Then, if $\varphi_k(t)$ is twice differentiable, the following large deviations results hold for $\gamma_n \to \gamma$ and n sufficiently large:

$$\alpha_n \approx \zeta_{n,0}(\gamma)\exp(-n\, I_0(\gamma)), \tag{9}$$

$$\beta_n \approx \zeta_{n,1}(\gamma)\exp(-n\, I_1(\gamma)), \tag{10}$$

where $I_k(\gamma)$ is referred to as the *rate function* and is given by the Fenchel-Legendre transform of the limit of the scaled LMGF[7] in (6). The terms $\zeta_{n,k}(\gamma)$ in (9)–(10) model sub-exponential behaviours and can be computed with the so-called "exact asymptotics," (see, e.g., [62]) related to the saddlepoint approximation (see, e.g., [67]). The Fenchel-Legendre transform rules the convergence of the tails of the

[5] More details are given in Section VI, and the mathematical conditions for the weak convergence are provided in Appendix B.

[6] The LMGF is also referred to as the cumulant generating function. Moreover, the dependence in (6) on the parameter n can be generalized. In particular n can be replaced by a generic deterministic sequence a_n (see details in [65]). The sequence a_n is referred to as the "convergence speed," and the speed is linear when $a_n = n$.

[7] In general, the rate function is given by the infimum value of the Fenchel-Legendre transform in the decision region Γ_k of $\mathcal{H}_k$; more details on this are provided in Section V.

TABLE 1 Error Probability Approximations

Panel A: Approximations

	Small deviations	Large deviations
α_n	$Q\left(\frac{\gamma_n - \mu_{n,0}}{\sigma_{n,0}}\right)$	$\zeta_{n,0}(\gamma)\, e^{-n\, I_0(\gamma)}$
β_n	$Q\left(\frac{\mu_{n,1} - \gamma_n}{\sigma_{n,1}}\right)$	$\zeta_{n,1}(\gamma)\, e^{-n\, I_1(\gamma)}$

Panel B: Asymptotic behaviour of parameters

	Small deviations			Large deviations
Finite sample-size	γ_n	$\mu_{n,k}$	$\sigma_{n,k}$	$\zeta_{n,k}$
Asymptotic behaviour	$\to \gamma$	$\to \mu_k$	$\sim \frac{\sigma_k}{\sqrt{n}}$	$\sim \frac{1}{\sqrt{n}}$

distribution of $T^{(n)}$, as formalized in (9)–(10). More details about the derivation of $\zeta_{n,k}$ are provided in Section IV-D; specifically, see (32) for the approximation of the error probabilities, as in (9)–(10).

Both the small deviations and the large deviations principles can be seen as two sides of the same medal; indeed, under mild regularity conditions (see, e.g., [68] and the discussion in [65]), the CLT can be obtained as a consequence of the LDP. In other words, the considerations that lead to the LDP are strictly related to those of the CLT.[8] For a quick consultation and comparison, we have collected in Table 1 the expressions of the approximate error probabilities based on the small deviations (7)–(8), and large deviations (9)–(10) principles.

Finally, it is worth mentioning that the small deviations approximation is useful to fix one of the error probabilities to a desired value, with the other one converging to zero at the maximum achievable rate given the convexity of the rate functions; more details on this aspect will be given in Section III-B and V-D. To give an idea, let us select $\gamma_n \to \mu_0$, obtaining $I_0(\mu_0) = 0$ and $I_1(\mu_0) \geq I_1(\gamma)$ for all $\mu_0 \leq \gamma \leq \mu_1$. This operational situation corresponds to the point on the y-axis (abscissa null) of Fig. 8.

Our goal is to derive the LDP for ML-based statistical hypothesis testing. We show that it is possible to compute numerically the error rate I_k, as well as the sub-exponential terms $\zeta_{n,k}$, from the characterization set $\mathcal{Z}$ without assuming knowledge of the original distributions.

We consider the following important scenarios, which are relevant in a number of applications, see, e.g., [38], [54], [55], [69], [70], [71], [72], [73], [74]. The first is perhaps simplistic, but it is convenient to motivate our development.

1) We begin by assuming, under $\mathcal{H}_k, k = 0, 1$, Independent and Identically Distributed (IID) observations $\mathcal{X}^{(n)} = (x_i)_{i=1}^n$, with simple hypotheses $x_i \sim f_k(x)$. We derive a decision statistic $T_{\boldsymbol{\omega}}^{(n)}$, which is the sample mean of the data in $\mathcal{X}^{(n)}$ processed according to the elementwise D3F, $t_{\boldsymbol{\omega}}(\cdot)$, as in the conceptual diagram illustrated in Fig. 3; this scenario is described in Section III and IV. In [75], we analyze the classification performance of a D3F using the popular MNIST (Modified National Institute of Standards and Technology) database.
2) We generalize the previous scenario to the case of composite hypotheses, with conditionally independent data given a parameter $\theta \in \Theta$. We have $x_i \sim f(x|\theta_0)$ with $\theta_0 \notin \Theta$ under $\mathcal{H}_0$ and $x_i \sim f(x|\theta)$ under $\mathcal{H}_1$, with θ distributed according to a prior discrete distribution w_θ, which can be inferred from the training data. In this case, we train a different elementwise D3F for each possible value of θ. Then, the complete decision statistic has a structure that involves the elementwise D3F applied to each sample in $\mathcal{X}^{(n)}$ for each possible value of θ, as illustrated by the conceptual diagram in Fig. 5; this scenario is described in Section V-B.
3) We further generalize to the case of dependent data under $\mathcal{H}_0$ and $\mathcal{H}_1$. This setting is particularly appealing also in the case of conditionally independent data and too large size of the space of θ, making the scheme in 2) intractable. Specifically, we focus on a scenario in which the input data $\mathcal{X}^{(n)}$ are binary images that may or may not contain a target, which occupies more than one pixel ("extended target"). The hypothesis $\mathcal{H}_0$ represents the absence of the target, and $\mathcal{H}_1$ represents its presence. The position, shape and orientation of the target are unknown in advance. In this scenario, the parameter n represents the number of pixels occupied by the target in the image. Some examples of the images used in this scenario are reported in Fig. 2, whereas the conceptual diagram of the D3F is illustrated in Fig. 6. Further details are available in Section V-C. In this scenario, using real-world data acquired by a high-resolution marine radar, we prove the effectiveness of the proposed architecture as well as the capability to predict its decision performance via the developed theoretical tools.

In the first two scenarios, we compare the performance of the D3F, implemented as a suitable fully-connected Neural Network (NN), with the ideal test based on the LLR. In the third scenario, the D3F is implemented as a Deep CNN (DCNN), and the D3F performance cannot be compared with the ideal LLR test because such a test is not available.

It is worth mentioning that the first two architectures implicitly exploit the statistical dependency structure of the input data. Our design approach, inspired by the LLR functional form, can be be framed within the model-based ML context, similar to, e.g., [76], [77], where ML methods take advantage of the statistical modeling of the considered problem, enhancing their data-driven vocation.[9]

[8] In the most general settings, the mathematical conditions required by the CLT and those required by LDP are different, being also mathematically different the two convergences. However, as already pointed out, in many relevant situations LDP and CLT are both verified.

[9] The nomenclature "model-based ML" was originally introduced by Bishop in [61] to define an approach to ML where all the assumptions about the problem domain are made explicit in the form of a model. Indeed, it takes

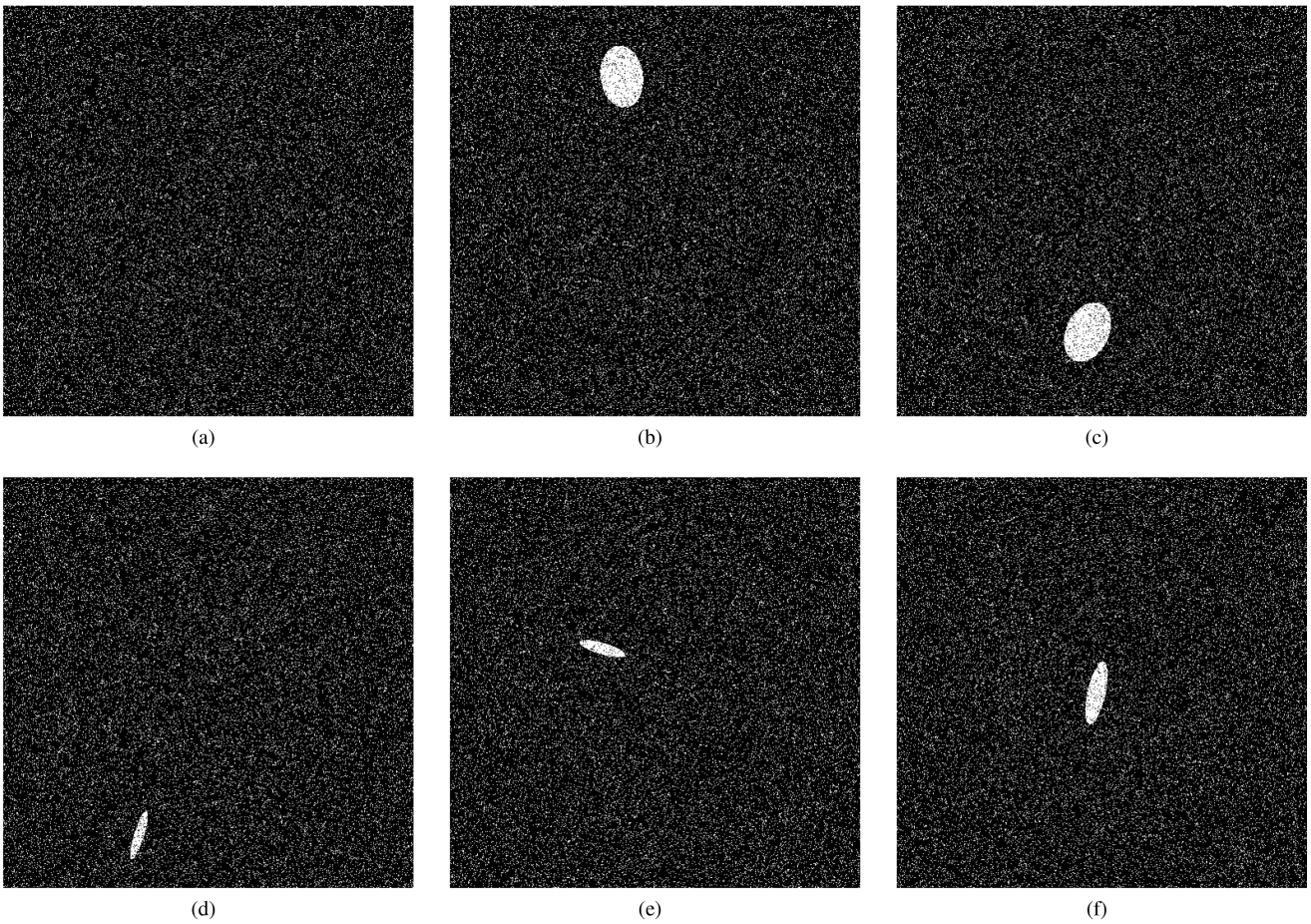


FIGURE 2. Notional examples of binary images that represent the input data: (a) under $\mathcal{H}_0$ (absence of target); and (b)–(f) under $\mathcal{H}_1$, with the target present in different locations with different orientations and shapes.

An important remark is related to the computation of the error rate functions and the sub-exponential terms, which are seldom available in closed form, even when the analytical distribution of the data is available. In our framework, given that the distribution of the data may not be available, we propose to estimate the scaled LMGF (6), the error rates, and the sub-exponential terms exploiting the characterization set. This is a distinct feature of the present paper.

Moreover, we will show that the numerical stability of such estimation can be improved with Monte Carlo Methods (MCMs), as discussed in depth in [65], [66]. Specifically, the use of MCMs requires principled generation of synthetic data from the so-called *tilted* distribution via the Metropolis–Hastings (MH) algorithm [65], [66]. It is worthwhile noting that the generation of synthetic data is a well-explored topic in the ML literature, see, e.g., variational autoencoders (VAEs) [78] and generative adversarial networks (GANs) [79], [80]. In the framework of model-based ML, instead, the generation of synthetic data is straightforward, given that the statistical model is assumed to be (at least partially) known. In this regard, it is also possible to train the D3F with real-world data, and then characterize the test with synthetic data assuming, for instance, different statistical models, or relevant parameters. The opposite option is also workable, with the training performed by synthetic data and the testing carried out by real data. Indeed, this is the approach taken for the third scenario in the above list: the D3F is trained on synthetic data, an example of which is reported in Fig. 2, but the performance is tested on real-world radar data.

the wider meaning of a bespoke method, formulated for new applications, where the solution is expressed through a compact modelling language, and the corresponding custom ML code is then generated automatically.

III. SMALL DEVIATIONS OF THE D3F WITH IID OBSERVATIONS

A. PROBLEM FORMULATION AND LEARNING MECHANISM

Let us consider the observations $\mathcal{X}^{(n)} = (x_i)_{i=1}^n$, where the x_i's are IID according to $f_0(x)$ under hypothesis $\mathcal{H}_0$ and $f_1(x)$ under hypothesis $\mathcal{H}_1$. The i-th observation $x_i \in \mathbb{X}$ is representative of a generic random variable, such as a real-valued random variable ($\mathbb{X} = \mathbb{R}$) as well as a binary pixel ($\mathbb{X} = \{0, 1\}$), or an optical image ($\mathbb{X} = \mathbb{R}^d$). The goal is to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$ based on the observed dataset $\mathcal{X}^{(n)}$. It is well known that the optimal decision statistic $L^{(n)}$ is the LLR [64], which takes the form of a summation in the

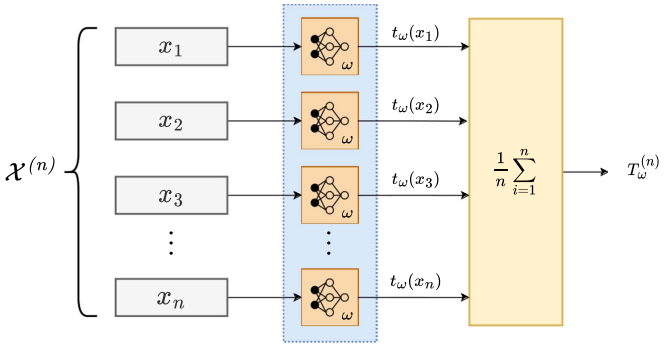


FIGURE 3. Conceptual diagram of the D3F $T_\omega^{(n)}$ (12) with IID observations. The elementwise D3F $t_\omega(x_i)$, $i = 1, 2, \ldots, n$, is pictorially represented as the output of a NN whose parameters are ω for each i (the NN is depicted in Fig. 4).

case of independent observations:

$$L^{(n)} = \frac{1}{n}\sum_{i=1}^{n} l(x_i) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_1(x_i)}{f_0(x_i)}. \tag{11}$$

As already discussed, in many practical applications, both distributions $f_k, k = 0, 1$, are unknown, and therefore the LLR test cannot be applied directly (in the present work, the LLR will be used as a term of comparison). However, given the independence assumption among the entries of $\mathcal{X}^{(n)}$, and inspired by the functional form of the optimal detector (11), in the following we construct the D3F exploiting a similar structure[10] as in (11). Precisely, the D3F for IID observations is as follows

$$T_{\boldsymbol{\omega}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} t_{\boldsymbol{\omega}}(x_i), \tag{12}$$

where $t_{\boldsymbol{\omega}}(x_i)$ represents a statistic of the datum x_i, parameterized by $\boldsymbol{\omega}$, and is denoted as the elementwise D3F. A pictorial block-diagram of the D3F (12) is provided in Fig. 3, therein the orange blocks are representative of the processing performed by the elementwise D3F $t_{\boldsymbol{\omega}}(x_i)$.

Evidently, if $t_{\boldsymbol{\omega}}(x_i) = l(x_i)$, the D3F boils down to the optimal detector (11); however, it demands the knowledge of the distributions f_k, $k = 0, 1$. Such distributions are often not available, and the idea is to appropriately learn $t_{\boldsymbol{\omega}}(x)$ from a labeled training set $\mathcal{Y} = \{\mathcal{Y}_0 \cup \mathcal{Y}_1\}$. Specifically, $\mathcal{Y}_k = \{y_{1,k}, y_{2,k}, \ldots, y_{m_y,k}\}$, $k = 0, 1$, whose entries are IID with $y_{j,k}$ distributed according to $f_k(\cdot)$, and m_y is the number of samples available for training under each hypothesis. The training set $\mathcal{Y}$ is assumed independent of the observed dataset $\mathcal{X}^{(n)}$ [81].

[10]For the case of *dependent* observations with an unknown dependency structure the additive functional form is in general too restrictive. In such situations the overall D3F architecture needs to be learned during the training stage, see details in Section V-C.

An example of elementwise D3F can be made starting from a NN. In our context, the NN processes the datum x_i and returns as output the pair $h_0(x_i)$ and $h_1(x_i)$. The NN is characterized by the parameters $\boldsymbol{\omega} = \{\omega_{j\ell}^{(p)}\}$, where $\omega_{j\ell}^{(p)}$ is the weight of the j-th entry of the input to the p-th NN layer, contributing to the ℓ-th entry of the layer output. This output is indicated with $h_\ell^{(p)}$, while $\omega_{0\ell}^{(p)}$ represents the bias term. A NN with a single hidden layer, scalar input x_i and two outputs is described by the following equations [81]

$$\begin{aligned} h_\ell^{(1)} &= g_1\left(\omega_{1\ell}^{(1)} x_i + \omega_{0\ell}^{(1)}\right) \\ h_k^{(2)} &= \sum_{j=1}^{M} \omega_{jk}^{(2)} h_j^{(1)} + \omega_{0k}^{(2)} \\ h_k &= g_{\sigma,k}\left(h_0^{(2)}, h_1^{(2)}\right), \end{aligned} \tag{13}$$

where M is the network width, and $g_1(\cdot)$ and $g_{\sigma,k}(\cdot)$ are suitable nonlinearities [81]. The equations above are valid for a scalar input, i.e., $x_i \in \mathbb{R}$. More in general, for a d-dimensional input $x_i \in \mathbb{R}^d$, whose j-th entry is denoted by $x_{i,j}$, a NN with N hidden layers is described by the following equations [81]

$$\begin{aligned} h_{\ell_1}^{(1)} &= g_1\left(\sum\nolimits_{j=1}^{d} \omega_{j\ell_1}^{(1)} x_{i,j}\right) \\ h_{\ell_2}^{(2)} &= g_2\left(\sum\nolimits_{j=1}^{M_1} \omega_{j\ell_2}^{(2)} h_j^{(1)}\right) \\ &\vdots \\ h_{\ell_p}^{(p)} &= g_p\left(\sum\nolimits_{j=1}^{M_{p-1}} \omega_{j\ell_p}^{(p)} h_j^{(p-1)}\right) \\ &\vdots \\ h_k^{(N+1)} &= \sum\nolimits_{j=1}^{M_N} \omega_{jk}^{(N+1)} h_j^{(N)} \\ h_k &= g_{\sigma,k}\left(h_0^{(N+1)}, h_1^{(N+1)}\right), \end{aligned} \tag{14}$$

where $p = 1, \ldots, N$ is the hidden layer index, $\ell_p = 1, \ldots, M_p$ identifies the neurons of the p-th hidden layer, and h_k, with $k = 0, 1$, indicates the network output (binary classification); the bias terms are omitted for notational simplicity. An example of such a network with two hidden layers is illustrated in Fig. 4.

A typical choice for the pair $g_{\sigma,0}(\cdot)$ and $g_{\sigma,1}(\cdot)$ is the softmax function; in this case, the network outputs $h_k(x_i)$ can be interpreted as the posterior class probabilities, i.e., the probabilities that x_i originated from $\mathcal{H}_k$. For this reason, $h_k(x_i)$ is indicated as $p_{\boldsymbol{\omega}}^{(\mathcal{H}_k)}(x_i)$, where we make explicit the dependency from the NN parameters $\boldsymbol{\omega}$. Assuming a uniform prior (such as in a balanced training set), the elementwise D3F statistic can be defined as the log ratio of the two network outputs, i.e.,

$$t_{\boldsymbol{\omega}}(x_i) = \log \frac{h_1(x_i)}{h_0(x_i)} = \log \frac{p_{\boldsymbol{\omega}}^{(\mathcal{H}_1)}(x_i)}{p_{\boldsymbol{\omega}}^{(\mathcal{H}_0)}(x_i)}, \tag{15}$$

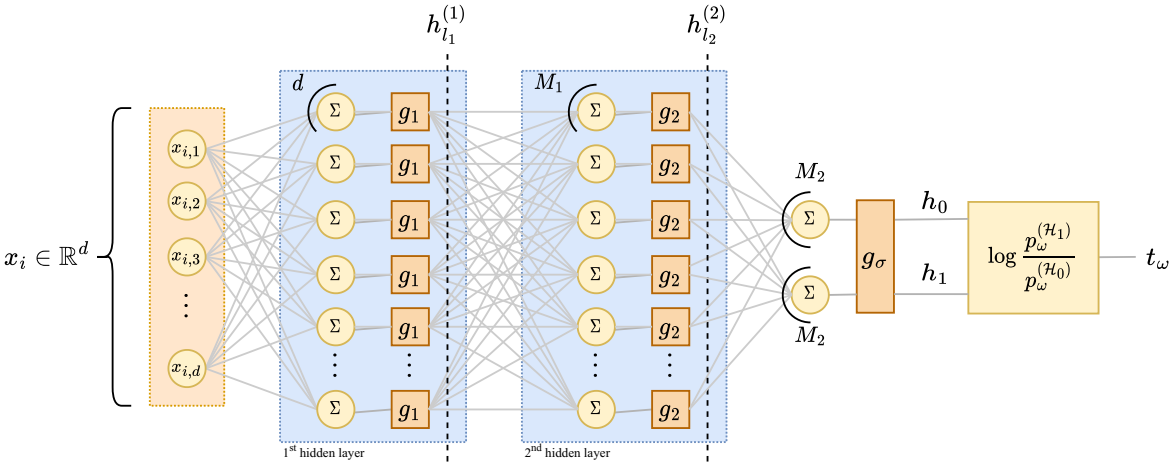


FIGURE 4. Illustration of the elementwise D3F, given by a generic fully-connected neural network with multidimensional input $x_i \in \mathbb{R}^d$ and two hidden layers. The bias terms in each hidden layer are omitted for simplicity. The final output $t_\omega(x_i)$ is the elementwise D3F, which processes the i-th observation x_i.

and can be interpreted as an approximated version of the elementwise LLR $l(x_i)$ in (11).[11]

The elementwise decision statistic $t_{\boldsymbol{\omega}}(x)$ is parameterized in $\boldsymbol{\omega}$, which can be learned (or estimated) by minimizing a suitable empirical loss function $\mathcal{E}$ [81]

$$\boldsymbol{\omega} = \arg\min_{\boldsymbol{\omega}'} \sum_{j,k} \mathcal{E}_k \left(t_{\boldsymbol{\omega}'}(y_{j,k}) \right), \tag{16}$$

where the summation is justified by the independence assumption among the observations of the training set. In binary decision problems, a suitable loss function is the binary cross-entropy loss, defined as [81]

$$\begin{aligned} \mathcal{E}_k \left(t_{\boldsymbol{\omega}}(y_{j,k}) \right) = & -k \log \left(p_{\boldsymbol{\omega}}^{(\mathcal{H}_1)}(y_{j,k}) \right) \\ & - (1-k) \log \left(p_{\boldsymbol{\omega}}^{(\mathcal{H}_0)}(y_{j,k}) \right). \end{aligned} \tag{17}$$

Alternative strategies to learn suitable elementwise D3F for binary detection are available in the recent literature; see, e.g., [82], [83]. The choice of the loss function and the resulting D3F depend heavily on the specific problem. For generality, the following analysis is not restricted to particular choices. From now on, the elementwise D3F $t_{\boldsymbol{\omega}}(x)$ will be considered a generic decision statistic function dependent on the parameter vector $\boldsymbol{\omega}$, learned from the dataset $\mathcal{Y}$.

As mentioned in Section II, we are going to exploit two asymptotic frameworks to study the classification performance of the D3F test; the first is based on the CLT [64], [84] for setting an asymptotic level of false alarm or missed detection; the second is based on the LDP [62], [63].

[11] When $g_{\sigma,k}$ is the softmax function, it can be easily verified that the computation of the data-driven decision statistic (15) amounts at computing the difference between the so-called logit values of class $\mathcal{H}_1$ versus $\mathcal{H}_0$, namely $h_0^{(N+1)}$ and $h_1^{(N+1)}$ in (14). The logit values are the raw (non-normalized) predictions that a classification model generates, which are ordinarily then passed to the softmax.

B. ASYMPTOTIC PROPERTIES BASED ON THE CLT

Let us define μ_k and σ_k, respectively, as the mean and the standard deviation, under hypothesis $\mathcal{H}_k$, of the elementwise statistic, which is $l(x_i)$ in the case of the LLR and $t_{\boldsymbol{\omega}}(x_i)$ in the case of the D3F. The aforementioned expectations are taken with respect to the observations $\mathcal{X}^{(n)}$, and with reference to the D3F, the parameters $\boldsymbol{\omega}$ are fixed, being learned during the training phase. Because of the additive nature of the decision statistics (11) and (12), referred to as $T^{(n)}$, we can invoke the CLT after a suitable normalization [64], [84]

$$\begin{aligned} \widetilde{T}_k^{(n)} &= \sqrt{n}\,(T^{(n)} - \mu_{n,k}), \\ \mu_{n,k} &= \mathbb{E}\left[T^{(n)} \,|\mathcal{H}_k\right] = \mu_k, \\ \sigma_{n,k} &= \mathbb{S}\left[T^{(n)} \,|\mathcal{H}_k\right] = \frac{\sigma_k}{\sqrt{n}}, \end{aligned} \tag{18}$$

under the hypothesis $\mathcal{H}_k$, $k = 0, 1$. Assuming that mean and variance are finite, then the CLT holds, and the normalized statistic $\widetilde{T}_k^{(n)}$ converges in distribution to a Gaussian random variable [64], [84]

$$\widetilde{T}_k^{(n)} \xrightarrow{d} \mathcal{N}(0, \sigma_k^2), \quad \text{under } \mathcal{H}_k,\ k = 0, 1. \tag{19}$$

In the case of the LLR, $T^{(n)} = L^{(n)}$, $\mu_1 = \mathcal{D}(f_1||f_0)$ and $\mu_0 = -\mathcal{D}(f_0||f_1)$, where $\mathcal{D}(p||q)$ is the Kullback-Leibler (KL) divergence between p and q [64]. For the D3F, $T^{(n)} = T_{\boldsymbol{\omega}}^{(n)}$, and the moments μ_k and σ_k^2 do not exhibit closed form expressions; thus they need to be computed numerically exploiting the characterization set.

Remark: Under the summation form in (12) and given the independence assumption, the convergence shown in (19) stems from the well known CLT. However, as discussed in Section II, under mild regularity conditions the CLT can be derived as a consequence of the LDP [68] without the additive structure and the statistical independence among the data. In this respect we will show (exploiting a numerical analysis) that such convergence seems to hold even if the data are not statistical independent and the D3F statistic does not exhibit

a summation structure but a more complicated function provided by a DCNN.

C. THRESHOLD SETTING

Let us consider the test (1) and the error probabilities (2), where $T^{(n)} = L^{(n)}$ for the LLR and $T^{(n)} = T_{\boldsymbol{\omega}}^{(n)}$ for the D3F. The threshold γ_n is usually chosen to operate at a desired false alarm probability level. Exploiting the CLT convergence (19), we can set the asymptotic false alarm probability to a given value $0 < \alpha < 1$, when n is large enough, as follows:

$$\gamma_n = \mu_{n,0} + \sigma_{n,0} Q^{-1}(\alpha), \tag{20}$$

where $\mu_{n,0}$ and $\sigma_{n,0}$ are defined in (18) and $Q^{-1}(\cdot)$ is the inverse of the Q-function, i.e., $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{u^2}{2}\right) du$. In fact, according to the CLT convergence in (19) and the threshold in (20), we have that[12]

$$\lim_{n\to\infty} \alpha_n = \alpha. \tag{21}$$

The aforementioned convergence (21) is clearly in agreement with the small deviations approximation in Table 1, being both of them derived from the CLT. Indeed, from Table 1 (small deviations column), with γ_n defined as in (20), we get that $\alpha_n \approx \alpha$.

In (20) we have implicitly assumed known $\mu_{n,0} = \mu_0$ and $\sigma_{n,0} = \sigma_0/\sqrt{n}$; clearly, in our setting both μ_0 and σ_0 are unknown and have to be estimated from the characterization set.

The expression of the threshold (20) is the summation of two terms: $\mu_{n,0} = \mu_0 = O(1)$ and $\sigma_{n,0} Q^{-1}(\alpha) = O(1/\sqrt{n})$. As it will be clear in the following section, the second term is negligible in terms of error rate under $\mathcal{H}_1$, but it is instrumental to ensure that the asymptotic false alarm level is α.[13]

An alternative choice for setting the threshold is $\gamma_n = \gamma$, with $\mu_0 < \gamma < \mu_1$. For the LLR test, based on the theory of large deviations [62], [63] this choice would cause both the error probabilities to vanish exponentially. Our goal is to study the conditions for the D3F to exhibit exponentially vanishing error probabilities in n, and compute the convergence rate of the test. To this end, in the next section we will still assume independence of the observations, but we will then remove this assumption in Section V.

IV. LARGE DEVIATIONS OF THE D3F WITH IID OBSERVATIONS

Large deviations analysis of statistical hypothesis testing is well established in the statistics literature (see, e.g., [62], [63]), and employed in several applications [85]; for instance, in sensor networks based on the running consensus paradigm [71], [86]. The assumption of IID observations is quite common in the sensor network literature and useful to establish theoretical results, with large deviations convergences assessed in [54], [87], [88], whereas in adaptive networks [89] large deviations are demonstrated for the steady-state distribution (slow adaptation regime) [55], [56]. In the following, the fundamental theorems are reviewed and then applied to the test error probabilities (2).

A. LARGE DEVIATIONS PRINCIPLE

Consider a sequence of real-valued IID random variables $z_1, z_2, \ldots$, and their sample mean $S^{(n)} = \frac{1}{n}\sum_{i=1}^n z_i$. Cramér's theorem establishes the conditions for the sample mean to "rarely" deviate from its expected mean $\mu = \mathbb{E}[z_1]$. In the case that $\mathbb{P}[S^{(n)} \geq \gamma]$, with $\gamma > \mu > 0$, vanishes exponentially with n, we say that $S^{(n)}$ obeys the LDP. The exponential rate of convergence depends on the threshold γ, and is often referred to as the rate function, which is provided by Cramér's theorem.

Theorem IV.1 (Cramér's theorem [63]): consider a sequence of IID random variables $z_1, z_2, \ldots$, with a LMGF

$$\varphi(t) = \log \mathbb{E}\left[\exp(t\, z_1)\right] < \infty \qquad \forall t \in \mathbb{R}. \tag{22}$$

Let $S^{(n)} = \frac{1}{n}\sum_{i=1}^n z_i$. Then, for all $\gamma > \mu$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left[S^{(n)} \geq \gamma\right] = -I(\gamma), \tag{23}$$

where $I(x)$ is the rate function, given by the Fenchel-Legendre transform of the LMGF, i.e.,

$$I(x) = \sup_{t\in\mathbb{R}} \left[x\, t - \varphi(t)\right]. \tag{24}$$

We will not report here the proof of Cramér's theorem, for which we refer the reader to [63], but it is worth mentioning that it is based on the *squeeze theorem*, where the tightest Chernoff bound represents the upper bound. Such an upper bound is easily derived. Let us assume that $\gamma > \mu$ and consider the Chernoff bound on $S^{(n)}$, i.e.,

$$\mathbb{P}\left[S^{(n)} \geq \gamma\right] = \mathbb{P}\left[\sum_{i=1}^n z_i \geq n\gamma\right] \leq \frac{\mathbb{E}\left[\exp\left(t z_1\right)\right]^n}{\exp\left(nt\gamma\right)}, \tag{25}$$

for $t \geq 0$. By applying the logarithm to both sides of the previous equation, we obtain

$$\log \mathbb{P}\left[\sum_{i=1}^n z_i \geq n\gamma\right] \leq n \log \mathbb{E}\left[\exp\left(t z_1\right)\right] - nt\gamma. \tag{26}$$

Since the bound holds for every $t \geq 0$, we can rewrite the previous equation as

$$\frac{1}{n} \log \mathbb{P}\left[S^{(n)} \geq \gamma\right] \leq -\sup_{t\geq 0} \left(\gamma t - \overbrace{\log \mathbb{E}\left[\exp\left(t z_1\right)\right]}^{\varphi(t)}\right)$$
$$= -\underbrace{\sup_{t\in\mathbb{R}} \left(\gamma t - \varphi(t)\right)}_{I(\gamma)},$$

[12] If the threshold is set with (20), the miss detection probability of the LLR test vanishes exponentially fast with the best possible exponent, which is given by the KL divergence, see Stein's Lemma [62]. This behaviour is represented in Fig. 8 that will be discussed later.

[13] See the numerical simulations in Section VII-A, where the false alarm level is converging to a desired value for the case of IID observations; to this end, we use (20) to set the threshold in the uppermost plots of Fig. 9 for both the LLR and D3F tests.

where the last inequality stems from the fact that $\gamma t - \varphi(t)$ is a concave function whose derivative at $t = 0$ is $\gamma - \mu > 0$.

Remark: In the general formulation of the LDP [55], [62], [63], the convergence rate associated to $\mathbb{P}[S^{(n)} \in A]$ is given by the infimum point of the Fenchel-Legendre transform within the region A, namely $\inf_{x \in A} I(x)$. In the case of Cramér's theorem, we have $A = [\gamma, \infty)$. Let us observe that the unconstrained minimum of $I(x)$ is attained at $x = \mu$ (with $I(\mu) = 0$). As a consequence, being $I(x)$ strictly convex (see the discussion in [63]), if $\gamma \geq \mu$, then $\inf_{x \geq \gamma} I(x) = I(\gamma)$.

The application of Cramér's theorem will provide for the error rate functions of the error probabilities in (2).

B. LDP OF THE LLR TEST

We report the error rate function of the LLR test in the following theorem [62], [63] for continuous probability distributions. For discrete probability distributions, the interested reader is referred to Blahut [51] for binary hypothesis testing, and to Gutman [52] for M-ary hypothesis testing with unknown distributions.

Theorem IV.2 (Optimal error rate function): Let $\gamma \in (\mu_0, \mu_1)$, with $\mu_0 = -\mathcal{D}(f_0||f_1)$ and $\mu_1 = \mathcal{D}(f_1||f_0)$ being the KL divergences introduced in the previous section, then the error probabilities of the LLR test in (1) with $T^{(n)}$ given by (11) and $\gamma_n = \gamma$ exhibit the following error rate functions

$$\lim_{n \to \infty} \frac{1}{n} \log \alpha_n = -I_0(\gamma), \qquad \lim_{n \to \infty} \frac{1}{n} \log \beta_n = -I_1(\gamma), \quad (27)$$

where $I_k(\gamma)$ is the Fenchel-Legendre transform of the LMGF of the elementwise LLR $l(x)$ under $\mathcal{H}_k$.

The proof of the theorem can be easily derived from Theorem 4.1. It can be shown that the Fenchel-Legendre transform $I_1(\gamma)$ is related to $I_0(\gamma)$ via $I_1(\gamma) = I_0(\gamma) - \gamma$ (see details in [62], [63]). Besides, according to Stein's Lemma (see details in [62]), forcing $\alpha_n < \epsilon$ the best exponent for β_n is given by the KL divergence $\mathcal{D}(f_1||f_0)$; this exponent can be achieved by setting γ_n as in (20).

C. LDP OF THE D3F TEST

In this subsection, we provide the mathematical conditions that the D3F should satisfy in order to exhibit non-zero error rate functions. Specifically, let us observe that the D3F produces a transformation of the observed samples as follows

$$\tau_i = t_{\boldsymbol{\omega}}(x_i), \quad \forall i = 1, 2, \ldots, n. \quad (28)$$

The LDP of the D3F test can be established applying Cramér's theorem to the sequence of IID random variables τ_i.

Theorem IV.3 (Error rate function of the D3F test): Let $\gamma \in (\mu_0, \mu_1)$, with $\mu_k = \mathbb{E}[\tau_i|\mathcal{H}_k]$, with $k = 0, 1$, where the expectation is taken with respect to the measurements x_i. Then, the error probabilities of the D3F test defined in (1), with $T^{(n)}$ given in (12) and $\gamma_n = \gamma$, exhibit the following error rate functions[14]

$$\lim_{n \to \infty} \frac{1}{n} \log \alpha_n = -I_0(\gamma), \qquad \lim_{n \to \infty} \frac{1}{n} \log \beta_n = -I_1(\gamma), \quad (29)$$

where $I_k(\gamma)$ is given by the Fenchel-Legendre transform of the LMGF of $\tau_i = t_{\boldsymbol{\omega}}(x_i)$ under $\mathcal{H}_k$, with $k = 0, 1$.

Proof: The result can be established resorting to Theorem 4.1. For the case of α_n we only need to substitute $S^{(n)}$ in Theorem 4.1 with the decision statistic (12), and the random variables $\tau_i = t_{\boldsymbol{\omega}}(x_i)$ in place of z_i in Theorem 4.1. For the case of β_n, this error probability can be written as follows

$$\beta_n = \mathbb{P}\left[-T^{(n)}_{\boldsymbol{\omega}} > -\gamma \,|\mathcal{H}_k\right]. \quad (30)$$

Consequently, it is sufficient to invoke Theorem 4.1 with $z_i = -\tau_i$, and consider the Fenchel-Legendre transform $\widetilde{I}_1(x)$ of the LMGF of $-\tau_i$ (under $\mathcal{H}_1$) evaluated at $x = -\gamma$. Finally, it is straightforward to verify that $\widetilde{I}_1(-\gamma)$ is equivalent to the Fenchel-Legendre transform of the LMGF of τ_i at γ. □

The interpretation of Theorem 4.3 is that, neglecting the sub-exponential terms, the performance of the test based on the D3F depends only on $I_k(\gamma)$, which is non-zero for $\mu_0 < \gamma < \mu_1$. Even if the LMGF is not available, we can estimate it from the characterization set and thus compute the Fenchel-Legendre transform (see Section VI), which eventually leads to $I_k(\gamma)$. Moreover, resorting to the generalization of Cramér's theorem, namely the Gärtner-Ellis theorem [62], [63], which will be enunciated in the next section, it is not difficult to show that (23) holds true even if γ is replaced by γ_n, provided that $\gamma_n \to \gamma$. Otherwise stated, if $\gamma_n \to \gamma$, then $\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}[S^{(n)} \geq \gamma_n] = -I(\gamma)$. As a consequence, considering γ_n defined as in (20), we have that $\gamma_n \to \mu_0$ and $\frac{1}{n} \log \alpha_n \to I_0(\mu_0) = 0$, which complies with the fact that $\alpha_n \to \alpha > 0$ by construction. We expect then that β_n vanishes exponentially, i.e., $I_1(\mu_0) > 0$.[15]

The rate functions describe the error probabilities' scaling laws to zero; however, in practical applications it is important to have a good approximation of the entire error probability curves with respect to n. This happens, for instance, when the asymptotic false alarm probability is controlled and fixed to a level α while we want to approximate β_n; in this case, the error rate of β_n is exactly the same for all non-null values of α, even if the curves themselves would be quite different. In the next subsection we will introduce a method to approximate the entire error probability curve thanks to the so-called "exact asymptotics," which is closely related to the saddlepoint approximation.

[14] Note that there is a slight abuse of notation given that the D3F error rates (29) are formally different from the LLR error rate (27).

[15] Numerical simulations in the case of IID observations are provided in Fig. 8, and the details are provided in Section VII-A.

D. EXACT ASYMPTOTICS AND SADDLEPOINT APPROXIMATION

Theorem 4.4, provided in [62], is important to approximate the error probability curves for finite sample-size regime of n.

Theorem IV.4 (Bahadur and Rao): Let $S^{(n)} = \frac{1}{n}\sum_{i=1}^{n} z_i$, where z_i are IID real valued random variables with LMGF $\varphi(t) = \log \mathbb{E}[\exp(t\, z_1)]$. If the law of z_1 is non-lattice,[16] then

$$\lim_{n\to\infty} J_n \mathbb{P}\left[S^{(n)} \geq \gamma\right] = 1, \tag{31}$$

where $J_n = t_\gamma \sqrt{2\pi n \varphi''(t_\gamma)}\, \exp(n\, I(\gamma))$, t_γ is the solution of the equation $\gamma = \varphi'(t_\gamma)$, and $I(x)$ is the Fenchel-Legendre transform of the LMGF (24).

We shall assume that $\gamma > \mu$, where μ is the expected mean of z_i; if this is not the case, as already discussed, the rate is zero. Then, the following approximation for the probability of a rare event, i.e., $\mathbb{P}[S^{(n)} \geq \gamma]$, referred to as "exact asymptotics" or "saddlepoint approximation," can be derived from Theorem 4.4

$$\mathbb{P}[S^{(n)} \geq \gamma] \approx \underbrace{\zeta_n}_{\substack{\text{sub-exponential}\\ \text{terms}}} \underbrace{\exp\left(-n\, I\left(\gamma\right)\right)}_{\text{exponential term}}$$
$$\zeta_n = \left(t_\gamma \sqrt{2\pi n \varphi''(t_\gamma)}\right)^{-1}, \quad t_\gamma : \varphi'(t_\gamma) = \gamma, \tag{32}$$

where t_γ is the solution of the equation $\gamma = \varphi'(t_\gamma)$, representing the solution to the optimization problem involved in the definition of the Fenchel-Legendre transform of the LMGF (24). The approximation (32) takes into account both the sub-exponential contribution and the exponential term provided by the LDP.[17] The term ζ_n in (32) can be replaced by an asyptotically equivalent term c_n, with $c_n/\zeta_n \to 1$ (see details in the proof of [62, Theorem 3.7.4]). The term c_n is defined as follows [62]

$$c_n = \int_0^\infty e^{-t} \left[Q(0) - Q\left(\frac{\zeta_n}{\sqrt{2\pi}}\, t\right)\right] dt, \tag{33}$$

where $Q(\cdot)$ is the Q-function.[18]

An alternative derivation of (32) can be obtained by the saddlepoint approximation (see Daniels' pioneering work in [92] and Reid's overview in [67]), where the value t_γ is often referred to as the saddlepoint [67]. It is worthwhile noting that in both Theorem 4.4 and the saddlepoint approximation the higher order terms are neglected as not asymptotically relevant.

The approximation (32), either referred to as the saddlepoint approximation or exact asymptotics, does not need the IID assumption to provide a faithful description [90]. Indeed, we exploit their general formulation to handle the case of dependent observations, as detailed in the next section.

Equation (32) lays the ground for the approximation of α_n and β_n in (9)–(10) in the finite sample-size regime of n. Indeed, α_n in (9) refers to $\mathbb{P}[S^{(n)} \geq \gamma]$ by definition, with $S^{(n)}$ given by $T^{(n)}$ (12). Moreover, similar to the proof of Theorem 4.3, the error probability β_n can be studied analyzing $\mathbb{P}[-T^{(n)} > -\gamma|\mathcal{H}_1]$, as in (30). In this case, while the rate function is still given by the Fenchel-Legendre transform of the LMGF of τ_i in γ, the other parameters involved in the exact asymptotics demand the LMGF of $-\tau_i$ in $-\gamma$. The obtained approximations of the error probabilities α_n and β_n are reported in Table 1 (large deviations column).

When the threshold is set as in (20), the approximate error probability (32) can be further refined to take into account the decaying term $1/\sqrt{n}$ in the threshold (20) embedded in $\sigma_{n,0}$. A possible way to compute this refinement is to exploit the saddlepoint approximation at each value of n, which is valid in view of the "non-asymptotic" regime of the saddlepoint approximation [90]. Basically, in (32) we replace $I(\gamma)$ with $I_n(\gamma_n)$, which clearly converges to $I(\gamma)$ for n large enough; we proceed similarly with all the other related parameters in ζ_n. This refinement is especially relevant in the case of dependent data considered in the next section and further discussed in Section VI.

V. LARGE DEVIATIONS OF THE D3F STATISTICAL HYPOTHESIS TESTING IN THE GENERAL CASE

A. THE GÄRTNER-ELLIS THEOREM AND THE DUALITY PROPERTY

Consider a generic sequence of random quantities $\mathcal{X}^{(n)}$ and the resulting sequence of decision statistics $T^{(n)} = T^{(n)}(\mathcal{X}^{(n)})$. The generalization of Cramér's theorem is given by the Gärtner-Ellis theorem [62], [63], where $\varphi(t)$ is the limit of the scaled LMGF (6).

Theorem V.1 (Gärtner-Ellis theorem): Consider a sequence of random variables $z_1, z_2, \ldots$, with an asymptotic scaled LMGF

$$\varphi(t) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp(n\, t\, z_n)\right] < \infty \qquad \forall t \in \mathbb{R}. \tag{34}$$

If $\varphi(t)$ is differentiable in $\mathbb{R}$, then z_n obeys the LDP

$$\lim_{n\to\infty} \log \mathbb{P}\left[z_n \in A\right] = -\inf_{x\in A} I(x), \tag{35}$$

where A is a closed subset of $\mathbb{R}$ and the rate function $I(x)$ is given by the Fenchel-Legendre transform of $\varphi(t)$.[19]

[16] The random variable z_1 has a lattice law if for some z_0 and d, the random variable $d^{-1}(z_1 - z_0)$ is (a.s.) an integer number, and d is the largest number with this property. The formulation of the theorem for lattice law is available in [62], but not reported here for brevity.

[17] An alternative is the Lugannani-Rice formula (omitted for brevity); see, e.g., [90]. We have observed that, in the scenarios studied in this work, the Lugannani-Rice formula is numerically equivalent to (32). In the case of discrete distributions a similar expression for the non-asymptotic regime analysis is provided in [91].

[18] The numerical simulation in which we test the error probability approximation given by the exact asymptotics (32), in the case of IID observations, is provided in the lowermost plots of Fig. 9 for both the LLR and the D3F tests.

[19] We adopt a simplified version of the Gärtner-Ellis theorem compared to [62], [63]. The main difference is that in this paper the "limit inferior" and the "limit superior" of the general theorem are assumed equal, and hence the definition of the large deviation principle involves a simple limit. This simplification is tailored with the engineering scope of our paper and is aligned with more application oriented literature; see, e.g., [55], [65].

$$L^{(n)} = \frac{1}{n}\log w_{\theta_*}\prod_{i=1}^{n}\frac{f(x_i|\theta_*)}{f(x_i|\theta_0)} + \frac{1}{n}\log\left(1+\frac{\sum_{\theta\in\Theta,\theta\neq\theta_*} w_\theta \prod_{i=1}^{n} f(x_i|\theta)}{w_{\theta_*}\prod_{i=1}^{n} f(x_i|\theta_*)}\right)$$

$$= \underbrace{\frac{1}{n}\log w_{\theta_*}}_{\to 0} + \underbrace{\frac{1}{n}\log\prod_{i=1}^{n}\frac{f(x_i|\theta_*)}{f(x_i|\theta_0)}}_{L_{\theta_*}^{(n)}} + \frac{1}{n}\log\left(1+\underbrace{\sum_{\theta\in\Theta,\theta\neq\theta_*}\frac{w_\theta}{w_{\theta_*}}\exp\left[-n\left(\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_*)}{f(x_i|\theta)}\right)\right]}_{R_n}\right), \tag{37}$$

An important property of the Fenchel-Legendre transform holds when $\varphi(t)$ is differentiable, strictly convex, and diverging at $\pm\infty$ (see the discussion in [65] and references therein). In this case, the Fenchel-Legendre transform reduces to the Legendre transform, see [65], and there exists a unique root of $\varphi'(t) = x$, which is denoted by t_x, leading to

$$I(x) = t_x x - \varphi(t_x). \tag{36}$$

Therefore, in this case the slopes of $\varphi(t)$ are one-to-one related to the slopes of $I(x)$. This property is referred to as the duality property of the Legendre transform (see more details in [65]) and is automatically verified for IID observations in Section IV. In the case that $\varphi(t)$ is twice differentiable and is strictly convex ($\varphi''(t) > 0$), then $I(x)$ must also be strictly convex, given that differentiating (36) we have $I''(x) = \frac{1}{\varphi''(t_x)}$, and the curvature of $I(x)$ is the inverse curvature of $\varphi(t)$ [65]. Given that $I(x)$ is strictly convex and attains a global minimum in the asymptotic mean of z_n, then the infimum in (35) can be easily handled as in Cramér's theorem. If $A = [\gamma, \infty)$, then the LDP in (35) is given by $I(\gamma)$ for γ larger than the asymptotic mean of the sequence z_n.

In the following, we will exploit the Gärtner-Ellis theorem under both hypotheses for a D3F, which can take a generic structure, not necessarily a summation. In the first case study, we consider data that are conditionally independent eqaution (37) shown at the top of this page.

B. CONDITIONALLY INDEPENDENT OBSERVATIONS: COMPOSITE HYPOTHESIS TESTING

Let us consider the case of a composite hypothesis under $\mathcal{H}_1$. Precisely, we assume that the observations $\mathcal{X}^{(n)} = (x_i)_{i=1}^n$ are conditionally independent given a parameter $\theta \in \Theta$ under $\mathcal{H}_1$. The hypothesis under $\mathcal{H}_0$ is simple (the generalization to the composite case is similar). Then, we have $x_i \sim f(\cdot|\theta)$ under $\mathcal{H}_1$ and $x_i \sim f(\cdot|\theta_0)$ under $\mathcal{H}_0$ with $\theta_0 \notin \Theta$.

The general framework of composite hypothesis testing is available in [62], where an universal hypothesis testing procedure is proposed, and studied in terms of large deviations. The asymptotic optimality of the popular Generalized Likelihood Ratio Test (GLRT) is investigated in [53], and its conditions in the Neyman-Pearson sense are studied and discussed. In our framework, where the distributions are unknown, we cannot rely on the GLRT approach or the universal detector. Besides, for the practical needs of the machine learning procedure it is convenient to assume a prior on the parameter θ, according to a Bayesian formulation of the hypothesis testing problem. Specifically, we adopt an embedded learning architecture to mimic the LLR with the aforementioned prior. In this setting the LLR is given by:

$$L^{(n)} = \frac{1}{n}\log\frac{f(\mathcal{X}^{(n)}|\mathcal{H}_1)}{f(\mathcal{X}^{(n)}|\mathcal{H}_0)} = \frac{1}{n}\log\frac{\sum_{\theta\in\Theta} w_\theta\prod_{i=1}^{n} f(x_i|\theta)}{\prod_{i=1}^{n} f(x_i|\theta_0)}, \tag{38}$$

where we have assumed a finite support of θ, namely $|\Theta| < \infty$, and a non trivial prior $w_\theta > 0$. Clearly, the case of an infinite support of θ can also be handled making further regularity assumptions on the prior. In the case that the parameter is continuous, the summation is replaced by an integral.

For n sufficiently large, under $\mathcal{H}_1$ and given the true value of θ, indicated with θ_*, the mixture distribution in the numerator of (38) is "close," in terms of KL divergence, to the actual conditional distribution of the observations (see also the discussion in [93], [94]).

At this point, to handle the LDP, we need to compute the limit of the scaled LMGF (34) associated with the LLR, under both the hypotheses. Specifically, we assume that under $\mathcal{H}_1$ all the data are drawn from $f(\cdot|\theta_*)$, where θ_* indicates the true value of θ, while under $\mathcal{H}_0$ all the data are drawn from $f(\cdot|\theta_0)$. As shown in Appendix A, assuming some regularity conditions, the limit of the scaled LMGF of the LLR (38) is then given by:

$$\varphi_k(t) = \lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\exp(n\,t\,L^{(n)})\,|\mathcal{H}_k\right] = \begin{cases}\log\mathbb{E}\left[\exp\left(t\,\log\frac{f(x_1|\theta_*)}{f(x_1|\theta_0)}\right)\right] & \text{under } \mathcal{H}_1,\\ \log\mathbb{E}\left[\exp\left(t\,\log\frac{f(x_1|\theta_m)}{f(x_1|\theta_0)}\right)\right] & \text{under } \mathcal{H}_0,\end{cases} \tag{39}$$

where the expectation in $\varphi_k(t)$ is taken assuming that the data are generated according to $f(\cdot|\theta)$ with $\theta = \theta_* \in \Theta$ under $\mathcal{H}_1$ and $\theta = \theta_0$ under $\mathcal{H}_0$; moreover, θ_m is the closest value of θ to θ_0 in terms of the KL divergence. (39) reveals that the prior distribution w_θ does not affect the asymptotic LMGF, which is in agreement with classic Bayesian formulations (see, e.g., [95]).

Let us introduce the D3F for conditionally independent observations. Assume that the elementwise D3F $t_{\boldsymbol{\omega}_\theta}(x, \theta)$,

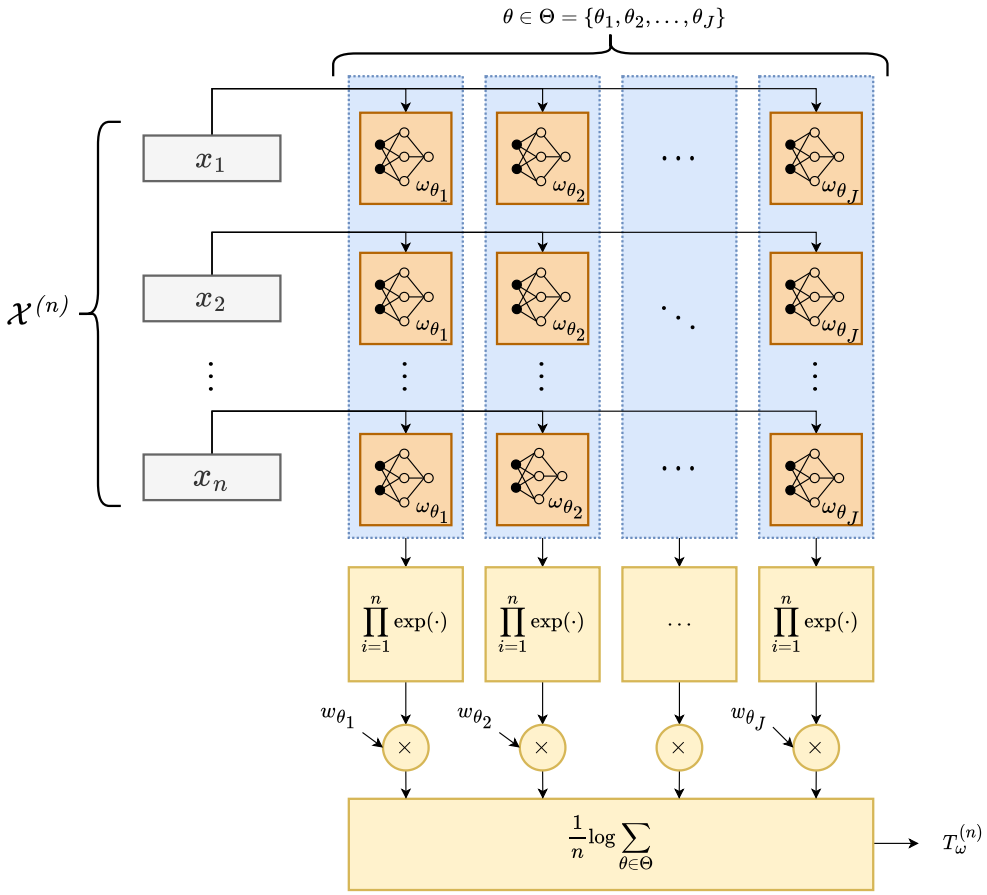


FIGURE 5. Conceptual diagram of the D3F $T_{\omega}^{(n)}$ with conditionally independent observations, see (40). The elementwise D3F $t_{\omega_\theta}(x, \theta)$, conditioned to a specific value of θ, is pictorially represented as a NN.

which approximates the elementwise LLR $\log \frac{f(x|\theta)}{f(x|\theta_0)}$ for a given value of the parameter θ, is available from the training stage (e.g., see Fig. 7 for an illustration of the elementwise D3F in the case of a shift-in-mean problem with Laplace and Gaussian observations). For instance, the elementwise D3F $t_{\omega_\theta}(x, \theta)$ can be derived as for the simple hypotheses, introduced in Section III, assuming that a specific value of θ is in force under $\mathcal{H}_1$. This means that the training stage will be repeated a number of times equal to the cardinality of Θ, and there are as many training sets, one for each specific value of the parameter $\theta \in \Theta$.

Assuming a structure like that in (38), the D3F counterpart is the following:[20]

$$T_{\boldsymbol{\omega}}^{(n)} = \frac{1}{n} \log \sum_{\theta \in \Theta} w_\theta \prod_{i=1}^{n} \exp\left(t_{\boldsymbol{\omega}_\theta}(x_i, \theta)\right), \qquad (40)$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_\theta)_{\theta \in \Theta}$; the D3F structure is illustrated in Fig. 5. In (40), the prior w_θ has to be estimated from the training set, or assumed uniform; in both the cases, the prior w_θ is not relevant from an asymptotic point of view.

Given that the D3F is constructed to have the same structure as the LLR, we can interrogate whether the scaled LMGF convergence holds similar to (39). In this respect, let us consider that the LLR can be written as in (37), where the true parameter is θ_* and the data are generated according to $f(\cdot|\theta_*)$ under $\mathcal{H}_1$. The key aspect to consider is that R_n, defined in (37), vanishes for n large enough (see the derivation in Appendix A). Under $\mathcal{H}_1$ the terms at the exponent in R_n converge to the KL divergence between θ_* and θ when θ_* is the true parameter; then, R_n vanishes exponentially fast with n. With a similar argument we can compute the term R_n for the D3F, that is the following

$$\sum_{\theta \in \Theta, \theta \neq \theta_*} \frac{w_\theta}{w_{\theta_*}} \exp\left[-n\left(\frac{1}{n}\sum_{i=1}^{n} t_{\omega_{\theta_*}}(x_i, \theta_*) - t_{\omega_\theta}(x_i, \theta)\right)\right]. \qquad (41)$$

Given that the observations x_i are IID we have that both the sample means at the exponent in (41) converge a.s. to their

[20] From an implementation perspective, both the decision statistics (38) and (40) need to be computed with the well-known log-sum-exp trick [96, Sec. 3.5.3] to avoid numerical underflow.

expected values:

$$\frac{1}{n}\sum_{i=1}^{n} t_{\omega_\theta}(x_i,\theta) \xrightarrow{a.s.} \mathbb{E}\left[t_{\omega_\theta}(x,\theta)|\theta_*\right] = \mu(\theta|\theta_*). \tag{42}$$

If the expected value of the D3F under θ_* is larger than the expected values in $\theta \neq \theta_*$, namely $\mu(\theta_*|\theta_*) > \mu(\theta|\theta_*)$, then R_n vanishes to zero and the convergence is similar to that in (39). It is worthwhile to note that such a condition is not necessary to have a large deviations result for the D3F, but it is useful to evaluate the asymptotic scaled LMGF, and consequently the test performance.[21]

Remark: Another valid learning strategy could be to train the D3F to distinguish not only any given parameter θ from θ_0, but also to distinguish $\theta \in \Theta$ from any other value $\widetilde{\theta} \in \Theta$, with $\theta \neq \widetilde{\theta}$. A possible training strategy would be resorting to the multi-class cross-entropy cost function; see, e.g., [81]. This approach leads to the M-ary statistical hypothesis testing counterpart. In this context, the error exponent analysis of the LLR is available in [97].

C. DEPENDENT OBSERVATIONS: TARGET DETECTION IN BINARY IMAGES

In this subsection we describe a statistical decision problem, where the observations $\mathcal{X}^{(n)} = (x_j)_{j=1}^{N_c}$ are dependent, and such dependence cannot be easily modeled. We consider the problem of deciding if a target is present or absent in the image $\mathcal{X}^{(n)}$, where $n \leq N_c$ represents the size of the target when present, and N_c is the size of the image. It is worthwhile to note that, different from the previous scenarios, the size of the data $\mathcal{X}^{(n)}$ is N_c and not n.[22] The data x_i, $i = 1, \ldots, N_c$, are represented by the pixels of the image (or, in the radar/sonar context, resolution cells), where the index i represents the i-th pixel. We indicate with $\mathcal{H}_1$ the target-present hypothesis and with $\mathcal{H}_0$ the target-absent one. The target can be *extended* (i.e, $n > 1$), in the sense that it can occupy multiple cells; see, e.g., [73], [98] in the context of Extended Target Tracking (ETT), and [99], [100], [101] in the context of underwater object classification.

To fix ideas, we provide here a statistical model for the problem of deciding the presence/absence of an extended target, and we stress that this is an example and not intended to delimit a general use case. Let us assume that, if the target is present in the i-th pixel of the image, $\theta_i = 1$, otherwise $\theta_i = 0$. We assume that each pixel represents a binary observation, $x_i \in \{0, 1\}$ according to a Bernoulli distribution with success probability p_1 if the target occupies the i-th pixel, i.e., $\theta_i = 1$, or with success probability p_0 if $\theta_i = 0$. We have basically assumed that the image is the output of a preliminary detection stage performed at the pixel level, which is typical in radar/sonar processing. The probabilities p_1 and p_0 are respectively the pixel-wise detection probability and false alarm probability. The classic notion of SNR would be involved in the relationship between p_1 and p_0. Clearly, another option would be to formalize the pixel observation with a continuous distribution, assuming a suitable clutter and target modelling.

Let us indicate $\Theta^{(n)} = (\theta_i)_{i=1}^{N_c}$, where n is given by the size of the target, namely $n = \sum_{i=1}^{N_c} \theta_i$. Obviously, under the $\mathcal{H}_1$ hypothesis we have $n \geq 1$. Let us also assume for simplicity that under $\mathcal{H}_1$ the data are conditionally independent given the target position and shape $\Theta^{(n)}$. Otherwise stated, the distribution of the data is given by

$$f_1\left(\mathcal{X}^{(n)}\,|\Theta^{(n)}\right) = \prod_{i,\theta_i=1} p_1^{x_i}(1-p_1)^{1-x_i} \times \prod_{i,\theta_i=0} p_0^{x_i}(1-p_0)^{1-x_i}. \tag{43}$$

Analogously, the distribution of the data under $\mathcal{H}_0$ is given by

$$f_0\left(\mathcal{X}^{(n)}\right) = \prod_{i} p_0^{x_i}(1-p_0)^{1-x_i}. \tag{44}$$

Note that the $\mathcal{H}_0$ hypothesis is simple, i.e., independent on n because of the absence of the target.[23]

If we assume perfect knowledge about the target position and shape, it is easy to verify that the LLR rule is equivalent to a test between two sequences of Bernoulli random variables of length n. In this scenario, the aim is to generalize the detection to the case of *unknown* target position and shape. In principle, one could always implement the LLR as in (38), obtaining

$$L^{(n)} = \frac{1}{n}\log\frac{\sum_{\Theta^{(n)}\in\mathcal{T}^{(n)}} w_{\Theta^{(n)}} f_1\left(\mathcal{X}^{(n)}\,|\Theta^{(n)}\right)}{f_0\left(\mathcal{X}^{(n)}\right)}, \tag{45}$$

where $\mathcal{T}^{(n)}$ and $w_{\Theta^{(n)}}$ are respectively the space and the (discrete) distribution of all target positions and shapes under consideration. Modeling such space and distribution is clearly infeasible in a real-world application, especially if we consider that, in principle, the target can have any possible shape. A convenient choice for the tractability of the problem that is usually made in the ETT literature is to model the target as an ellipsoid, with the target shape posterior in some cases given by an inverse Wishart distribution (e.g., see [73], [98], [103] and references therein). In an ETT problem, the relevant parameters of the target, such as its position, velocity, and shape are estimated sequentially. Although the ellipsoid assumption is quite convenient, it is also somehow limiting. Different from the ETT literature, we focus primarily on the target detection task, where the D3F decision statistic is based on a DCNN, whose architecture is represented in Fig. 6. The

[21] Numerical simulations in the case of conditionally independent observations are reported in Section VII-B, Figs. 10 and 12.

[22] The parameter n is not anymore the size of the data, but it has the same role as before; it is just quantified in a different way. The key idea is that n still rules the detection performance, but the rate function changes to take into account the redundancy introduced by the dependence.

[23] The model (43)-(44) can be further generalized assuming non-uniform detection and false alarm probabilities; see, e.g., [102].

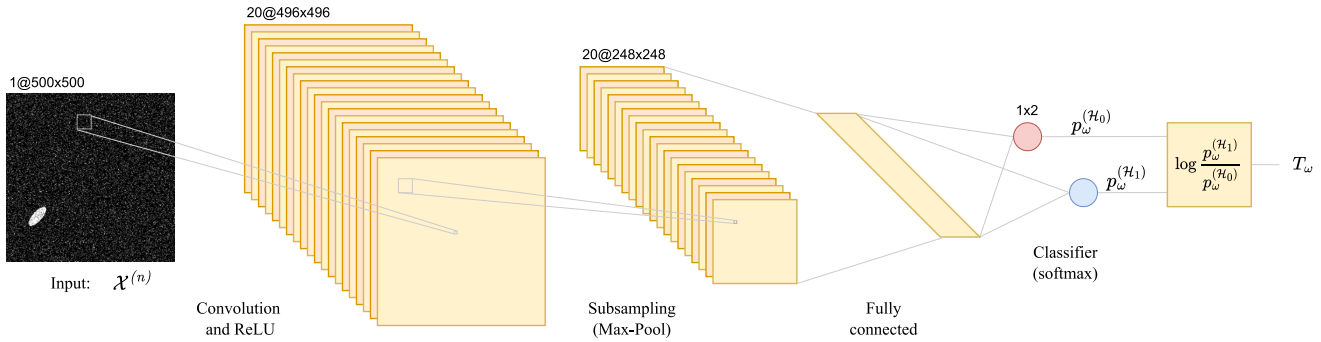


FIGURE 6. **Illustration of the employed CNN model. The architecture comprises three layers: one convolutional layer, one pooling layer, and one fully-connected classification layer. The convolutional layer uses a 5 × 5 convolutional kernel and has 20 neurons; the pooling layer uses a 2 × 2 pooling kernel. The last layer is a fully-connected classification layer with softmax activation function, whose output is the final $\mathcal{H}_0$ vs $\mathcal{H}_1$ classification.**

DCNN is a proper choice to capitalize the dependency structure embedded in the input data, which can be learned during the training.

We generate the synthetic training sets $\mathcal{Y}_k$ under $\mathcal{H}_k$, $k = 0, 1$, each composed by m_y images. Under $\mathcal{H}_1$, the images in $\mathcal{Y}_k$ are generated according to (43)–(44), with varying target positions, shapes and orientations; thus, the size of the target n_j varies for $j = 1, \ldots, m_y$. Under $\mathcal{H}_0$, only false alarms are generated randomly. The training set does not include any explicit information about the actual size, shape and position of the target; therefore the DCNN-based D3F does not process directly n or $\Theta^{(n)}$.

The D3F decision statistic is defined as

$$T_{\boldsymbol{\omega}}\left(\mathcal{X}^{(n)}\right) = \log \frac{p_{\boldsymbol{\omega}}^{(\mathcal{H}_1)}(\mathcal{X}^{(n)})}{p_{\boldsymbol{\omega}}^{(\mathcal{H}_0)}(\mathcal{X}^{(n)})}, \tag{46}$$

and is the log ratio of the softmax classifier probabilities computed by the DCNN, i.e., $p_{\boldsymbol{\omega}}^{(\mathcal{H}_k)}(\mathcal{X}^{(n)})$ of hypothesis $\mathcal{H}_k$, $k = 0, 1$, with $p_{\boldsymbol{\omega}}^{(\mathcal{H}_0)}(\mathcal{X}^{(n)}) + p_{\boldsymbol{\omega}}^{(\mathcal{H}_1)}(\mathcal{X}^{(n)}) = 1$. The $\boldsymbol{\omega}$ vector represents the network parameters learned in the training phase. The D3F $T_{\boldsymbol{\omega}}$ is the natural generalization of (12) for IID observations, and plays the role of the LLR (45).[24]

1) ASYMPTOTIC FRAMEWORK

The model (43)-(44) assumes that the size of the input data is larger than the target, i.e., $n \leq N_c$. Patently, the limit for n that diverges demands that an infinite input sequence is available. In this respect, it is implicitly assumed that the DCNN can process increasingly large images ($N_c \to \infty$). From a practical perspective, this amounts to assuming that the number of pixels in the image N_c is always sufficiently larger than the size of the target n, for any n of interest.

[24]Numerical simulations for the case of extended target detection are reported in Section VII-C; the analysis of real-world data is reported in Section VII-D (see Figs. 13 and 15).

The distribution of the input data under $\mathcal{H}_0$, and under $\mathcal{H}_1$ conditioned on the target $\Theta^{(n)}$, remains a sequence of independent Bernoulli random variables. As in Section V-B, both the detection performance and the asymptotic scaled LMGF shall be conditioned to a realization of the parameter under $\mathcal{H}_1$. In the current scenario, to evaluate the performance, we shall fix a shape for the target, for instance a circle, and a growing rule when n increases. The growing rule could be that the radius r_Θ of the circle diverges, with n that therefore also diverges (quadratically) with the radius.

Then, we consider the sequence of random variables $T_{\boldsymbol{\omega}}^{(n)}$ given by the normalized D3F decision statistic

$$T_{\boldsymbol{\omega}}^{(n)} = \frac{1}{n} T_{\boldsymbol{\omega}}\left(\mathcal{X}^{(n)}\right), \tag{47}$$

where $n = \sum_{i \geq 1} \theta_i$ increases, and the randomness of the observations $\mathcal{X}^{(n)}$ is given by the Bernoulli realizations in (43)-(44). Having fixed the shape of the target, we can generate the characterization set for each value of n that is of interest, and then we can compute numerically the asymptotic scaled LMGF and the relevant parameters of the exact asymptotics to construct the approximate error probability curves (2). More details on this are given in Section VI.

D. ASYMPTOTIC NORMALITY AND THRESHOLD SETTING

Assuming that the limit (34) exists, the decision statistic is then asymptotically normal around the asymptotic expected value, given by $\mu_k = \varphi_k'(0)$; see (8) and the discussion in Section II. Specifically, similar to (18), we can define the normalized statistic under $\mathcal{H}_k$:

$$\begin{aligned} \widetilde{T}_k^{(n)} &= \sqrt{n}\,(T^{(n)} - \mu_{n,k}), \\ \mu_{n,k} &= \mathbb{E}\left[T^{(n)} | \mathcal{H}_k\right], \\ \sigma_{n,k} &= \mathbb{S}\left[T^{(n)} | \mathcal{H}_k\right], \end{aligned} \tag{48}$$

where we indicate with $T^{(n)}$ the LLR (38), or the D3F statistic, i.e., (40) or (47). For n sufficiently large, $\mu_{n,k} \to \mu_k$ and $\sqrt{n}\,\sigma_{n,k} \to \sigma_k = \sqrt{\varphi_k''(0)}$, and by virtue of the results in [68],

we have that[25]

$$\widetilde{T}_k^{(n)} \xrightarrow{d} \mathcal{N}(0, \sigma_k^2), \quad \text{under } \mathcal{H}_k,\ k = 0, 1. \tag{49}$$

As in (20), we can then set the asymptotic false alarm probability $\alpha = \lim_{n\to\infty} \alpha_n$ as follows

$$\gamma_n = \mu_{n,0} + \sigma_{n,0} Q^{-1}(\alpha), \tag{50}$$

where $\mu_{n,0}$ and $\sigma_{n,0}$ are given in (48).[26]

1) CONDITIONALLY INDEPENDENT OBSERVATIONS

To grasp further insights about the asymptotic normality of the statistic in (48), let us consider the scenario with conditionally independent observations, discussed in Section V-B. Let us begin with $T^{(n)}$ given by the LLR in (38). The convergence can be established considering that the term R_n in (37) vanishes in probability. Specifically, under $\mathcal{H}_1$ and with a derivation similar to that reported in Appendix A, we can show that the LLR is given by the clairvoyant LLR (i.e., that has knowledge of θ_*, which is the true value of θ) plus a remainder term, referred to as $\widetilde{R}_n$, i.e.,

$$L^{(n)} = L_{\theta_*}^{(n)} + \widetilde{R}_n, \tag{51}$$

with $\widetilde{R}_n = \frac{1}{n}\log(w_{\theta_*}(1 + R_n))$ that converges to zero in probability (given that R_n vanishes in probability; see details in Appendix A). By virtue of Slutsky's theorem [84], we have that $L^{(n)}$ converges in distribution to the limit distribution of the ideal LLR $L_{\theta_*}^{(n)} = \frac{1}{n}\log\prod_{i=1}^{n}\frac{f(x_i|\theta_*)}{f(x_i|\theta_0)}$, which is the LLR of IID observations distributed under $\mathcal{H}_1$ with parameter θ_*. Thanks to the CLT, the normalized LLR of IID observations converges in distribution to a Gaussian distribution, as already expressed in (19). Analogously, a similar convergence holds under $\mathcal{H}_0$; indeed, considering (68), we have that $L^{(n)}$ converges in distribution to the limit distribution of $L_{\theta_m}^{(n)} = \frac{1}{n}\log\prod_{i=1}^{n}\frac{f(x_i|\theta_m)}{f(x_i|\theta_0)}$, which is the LLR of IID observations distributed under $\mathcal{H}_0$ with the parameter θ_m that minimizes the KL divergence between $f(x|\theta)$ and $f(x|\theta_0)$ over $\theta \in \Theta$. Summarizing, we have that $\sqrt{n}(L^{(n)} - \mu_{n,k})$ converges in distribution to a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma_k^2)$ under $\mathcal{H}_k$, with the expected values given by the following KL divergences

$$\mu_{n,k} \to \mu_k = \begin{cases} \mathcal{D}(f(x|\theta_*)||f(x|\theta_0)) & k = 1 \\ -\mathcal{D}(f(x|\theta_0)||f(x|\theta_m)) & k = 0, \end{cases}$$

and $\sigma_k = \mathbb{S}\left[\log\frac{f(x|\theta)}{f(x|\theta_0)}|\mathcal{H}_k\right]$, with $\theta = \theta_*$ under $\mathcal{H}_1$ and $\theta = \theta_m$ under $\mathcal{H}_0$. Note that the convergence of $\mu_{n,k}$ is coherent with the convergence of the scaled LMGF (39); indeed, it is easy to verify that $\mu_k = \varphi_k'(0)$ from the properties of the asymptotic scaled LMGF [65]. For the convergence of the D3F (40), we can proceed analogously, assuming that the D3F R_n term vanishes in probability to zero. However, this latter condition is not strictly necessary to establish the aforementioned convergence, as reported in Section VII. In other words, the convergence of R_n to zero is just a sufficient condition for the Gaussian convergence of the D3F.

2) TARGET DETECTION IN BINARY IMAGES

In the problem described in Section V-C, the LLR is practically infeasible. However, it is worth mentioning that it would converge asymptotically in distribution to a Gaussian distribution because of the aforementioned conditional independence of observations.

Let us consider the sequence of decision statistic $T_{\boldsymbol{\omega}}^{(n)}$ defined in (47). The DCNN training set used in this work is composed by different target shapes, positions, orientations, and size n, whose knowledge is not capitalized during the learning phase. In other words, the DCNN (and the resulting D3F) is not dependent on the parameter n; this dependence is only induced through the input observations. Given that under $\mathcal{H}_0$ the observations contain only false alarms, then the normalized statistic (48) is independent on $\Theta^{(n)}$. However, we have observed empirically, as illustrated in Fig. 14 [where we plot the histograms of the normalized statistic (47)], that the D3F (46) is approximately Gaussian-distributed under $\mathcal{H}_0$; in other words:

$$T_{\boldsymbol{\omega}}\left(\mathcal{X}^{(n)}\right) \sim \mathcal{N}(\mu_0, \sigma_0^2), \tag{52}$$

where, as for the other scenarios, both the parameters μ_0 and σ_0 can be estimated from the characterization set. Thanks to (52), the threshold can be again set as in (50):

$$\gamma = \mu_0 + \sigma_0 Q^{-1}(\alpha), \tag{53}$$

where α is the desired false alarm rate. Assuming that (52) holds, the normalized statistic (47) clearly vanishes under $\mathcal{H}_0$ for n large enough, and therefore the LDP does not hold under $\mathcal{H}_0$. The LDP can be instead verified under $\mathcal{H}_1$. Indeed, we also observe empirically that the normalized decision statistic (47) converges to a Gaussian distribution, as illustrated in Fig. 14.

Remark: It is possible to justify that the D3F statistic (52) under $\mathcal{H}_0$ is approximately Gaussian assuming that the width of any of the layers of the network diverges; a similar argument can be found in [48], [50]. Specifically, at the first DCNN layer, the samples at the input of the different non-linearities (i.e., at the output of the convolutional layers) can be modeled, under some mild assumptions, as Gaussian random variables. This is thanks to the CLT for independent (but non-identically distributed) samples [84] and the IID assumption of $\mathcal{X}^{(n)}$ under $\mathcal{H}_0$. Then, we conjecture that this mechanism propagates among the layers. Specifically, we suppose that the CLT can be still exploited to model the samples at the input of the $(i+1)^{\text{th}}$ non-linearity, if the i-th layer's width is sufficiently large and assuming that the correlations among

[25] Note that in the problem of extended target detection in previous section, the statistic under $\mathcal{H}_0$ does not depend on $\Theta^{(n)}$, and thus n.

[26] The histograms of the decision statistic under both hypotheses and for different values of n are reported in Fig. 11 for the case of conditionally independent observations and in Fig. 14 for the case of extended target detection. In both the figures the histograms appear quite close to the normal distributions, even for small values of n.

the i-th layer's output samples are weak enough to invoke the CLT.

It is worthwhile observing that the assumptions made in [48], [50] require the network weights being IID; conversely, here the weights are fixed and the network inputs are random. A further analysis of this convergence is left to future works; we focus instead here on the convergence in the parameter n.

VI. RATE FUNCTION ESTIMATION AND ERROR PROBABILITY APPROXIMATIONS

In the previous section we have established the main large deviations properties of the D3F test (1). In this section we detail a viable means to compute rate functions, approximate error probability curves, and evaluate the threshold parameters, exploiting the *characterization set* $\mathcal{Z}$. In some circumstances an overlap between the characterization set and the training set is possible. For this reason, in Appendix B we show that the estimators based on the sample mean, which will be detailed in the next subsections, converge in probability to the proper quantities of interest. In other words, the independence between characterization set and training set is not strictly necessary.

Before entering into details, let us consider that the characterization set depends on the specific parameter $\theta \in \Theta$ under consideration when the observations are conditionally independent (see the definition in Section V-B), as well as on the extended target $\Theta^{(n)}$ under consideration, defined in Section V-C.

A. RATE FUNCTION ESTIMATION

We propose two procedures to estimate the rate function. The former is based on the *direct estimation* of the LMGF, or its scaled version, and relies on solving numerically the Fenchel-Legendre transform thanks to a numerical representation of the (scaled) LMGF. The latter, which is more accurate, instead requires to sample from the exponentially tilted distribution via MCMs.

1) DIRECT ESTIMATION APPROACH

A statistically *consistent* estimator of the LMGF, i.e., $\log \mathbb{E}[\exp(t\,\tau(x))|\mathcal{H}_k]$, for the first scenario (Section III) can be obtained by replacing the expected value with the sample mean:

$$\hat{\varphi}_k(t) = \log \frac{1}{m_z} \sum_{j=1}^{m_z} \exp(t\,\tau_{j,k}), \tag{54}$$

where $\tau_{j,k} = t_{\boldsymbol{\omega}}(z_{j,k})$ is the elementwise D3F; the samples $z_{j,k}$ are assumed IID according to $\mathcal{H}_k$, $k = 0, 1$, and taken from the characterization set $\mathcal{Z}$, whose size is m_z for each hypothesis. Thanks to the law of large numbers, $\hat{\varphi}_k(t) \xrightarrow{a.s.} \varphi_k(t)$ for $m_z \to \infty$.[27] Then, it is possible to compute numerically the Fenchel-Legendre transform of the estimated LMGF $\hat{\varphi}_k(t)$, which leads to an estimate of the rate function. The same approach can be adopted when we deal with the scaled LMGF in (34), for different values of n:

$$\hat{\varphi}_{n,k}(t) = \frac{1}{n} \log \frac{1}{m_z} \sum_{j=1}^{m_z} \exp(n\,t\,T_{j,k}^{(n)}), \tag{55}$$

where $T_{j,k}^{(n)}$ is computed exploiting the characterization set, and we have again replaced the expected value with the sample mean and implicitly assumed the dependency on the parameter θ or $\Theta^{(n)}$. Clearly, a sufficiently large value of m_z is required to approximate the scaled LMGF; moreover it is also required that n is large enough to capture the asymptotic behaviour of interest. For each value of n, the approximate scaled LMGF can be exploited to compute the saddlepoint, as in (32), but with a rate function I_n given by the Fenchel-Legendre transform of the estimated version of the scaled LMGF for a given n. This approach is valid in view of the non-asymptotic regime; see, e.g., the discussion in [90]. We apply this strategy to compute the approximate error probability curves for the problem of deciding the presence/absence of an extended target in an image, discussed in Section V-C.

2) RATE FUNCTION ESTIMATION VIA THE EXPONENTIAL CHANGE OF MEASURE, AND SAMPLING FROM THE TILTED DISTRIBUTION

The direct estimation of the scaled LMGF can be inaccurate when m_z is not large enough to accommodate possibly high values of t and/or n. A valid alternative is use an importance sampling procedure to generate samples from the *exponentially tilted* distribution; such procedure is also referred to as exponential change of measure [66], [104], [105]. The approach is general enough to be applied to a generic sequence of data $\mathcal{X}^{(n)}$.

The main idea is to estimate directly the derivative of the (asymptotic) scaled LMGF, based on which the rate function estimation can be computed. To this end, let us observe that

$$\lim_{n\to\infty} \mathbb{E}_{f_{n,k}^{(t)}}\left[T^{(n)}(\mathcal{X}^{(n)})\right] = \varphi_k'(t), \tag{56}$$

provided that the input $\mathcal{X}^{(n)}$ is sampled from the tilted distribution, indicated with $f_{n,k}^{(t)}(\mathcal{X}^{(n)})$ under $\mathcal{H}_k$. The tilted distribution of $\mathcal{X}^{(n)}$ under $\mathcal{H}_k$ is defined as follows

$$f_{n,k}^{(t)}(\mathcal{X}^{(n)}) = \frac{\exp(n\,t\,T^{(n)}(\mathcal{X}^{(n)}))}{W_n^{(t)}}\, f_k(\mathcal{X}^{(n)}), \tag{57}$$

where $W_n^{(t)}$ is the distribution normalizing factor. The derivation is straightforward and is reported, e.g., in [66]. We

[27] If the characterization set is overlapped with the training set, then the strong convergence should be replaced with the weak convergence, see details in Appendix B.

assume that $\mathcal{X}^{(n)}$ can be sampled from the tilted distribution with the MH algorithm[28] or with an alternative strategy.

As a consequence, a viable estimator of $\varphi'_k(t)$ can be obtained by the sample mean of $T^{(n)}(\mathcal{X}_j^{(n)})$, where data $\mathcal{X}_j^{(n)}$, $j = 1, 2 \ldots, m'_z$, are drawn from the tilted distribution (57):

$$\hat{\varphi}'_{n,k}(t) = \frac{1}{m'_z} \sum_{j=1}^{m'_z} T^{(n)}(\mathcal{X}_j^{(n)}), \quad \mathcal{X}_j^{(n)} \sim f_{n,k}^{(t)}(\mathcal{X}^{(n)}). \tag{58}$$

The rate function is given by the Fenchel-Legendre transform, which can be efficiently solved numerically exploiting the estimated derivative of the scaled LMGF (58) thanks to the duality property discussed in Section V-A; see also (36). Specifically, the approximate point that solves the Fenchel-Legendre transform is given by the equation $\gamma = \hat{\varphi}'_{k,n}(t_\gamma)$. Then, based on (36), the rate function can be computed by integrating numerically the derivative of the scaled LMGF with the boundary condition that the scaled LMGF is null in $t = 0$.

Finally, it is worthwhile stressing that the number of data m'_z required by the MH-based method to achieve a reliable estimate of the rate function (which implicitly requires a large value of n) is significantly smaller compared to the direct estimation method (54)-(55); see details, e.g., in [66].

Remark: The main drawback of the method based on the exponential change of measure is that we need to sample from the tilted distribution. The MH-based method provides a viable means to accomplish this task, however it demands some knowledge of the input data distribution to generate samples used in the MH algorithm as well as to compute the so-called acceptance ratio. The possibility of sampling data and computing the acceptance ratio is perfectly acceptable in several scenarios, for instance in the context of GANs and VAE. Furthermore, as already discussed in Section II, a partial knowledge of the statistical model of the input data is one of the distinguishing features of model-based ML approaches, see, e.g., [76], [77].

B. ERROR PROBABILITY APPROXIMATIONS

The error probability approximations introduced in Section II and reported in Table 1 are based on the small and large deviations principles. The small deviations approximation relies on the Gaussian convergence of the decision statistic, see (19) and (49), and only requires the computation of the mean $\mu_{n,k}$ and the variance $\sigma^2_{n,k}$ of the decision statistic under each hypotheses, as it will be detailed in the next subsection.

The large deviations approximation is provided by the exact asymptotics, or the saddlepoint approximation, introduced in Section IV-D, for a given value of n. Specifically, this approximate expression of the error probability follows the structure given in (32), which is valid not only for IID observations, but also in more general cases [62]. To compute (9)-(10) we basically need to apply the approximation (32) to the error probabilities $\alpha_n = \mathbb{P}[T^{(n)} \geq \gamma_n | \mathcal{H}_0]$ and $\beta_n = \mathbb{P}[-T^{(n)} > -\gamma_n | \mathcal{H}_1]$, where $\gamma_n \to \gamma$ and γ is the asymptotic threshold, as detailed in Section IV-D. Thus, we need the following ingredients:

1) The saddlepoint $t_{\gamma,k}$ corresponding to the solution $\varphi'_{n,k}((-1)^k t_{\gamma,k}) = \gamma$ under $\mathcal{H}_k$, $k = 0, 1$;
2) The value of the rate function $I_k(\gamma)$, under $\mathcal{H}_k$, $k = 0, 1$;
3) The second derivative of the (scaled) LMGF in the saddlepoint, $\varphi''_{n,k}((-1)^k t_{\gamma,k})$ under $\mathcal{H}_k$, $k = 0, 1$.

The first two ingredients are automatically available from the estimation of the rate function, and the third one is easily obtained from the (scaled) LMGF (details omitted for brevity).

Special attention should be paid when the threshold moves with n, as the approximation can be inaccurate, especially for small values of n, as it will be discussed in Section VII. In these cases, we can resort to the "non-asymptotic" saddlepoint approximation [90], which basically consists in computing the previous three ingredients for each value of γ_n, instead of γ. This leads to compute the (scaled) LMGF (55), the saddlepoint $t_{\gamma_n,k}$, and an approximate rate $I_n(\gamma_n)$, all varying with n.

With this approach, i.e., leveraging the properties of the scaled LMGF for a given value of n, the approximate rate is not guaranteed to be strictly positive as the intervals $[\gamma_n, \infty)$ for α_n and $(-\infty, \gamma_n)$ for β_n may include the related expected value of the decision statistic, i.e., $\mu_{n,k}$. In such cases, we can exploit the small deviations approximation instead of the exact asymptotics. Clearly, the aforementioned Gaussian approximation is not expected to perform well when n diverges, as already pointed out in the previous sections, and as it will be verified in Section VII. However, under the assumption that $\mu_0 < \gamma < \mu_1$, it is expected that, for large enough values of n, the aforementioned critical situations do not occur.

C. ESTIMATION OF $\mu_{n,k}$ AND $\sigma_{n,k}$ TO SET THE TEST THRESHOLD

The threshold γ_n of the test in (1) can be selected following two strategies:

1) By fixing one of the error probabilities to a desired level, for instance the false alarm, see (20) and (50), where γ_n converges to μ_0; or
2) In such a way that both error probabilities vanish, with $\gamma_n = \gamma \in (\mu_0, \mu_1)$.

In the first strategy it is necessary to estimate $\mu_{n,k}$ and $\sigma_{n,k}$. They can be computed straightforwardly, as they represent the expected value and the standard deviation of the decision statistic under $\mathcal{H}_k$, respectively. Both $\mu_{n,k}$ and $\sigma_{n,k}$ can be estimated via the sample mean exploiting the characterization set. Specifically, given that $\mu_{n,k}$ is the expected value of the

[28] Details about the well-known MH algorithm implementation are omitted for brevity, however relevant information can be found, e.g., in [106] and about its application in the LDP context in [66].

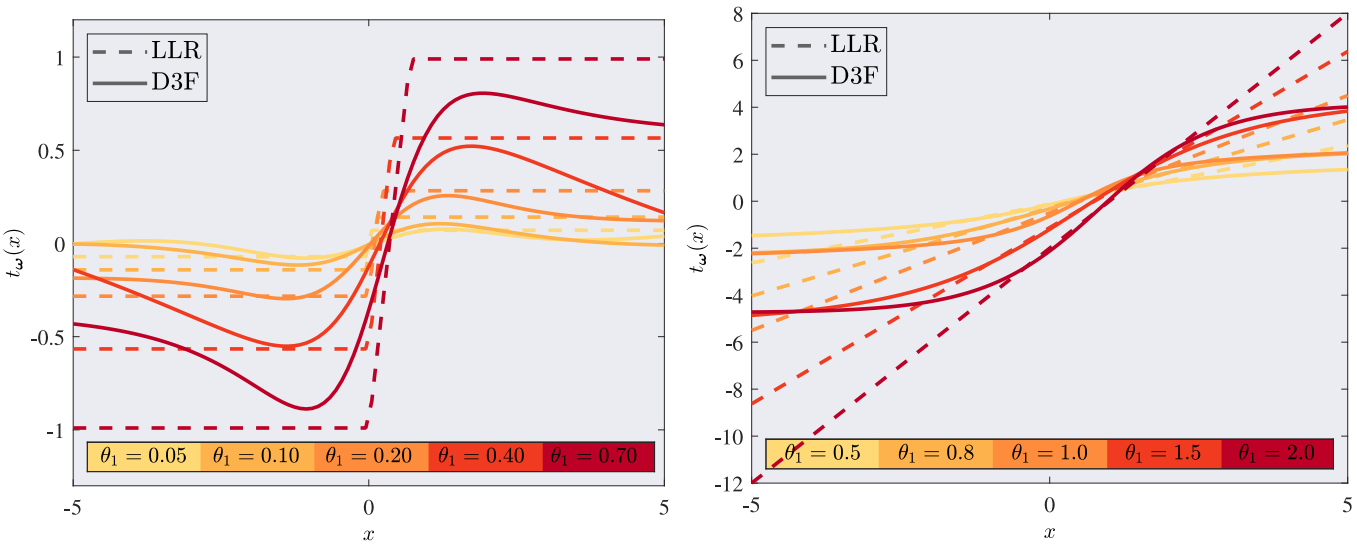


FIGURE 7. Elementwise D3F (solid curve) versus the input x. The elementwise D3F is defined as 2 layers NN trained with $m_y = 10^4$ independent samples under each hypothesis, distributed according to a Laplace (left) and Gaussian (right) with $\theta_0 = 0$ under $\mathcal{H}_0$ and different values of θ_1 under $\mathcal{H}_1$. The LLR function (dashed curve) is also reported for the different values of θ_1. Each color represents a specific value of θ_1.

decision statistic, the sample mean estimator is the following

$$\hat{\mu}_{n,k} = \frac{1}{m_z} \sum_{j=1}^{m_z} T_{j,k}^{(n)}, \tag{59}$$

where $T_{j,k}^{(n)}$ is provided by the characterization set. We can estimate $\sigma_{n,k}$ analogously. In the second strategy, we only need the asymptotic mean values μ_k, which can be estimated again with the sample mean, or as the derivative of the scaled LMGF, as reported in (8). Alternatively, from the rate functions we can compute the asymptotic means which are the points where the rates are nulls, i.e., $I_k(\mu_k) = 0$.

VII. EXPERIMENTAL RESULTS AND NUMERICAL SIMULATIONS

In this section, we provide a numerical analysis and the sanity check of the theoretical results previously stated. Before entering into details, we remind the distinction between the number of samples m_y under each hypothesis, namely the size of the training set, and the number of samples m_z of the characterization set, used to compute the means $\mu_{n,k}$ and standard deviations $\sigma_{n,k}$, as well as the rate functions and the approximate error probability curves. In the following simulation campaign we have chosen the size of the characterization set always equal to the number of Monte Carlo runs used to compute the empirical error probabilities. For the MH algorithm, we used the parameter $n = n_{MH}$, with $n_{MH} = 10^4$, to sample from the tilted distribution (57); this is done to compute the numerical version of (56). Then, in the exact asymptotic formula (32), the variable $n \geq 1$ is let varying to describe the error probability curve.

A. IID OBSERVATIONS

We start by analysing the case of IID observations presented in Secs. III and IV. Specifically, we consider a shift-in-mean detection problem with noise distributed according to a Laplace distribution. Let us denote by $\mathcal{L}(a, b)$ a (shifted) Laplace distribution with shift parameter a and scale parameter b, i.e., having the following probability density function

$$f_{\mathcal{L}}(x) = \frac{1}{2b} e^{-\frac{|x-a|}{b}}. \tag{60}$$

The goal is to test $x_i \sim \mathcal{L}(\theta_0, \sigma)$ under $\mathcal{H}_0$ versus $x_i \sim \mathcal{L}(\theta_1, \sigma)$ under $\mathcal{H}_1$. The elementwise D3F is trained with m_y independent observations under both $\mathcal{H}_0$ and $\mathcal{H}_1$. As loss function, the binary cross-entropy loss (17) is adopted. The D3F is defined as a simple fully connected NN with two hidden layers of 5 neurons each. The first hidden layer has a tanh activation function, and the second is followed by a softmax activation function; the elementwise D3F $t_{\boldsymbol{\omega}}(x)$ is given by the log ratio of the two softmax outputs; see (15). Fig. 7 illustrates the shape of the D3F $t_\omega(x)$ trained with data drawn from a Laplace (left panel) and Gaussian distribution (right panel), $m_y = 10^4$, for different values of θ_1 with $\theta_0 = 0$ and $\sigma = 1$, as well as the shape of the LLR computed for the same parameters. The scenario with Gaussian data is described in the next subsection. It is apparent that the D3F and the LLR functions are similar, but different for extreme values of x (i.e., $|x| \gg 0$). However, we will see later that these differences are not informative in terms of detection performance.

In Fig. 8 we analyse the scenario with $\theta_0 = 0$, $\theta_1 = 0.05$ and $\sigma = 1$. We show both the rate functions: $I_0(\gamma)$ under $\mathcal{H}_0$ and $I_1(\gamma)$ under $\mathcal{H}_1$ for the LLR (dashed line) and D3F (solid line) and two different values of m_y (represented with dark and light blue). The rates are computed by letting the threshold γ vary and are evaluated with the direct method described in

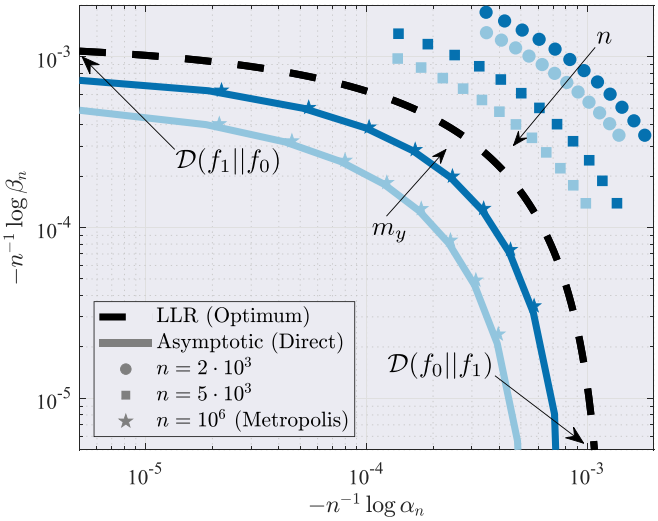


FIGURE 8. Rate function $I_1(\gamma)$ (y-axis) versus $I_0(\gamma)$ (x-axis) of the LLR (dashed black), see Theorem 4.2, and the D3F (solid lines), see Theorem 4.3. Two different D3F statistics are considered with $m_y = 10^3$ (light blue) and $m_y = 10^4$ (dark blue). We report also their respective finite sample-size version $-n^{-1}\log\alpha_n$ and $-n^{-1}\log\beta_n$, where the error probabilities α_n and β_n (markers) are computed for different values of n via Monte Carlo simulations. The maximum achievable rate is the KL divergence, attained by the LLR, when the rate under the other hypothesis is null. Data are IID and generated according to a Laplace distribution with $\theta_0 = 0$, $\theta_1 = 0.05$, and $\sigma = 1$. Number of Monte Carlo runs is 10^5 for the empirical curves when $n = 2\cdot 10^3$ and $n = 5\cdot 10^3$, while the curve for $n = 10^6$ is obtained sampling from the tilted distribution via the MH algorithm.

Section VI. The convergence is also assessed numerically for different values of n (i.e., $2\cdot 10^3$ and $5\cdot 10^3$) by computing the error probabilities α_n and β_n (markers) empirically via Monte Carlo simulations. The empirical curves for $n = 10^6$ cannot be computed via Monte Carlo simulations being the error probability prohibitively small.[29] For this reason, we resort to the framework of tilted distribution sampling [65]. We can see that $-n^{-1}\log\alpha_n \to I_0(\gamma)$ and $-n^{-1}\log\beta_n \to I_1(\gamma)$, as established by (27)–(29).

In Fig. 8, as expected, being the optimal detector, the LLR achieves the best rates. However, we observe that, increasing m_y from 10^3 to 10^4, the D3F curve gets closer and closer to the curve of the LLR. The extreme values of the LLR rate, where $I_0 = 0$ or $I_1 = 0$, provide respectively the best achievable rate, which is equal to the KL divergence $\mathcal{D}(f_0||f_1)$ when $I_1 = 0$ and $\mathcal{D}(f_1||f_0)$ when $I_0 = 0$. We analyze the test performance of both the decision statistics in one of these extreme points in Fig. 9, where $I_0(0) = 0$, setting the threshold as in (20) and thus obtaining $\alpha_n \to \bar{\alpha}$. The desired values of false alarm are set to 0.25 and 0.05, and, in this case study, we set the parameter under $\mathcal{H}_1$ to $\theta_1 = 2$, while $\theta_0 = 0$.

In Fig. 9, plots on the left refer to the LLR, while those on the right refer to the D3F, and the markers represent the empirical probabilities computed via Monte Carlo simulations. The upper panels illustrate the behaviour of α_n, which tends to the desired values (dashed lines) even for small values of n. The lower panels show instead the behaviour of β_n for the two desired asymptotic false alarm levels and their approximations. These approximations rely on either the exact asymptotics (32) (solid line), properly refined as described in Section VI-B, or the Gaussian convergence (19) (dashed line). We shall note that for values of n lower than 10, the Gaussian approximation is accurate, especially when the false alarm is smaller, while the exact asymptotics in this example uniformly provide a good approximation. However, it becomes evident as n increases that the Gaussian approximations converge to zero faster than the exact asymptotic curves, which instead match much better the empirical probability values. This behaviour of the Gaussian approximation is perfectly aligned with the theory discussed in the previous sections, as the Gaussian approximation is quite accurate when the interval, representing the event of interest, is not too far away from the mean where the decision statistic sequence is converging; alternatively, when it is too far away, the theory of large deviations shall be used to achieve a better accuracy. It is interesting to highlight that in Fig. 9, even if the convergence rate is equal for the two different asymptotic false alarm probabilities, the sub-exponential terms intervene to separate the curves. Moreover, the curves exhibit the same rate, meaning that for large n the curves converge to zero in parallel with the same slope.

B. CONDITIONAL INDEPENDENT OBSERVATIONS: COMPOSITE HYPOTHESIS TESTING

Let us now move on with the composite hypothesis testing problem described in Section V-B. This time we consider Gaussian observations $x_i \sim \mathcal{N}(\theta, \sigma)$; the Laplace example is similar and not reported for brevity. We assume $\sigma = 1$ under both the hypotheses. Under $\mathcal{H}_0$, we set the parameter $\theta_0 = 0$, whereas under $\mathcal{H}_1$ it can take two different possible values in $\Theta = \{\theta_1, \theta_2\}$; specifically, we set $\theta_1 = 0.25$ and $\theta_2 = 0.35$. The two parameter values are a priori equally probable. The D3F is trained exactly as for the IID case, but the training procedure is repeated for each value of θ under $\mathcal{H}_1$, as described in Section V-B. The number of training samples is $m_y = 10^3$ under each hypothesis and for each value in Θ. The shape of the elementwise D3F is illustrated in the right-hand side plot of Fig. 7. In Fig. 10, we report the convergence of β_n to zero (left-side panel) for the two possible values θ_1, θ_2 of the true parameter θ_* under $\mathcal{H}_1$, with an asymptotic false alarm fixed to 0.25, and resorting to the threshold selection γ_n described in Section V-D. We can see that the false alarm α_n (right side, top panel) converges quickly to the desired value even for small values of n for both the LLR and the D3F. We compare the exact asymptotic approximations (both the direct computation and the exponential tilting sampling methods described in Section VI-B) with the Monte Carlo simulation of the error probability β_n of the D3F (10^4 runs) and the LLR (10^7 runs).[30]

[29] To give a rough idea, with $n = 10^6$ and a rate $I = 10^{-4}$, the error probability would be $\sim \exp(-10^2) \approx 3.7 \times 10^{-44}$.

[30] The number of Monte Carlo runs of the LLR is larger than the D3F thanks to a convenient manipulation of the conditional LLR in (38).

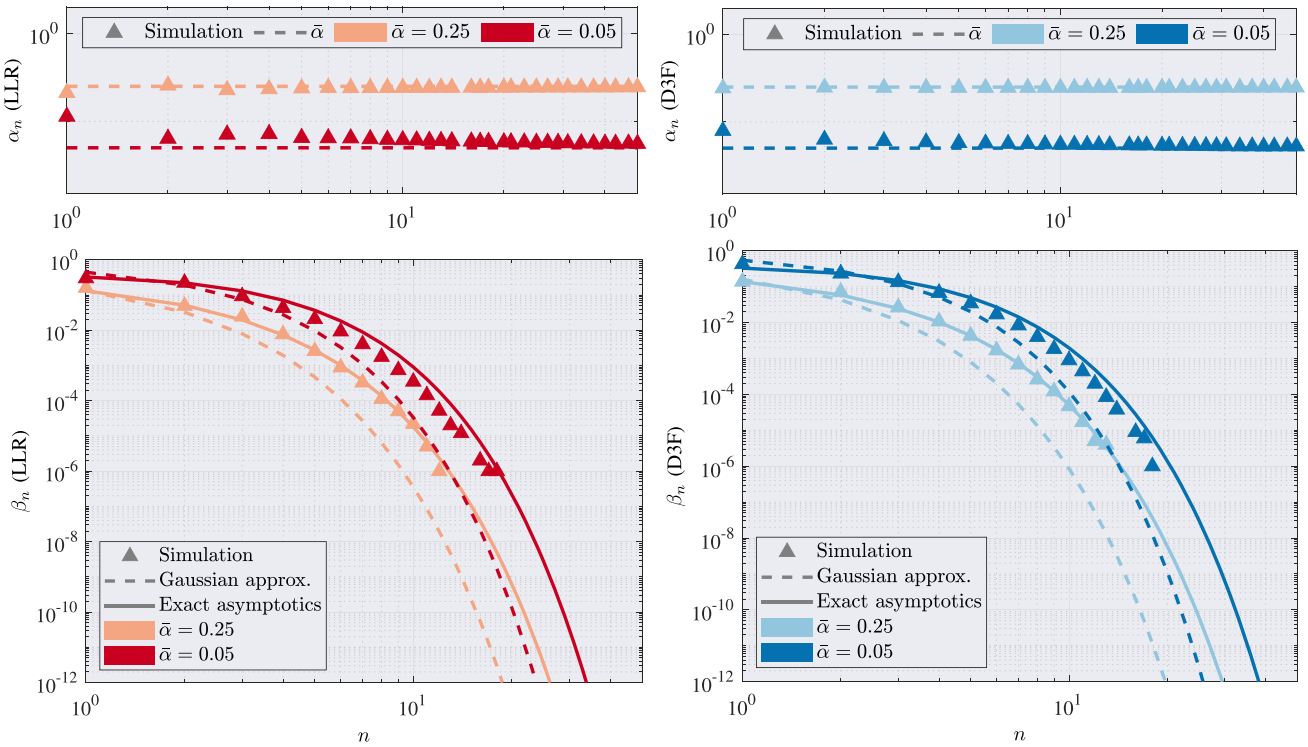


FIGURE. 9. Error probabilities α_n and β_n for IID Laplace observations. Plots in the left column refer to the LLR (shades of red), and those in the right column to the D3F (shades of blue). The plots in the top row illustrate the behaviour of α_n, which tends to the desired values $\bar{\alpha}$ equal to 0.25 (light red/blue) and 0.05 (dark red/blue). The plots in the bottom row show the behaviour of β_n (markers), computed via Monte Carlo simulations, and its two approximations: one based on the exact asymptotics (solid line), and the other one based on the Gaussian approximation (19) (dashed line), referred to small deviations in Section III-B. The parameters are $\theta_0 = 0$ and $\theta_1 = 2$, and the training set size is $m_y = 10^3$; the number of Monte Carlo runs is 10^6.

It is important to highlight that in this specific setting, different from the Laplace example, both the LLR and the D3F achieve the same performance in terms of error probabilities. This will be confirmed in the rate function analysis in Figure 12.

We also report the convergence of γ_n for both LLR and D3F (middle and bottom, respectively, right-side panels), and, as expected, $\gamma_n \to \gamma = \mu_{n,0}$. It is worthwhile to note that the agreement between the exact asymptotic approximations and the Monte Carlo simulations is less accurate when n is smaller than 10, and this is because the threshold γ_n is larger than $\mu_{n,1}(\theta)$. In the aforementioned condition, as explained in Section VI-B, the estimated rate is not guaranteed to be strictly positive; consequently, the saddlepoint approximation can be meaningless.

A pictorial analysis of the "small deviations" regime is reported in Fig. 11, which illustrates the convergence to the Gaussian distribution discussed in Section V-D. Left-column panels refer to the LLR, and right-column panels to the D3F. Each row refers to a different value of n, specifically 30, 207, 1439. The decision statistic distribution is simulated under $\mathcal{H}_0$ and $\mathcal{H}_1$ for each value of $\theta \in \Theta$, and we report in Fig. 11 the related histograms (colored area). Moreover, we also plot the approximated Gaussian distributions $\mathcal{N}(x; \mu_{n,k}, \sigma_{n,k})$ (solid lines) where $\mu_{n,k} \sim \mu_k$ and $\sigma_{n,k} \sim \sigma_k/\sqrt{n}$. There is almost a perfect match between the histograms and the Gaussian distributions for all the values around the means; however, as we already discussed, the agreement is expected to be increasingly less accurate on the tails.[31] We observe that the larger n, the more the distributions of the decision statistic concentrate around the asymptotic mean μ_k, with a decreasing variance, which confirms the convergence in (49). We also show how the threshold moves with n so that the asymptotic false alarm remains fixed at the desired level of 0.25, as in Fig. 10.

In Fig. 12, we set a fixed threshold $\gamma \in (\mu_0, \min_{\theta\in\Theta} \mu_1(\theta))$ in order to allow the convergence to zero of both the error probabilities (see left-hand side panels of Fig. 12). By doing so, trivial operating points are avoided. For this reason, we report in the right hand side panels of Fig. 12 the Fenchel-Legendre transform (computed by means of the exponential tilting method) of both the LLR (dark blue) and the D3F (light blue), under $\mathcal{H}_0$ and $\mathcal{H}_1$ for both values of θ. The LLR curves have a parabolic shape, in perfect agreement with (39). Indeed, the asymptotic scaled LMGF for conditionally independent Gaussian observations is given by the LMGF of the elementwise LLR, which is still Gaussian distributed; then, it is easy to compute the Fenchel-Legendre transform, which turns out to be a parabola (see, e.g., [55]). Interestingly, the D3F rate curves seem to be approximately an horizontal shift of the LLR ones. However, in order to have a meaningful

[31] The distributions' tails cannot be easily visualized in the histograms, but they are analyzed in terms of error probabilities in Figs. 10 and 12.

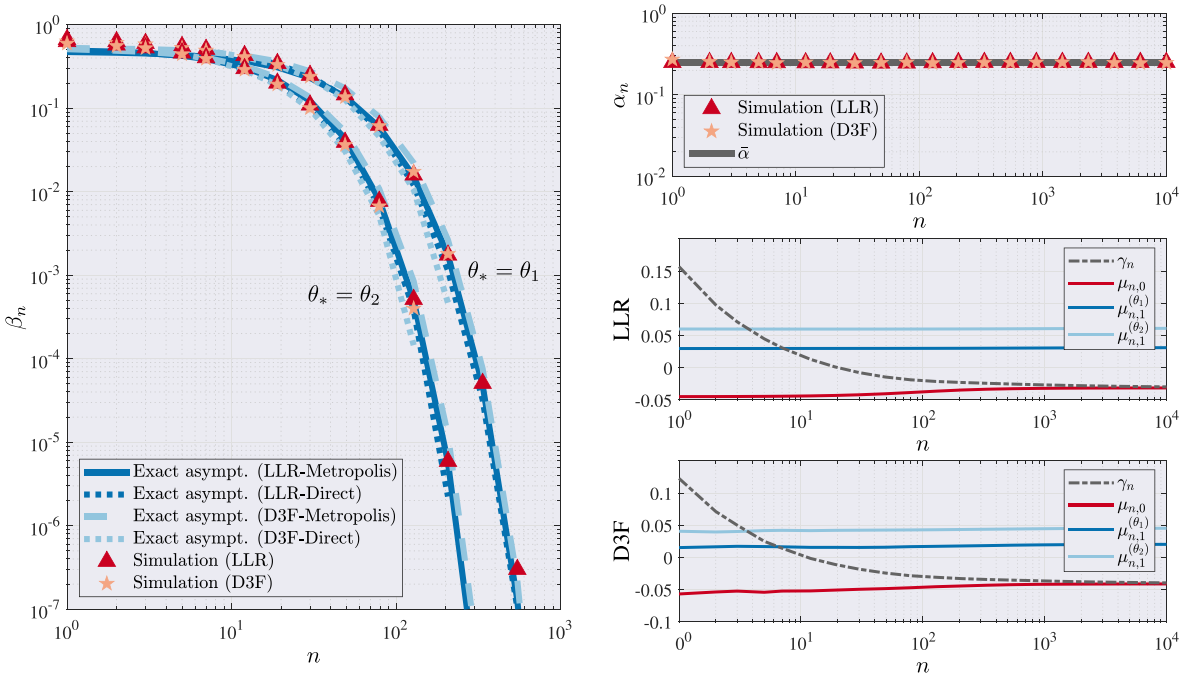


FIGURE 10. Composite hypothesis test with Gaussian observations. The D3F is trained with $m = 10^3$ samples. Leftmost panel: error probability β_n of the LLR and D3F; exact asymptotic approximation via exponential tilting sampling for the LLR (solid line) and the D3F (dotted line), Monte Carlo simulations of LLR (▲, 10^7 runs), and D3F (⋆, 10^4 runs) for each values of the true parameter θ_*. Rightmost upper panel: false alarm probability α_n, desired level $\bar{\alpha} = 0.25$ (solid line), LLR (▲) and D3F (⋆). Rightmost mid and bottom panels: threshold γ_n of LLR and D3F compared to the mean $\mu_{n,0}$ under $\mathcal{H}_0$ and the mean $\mu_{n,1}(\theta)$ under $\mathcal{H}_1$ where $\theta \in \Theta$.

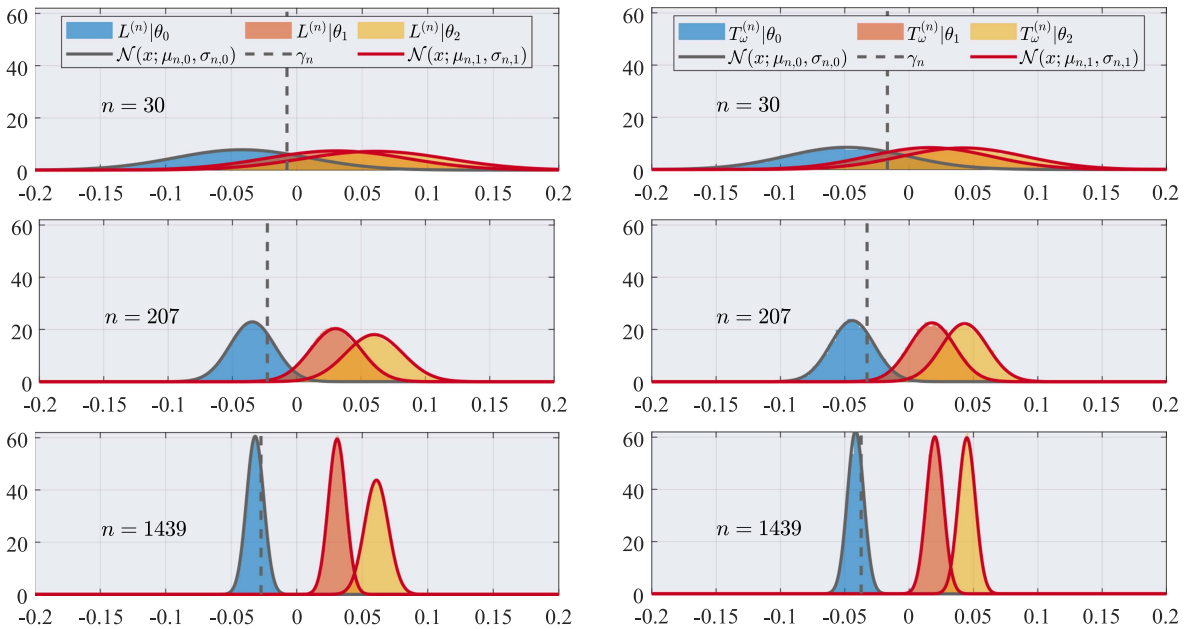


FIGURE 11. Empirical distribution of the decision statistics (left column: LLR, right column: D3F) in the case of composite Gaussian hypotheses, where $\theta_0 = 0$ under $\mathcal{H}_0$ (blue) and $\theta \in \{\theta_1, \theta_2\}$ under $\mathcal{H}_1$, with $\theta_1 = 0.25$ (orange) and $\theta_2 = 0.35$ (yellow). Gaussian distribution under $\mathcal{H}_0$ (solid gray line) and under $\mathcal{H}_1$ (solid red line), with parameters provided by the small deviations and estimated from the characterization set. Panels in each row refer to the same value of n. For the D3F case, the neural network was trained with $m = 10^3$ samples. The distributions of the LLR decision statistics have been estimated with 10^6 realizations, while those for the D3F statistics with 10^4 realizations. The threshold γ_n (dashed gray vertical line) moves with n so that α is fixed to 0.25.

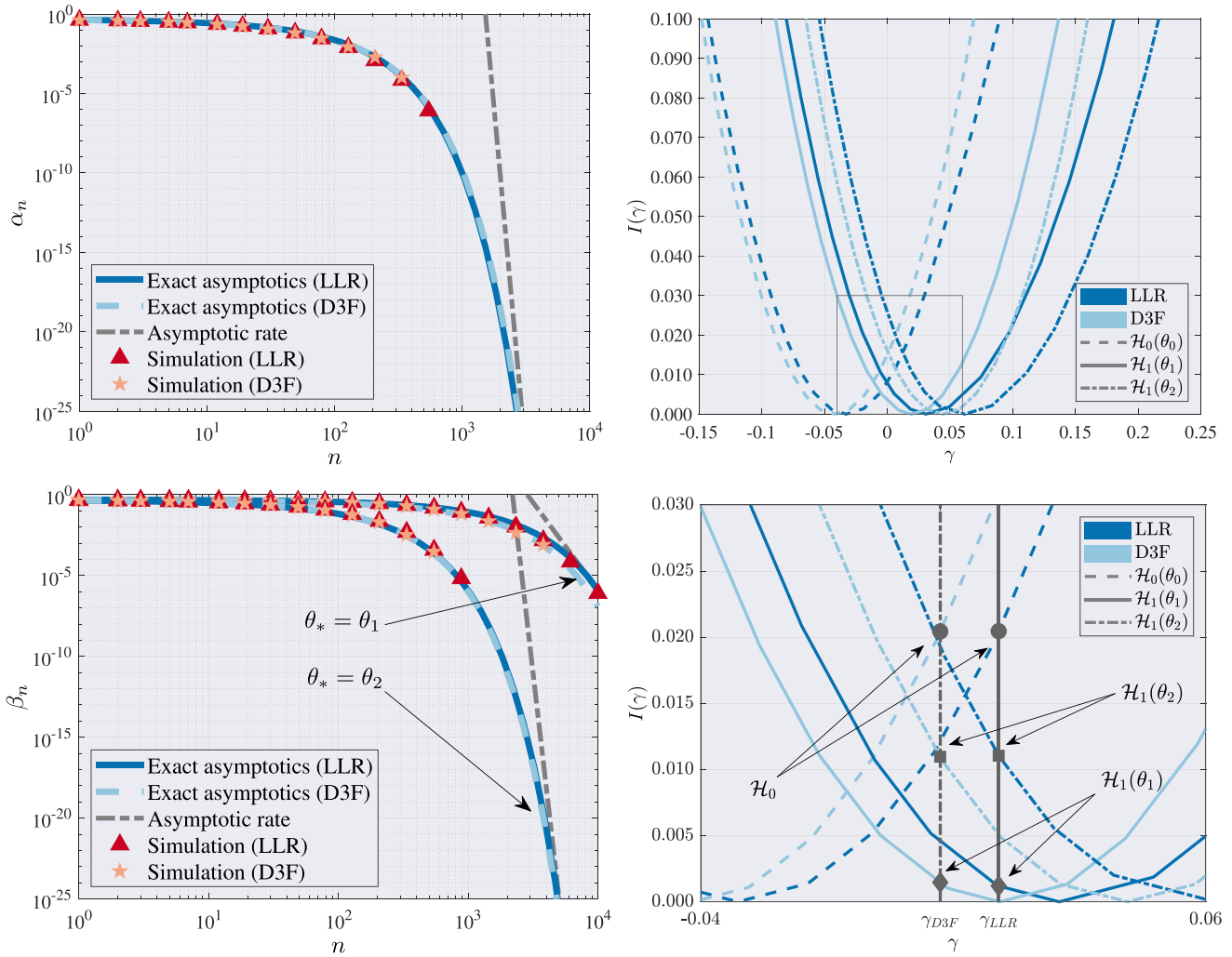


FIGURE 12. Composite hypothesis test with Gaussian observations. The leftmost plots show: the error probabilities α_n and β_n of the LLR and the D3F; the exact asymptotics computed via exponential tilting for the LLR (solid) and the D3F (dashed); the Monte Carlo simulations for the LLR (▲, 10^7 runs) and the D3F (⋆, 10^4 runs); the asymptotic rates (dot-dashed) given by (39). The rightmost plots show the rate functions under both hypotheses, computed via exponential tilting. The bottom-right plot is a close-up of the rectangular region represented by the box in the top-right plot and shows the test operational points given by the thresholds γ_{D3F} and γ_{LLR} for the D3F and LLR, respectively. The D3F is trained with $m = 10^3$ samples.

comparison, we also need to select two thresholds, indicated with γ_{LLR} and γ_{D3F} for the LLR and the D3F, respectively (see bottom-right panel in Fig. 12, which provides a zoom of the upper-right panel), so that the rates under $\mathcal{H}_0$ for the LLR and D3F are equal. Once the thresholds have been chosen, the rates under $\mathcal{H}_1$ are automatically determined for both the LLR and the D3F in each of the points θ_1 and θ_2. In the left-hand side panels of Fig. 12 we show α_n (top) and β_n (bottom). Specifically, we report the empirical error probabilities computed by means of Monte Carlo simulations, along with the asymptotic approximations, which again are in a very good agreement. Similar to the previous example (Fig. 10), the LLR and D3F achieve the same performance; this is also due to the fact that they have very similar rate functions, as reported in the Fenchel-Legendre transform curves in the right-side panels of Fig. 12. We also plot the theoretical asymptotic rate provided by (39), which intercepts the curves when n is sufficiently large. It is interesting to observe that α_n converges faster to zero than β_n, being the rate around 0.02, while β_n has two rates, the first is around 0.011 when $\theta_* = \theta_2$ and the second is below 0.005 when $\theta_* = \theta_1$. Our intuition suggests that the performance will always be worse for $\theta_* = \theta_1$, given that this value is closer to θ_0 than θ_2. Indeed, in the left-side part of the parabola the rate function of θ_2 is always larger than the rate function of θ_1 for both the LLR and the D3F, and this behaviour is confirmed in Fig. 10, where the error converges to zero faster when $\theta_* = \theta_2$; see the rate functions of $\mathcal{H}_1$ in the point in which the rate of $\mathcal{H}_0$ is null in μ_0 (around -0.04 for the LLR and -0.03 for the D3F). We have the opposite behaviour on the right side of the parabola, where the rate of θ_1 is larger than the rate of θ_2; however, this region is not relevant for our scopes, as there are always no large deviations for β_n under θ_1, and no large deviations for both values of θ when the threshold is larger than $\mu_1(\theta_2)$.

C. EXTENDED TARGET DETECTION

In order to corroborate the results in Section V-C, we consider now a more complex scenario, where the data are representative of the output of a high-resolution sensor, e.g., a radar, a sonar, or an optical camera. The goal is to decide about the presence/absence of a target in the surveillance region. In this scenario, the D3F is the output of a DCNN, whose input is an image of 500 by 500 pixels; see Fig. 6 for a schema of the DCNN architecture.

In this setup, the DCNN is trained starting from 5120 examples of elliptical targets and as many examples of rectangular targets, whose position and dimensions are generated randomly; targets are also rotated by a random angle (some notional examples of images are reported in Fig. 2). All the target shape parameters are selected independently, and the random generation allows for partial observations (target on the boundary only partially visible) to be represented in the training set. For each realization of the target shape parameters, 5 independent realizations of the background noise are generated. In this way, the resulting training set is composed by 76800 images, of which 25600 generated from the null hypothesis (absence of target), and 51200 from the alternative one (one half are targets with elliptical shape and the other half are targets with rectangular shape). All the images are generated in agreement with the model described in Section V-C. The training set is generated with pixel-wise detection and false alarm probabilities $p_1 = 0.9$ and $p_0 = 0.1$, respectively. The threshold is selected as in (53) to achieve approximately a desired false alarm level of 0.3, 0.075 and 0.005. The related miss detection probability is reported in the upper panel of Fig. 13, where we simulate a target with a circular shape, located in the center of the image, with n being the number of resolution cells occupied by the target. The miss detection probability is estimated via Monte Carlo simulations with 10^5 runs, and compared with the exact asymptotics via the direct estimation method and the Gaussian approximation.

To further shade light on the detector behaviour, we also perform a sensitivity analysis with respect to the pixel-wise detection and false alarm probabilities, p_1 and p_0, respectively. Specifically, we simulate the mismatched scenario where $p_1 = 0.8$ and $p_0 = 0.2$, reported in the upper panel of Fig. 13.

As observed in the previous subsections, the Gaussian approximation is quite good for small values of n, whereas (as it is expected) the empirical miss detection probabilities are closer to the exact asymptotic as n increases. The curves, obtained via exact asymptotics, are displayed starting from values of n where the approximation is meaningful, i.e., when the expected mean of the decision statistic is larger than the threshold (see the panel at the bottom of Fig. 13) for the matched scenario.

Clearly, the performance of the mismatched scenario is worse than the matched one given that the pixel-wise false alarm rate is higher (0.2 instead of 0.1) and the pixel-wise detection rate is lower (0.8 instead of 0.9). However, it is worthwhile highlighting that, even if the training and the actual input data exhibit a significant mismatch, the D3F is still able to perform target detection when the size of the target increases.

For each value of the desired false alarm level α, we estimate the actual false alarm level and plot it in the mid panel of

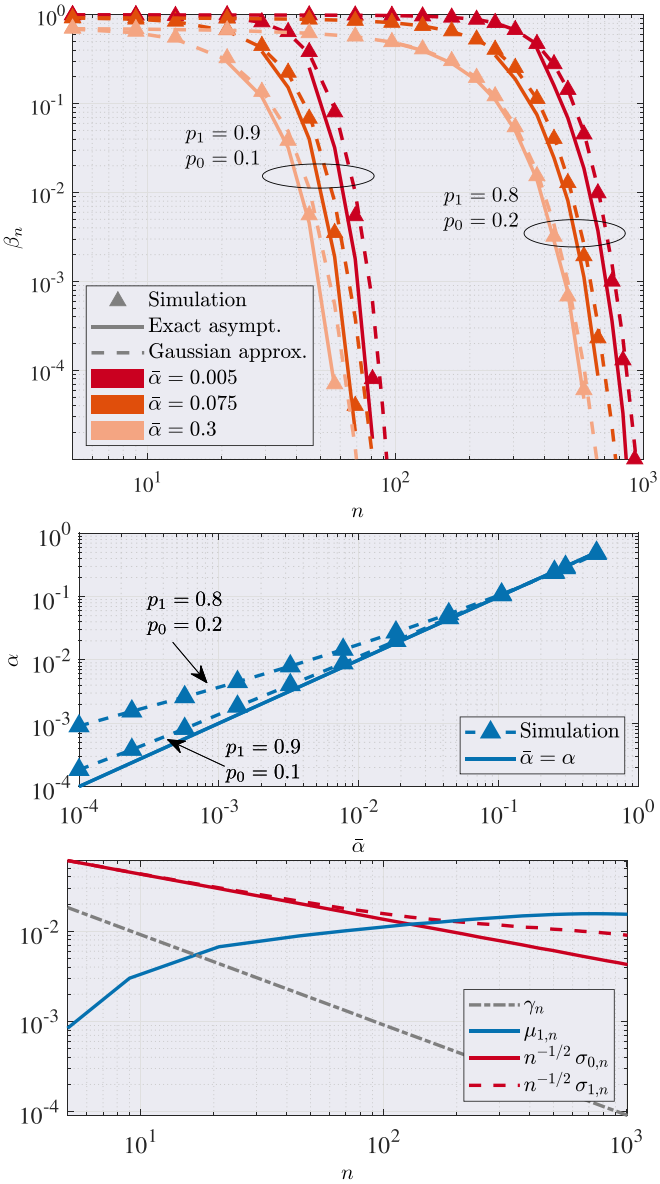


FIGURE 13. Detection of presence/absence of an extended target in a binary image. Top panel: miss detection probability β_n, with n being the number of resolution cells occupied by the target. The error probability is computed for three different false alarm levels via Monte Carlo simulations (▲) with 10^5 runs, and compared with the exact asymptotic and Gaussian approximations. Two scenarios are considered: in the former, the detection probability and false alarm probability of each cell resolution (pixel) are $p_1 = 0.9$ and $p_0 = 0.1$, respectively; in the latter, they are $p_1 = 0.8$ and $p_0 = 0.2$, respectively. Middle panel: desired false alarm level $\bar{\alpha}$ vs false alarm level observed in the Monte Carlo simulation (▲). Bottom panel: convergence of γ_n, $\mu_{1,n}$ and $\sqrt{n}\sigma_{k,n}$ for $k = 0, 1$, where $\gamma_n = \frac{\gamma}{n}$, with γ given in (53), while $\mu_{1,n}$ and $\sqrt{n}\sigma_{k,n}$ are the parameters of the normalized decision statistic (47).

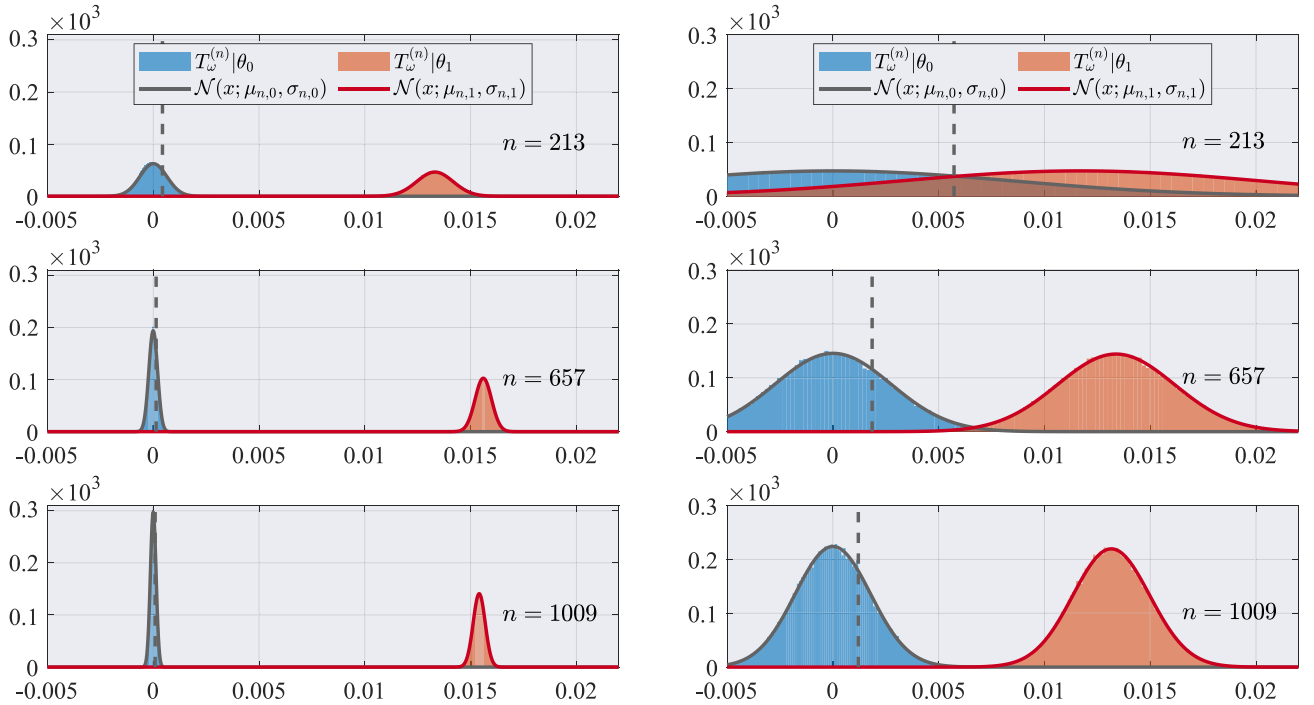


FIGURE 14. Empirical distribution of the normalized D3F statistics (47) for the problem of detecting an extended target in an image under $\mathcal{H}_0$ (blue) and $\mathcal{H}_1$ (orange). Left column, matched scenario: $P_1 = 0.9$, $P_0 = 0.1$. Right column, mismatched scenario: $P_1 = 0.8$, $P_0 = 0.2$. The D3F empirical distribution is compared with the Gaussian distributions under $\mathcal{H}_0$ (solid gray line) and $\mathcal{H}_1$ (solid red line). The Gaussian distribution parameters, $\mu_{n,k}$ and $\sigma^2_{n,k}$, are provided by the small deviations, and estimated from the characterization set. The D3F is translated to obtain $\mu_{n,0} = 0$. Panels in each row refer to the same value of n.

Fig. 13. We observe that the Gaussian approximation is less accurate as α decreases (between 10^{-4} and 10^{-3}) because the tail of the decision statistic is not necessarily Gaussian, and would be more accurately ruled by the related LMGF and its Fenchel-Legendre transform. The Gaussian approximation, discussed in Section V-D, is pretty accurate for both hypotheses as long as we focus on values of the decision statistic close to the mean, as we can observe in the histograms of the statistic (47), reported in Fig. 14 for both the matched (left-side panels) and mismatched scenarios (right-side panels). A fixed offset to the decision statistic is applied so that the resulting expected mean of decision statistic distribution, under $\mathcal{H}_0$, is zero. As explained in Section V-D, the decision statistic under $\mathcal{H}_0$ is independent on n [see (52)]. Indeed, the standard deviation $\sigma_{0,n}$ of the statistic (47) converges to zero faster than $\frac{1}{\sqrt{n}}$ (see the bottom panel in Fig. 13). However, the data are still well approximated by a Gaussian distribution, as illustrated in Fig. 14. On the other hand, the convergence under $\mathcal{H}_1$ is verified empirically in agreement with the LDP. Indeed, $\mu_{1,n}$ converges to an asymptotic value μ_1, whereas $\sqrt{n}\sigma_{1,n}$ converges to an asymptotic value σ_1, as reported in the panel at the bottom of Fig. 13.

D. EXTENDED TARGET DETECTION IN A REAL-WORLD SCENARIO USING X-BAND RADAR DATA

In this subsection we consider real-world data collected by a coherent high-resolution X-band maritime radar located in the Gulf of La Spezia, Italy. The radar is a low power, compact, lightweight, quickly deployable system, while still achieving high performance with relatively simple electronics. This is thanks to the use of pulse compression and linear frequency modulated continuous wave. The radar has an antenna mounted on a rotor with variable rotating speed and the possibility to lock and hold the position towards a specific direction with a 0.1° accuracy. The range resolution is approximately 1 m, while the angular resolution is 0.172° [107]. More details about the radar system, the signal processing chain, and the data collection are available in [103], [107], [108], [109].

For our analysis, we consider three acquisitions from two large vessels and a tugboat, illustrated in the top row of Fig. 15. The Automatic Identification System (AIS) is used as ground truth; this is an common approach in literature (see, e.g., [109], [110], [111]). The vessels observed in the radar images are T1, a container ship (MMSI: 351361000) being pushed by T2, a tugboat (MMSI: 247222500), and a passenger ship (MMSI: 255803790). From the radar frame, a rectangular region is selected (indicated with a red box); then, persistent clutter is masked out. A pixel-wise maximum likelihood detector is applied in each resolution cell, see more details in [109]. The result of this processing step are binary images, reported in the second row of Fig. 15. It can be seen that, for each radar frame, it is possible to detect correctly the extended targets, with some clutter still being present (in the upper right hand corner of the second row's central panel).

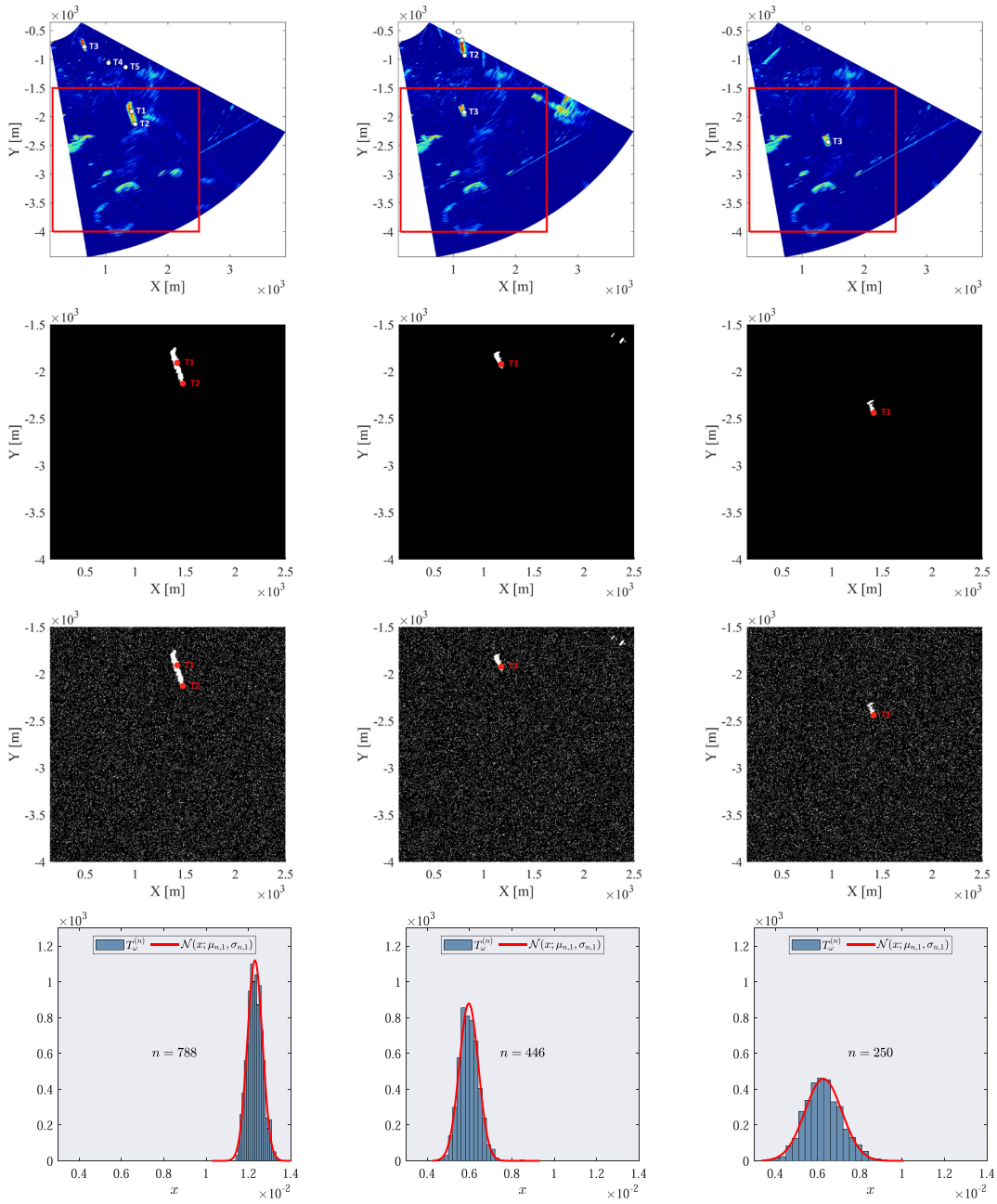


FIGURE 15. Analysis of the normalized D3F statistics (47) in the extended target scenario. Panels in the first row show three radar frames with AIS contacts overlaid. From these, a rectangular region (indicated with a red box) is extracted; then, persistent clutter is masked out and a detector applied. The output of this processing stage is illustrated in the second row, with the targets lighting up $n = 788$, $n = 446$ and $n = 250$ respectively, from left to right. Since the false alarm level is extremely low, synthetic clutter is introduced (third row). Finally, the empirical distribution of the normalized D3F statistic is computed from 1000 independent realizations and showed in the panels on the last row; the superimposed red curves are Gaussian distributions, showing that the D3F statistic is following the small deviations principle in all three cases.

The false alarm rate is very low, creditably so from the surveillance perspective but not so useful to explore our performance prediction results. For this reason we introduce at this point some synthetic clutter, as in the model (43)–(44). Examples of binary images with synthetic clutter added are reported in the third row of Fig. 15. These represent the input to compute the normalized D3F statistic (47) using the same DCNN, trained with synthetic data, described in the previous subsection. It is interesting to notice that the DCNN is able to detect the target correctly, even if it was not trained directly on radar data.

Finally, from the generation of independent pixel-wise false alarms, with $p_0 = 0.1$, we have multiple independent realizations of the normalized D3F statistic for each radar acquisition. Then, we can construct histograms of the decision statistic (last row of Fig. 15), which again appears

evidently Gaussian-distributed, thus confirming the convergence already discussed in Section V-D with real-world data.

VIII. CONCLUSION

In this paper, we have proposed a novel method to analyze the performance of a Machine Learning (ML) classification technique. The ML classifier is based on a suitable decision statistic, referred to as the Data-Driven Decision Function (D3F), which is learned from training data, and its performance is defined in terms of error probabilities and their convergence rates.

We have provided the mathematical conditions, based on the theory of large deviations, for the D3F to exhibit an error probability that vanishes exponentially with a parameter n, which represents the amount of information accumulated about the classification problem. Three different classification problems have been investigated: *i*) independent and identically distributed data with simple hypotheses, *ii*) conditionally independent data with composite hypotheses, and *iii*) (weakly) dependent data with an unknown dependency structure. In the first two settings, the parameter n is the number of observations available for testing; in the third setting, it represents the size of an extended target in a radar image.

The mathematical conditions can be verified numerically exploiting the available set of data or by generating synthetic data. Thanks to the theory of large deviations, it has been shown that such conditions depend on the Fenchel-Legendre transform of the cumulant-generating function of the D3F. Moreover, we have derived approximations for two asymptotic regimes of the D3F statistic. In the first one, referred as to as the *small deviations* regime (related to the central limit theorem), it is possible to establish the convergence of the normalized D3F statistic to a Gaussian distribution for values of n large enough. This property can be used to set a desired, asymptotic false alarm probability, which turns out to be accurate also for operationally-relevant values of n. In the second regime, referred to as the *large deviations* regime, we have established the conditions for the error probabilities to vanish to zero exponentially. Then, exploiting the exact asymptotics, which is a refinement of the large deviations theory, we have also provided an accurate approximation of the error probability curve as function of n. All the theoretical findings are corroborated and supported by extensive numerical simulations, as well as by an analysis of real-world data acquired by an X-band marine surveillance radar system.

We believe that the proposed method is especially relevant in the context of model-based ML, where the statistical structure of problem under investigation is partially known or can be easily simulated.

APPENDIX A
LIMIT OF THE LLR'S SCALED LMGF WITH CONDITIONALLY INDEPENDENT OBSERVATIONS

Let us compute the asymptotic scaled LMGF of $L^{(n)}$ (38) under $\mathcal{H}_1$ when the true parameter is θ_* and the data are generated according to $f(\cdot|\theta_*)$. Hereafter, we omit the conditioning on the hypothesis in the expected values to simplify the notation. We can write the equality (37) and note that the first term converges to zero assuming a non-trivial prior $w_{\theta_*} > 0$. The terms at the exponent in R_n converge to the KL divergence between θ_* and θ when θ_* is true. Specifically, given that the data are IID, the following holds

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_*)}{f(x_i|\theta)} \xrightarrow{a.s.} \mathcal{D}(\theta_*||\theta) = \mathbb{E}\left[\log\frac{f(x_i|\theta_*)}{f(x_i|\theta)}\right] > 0, \tag{61}$$

where the convergence is a.s. assuming mild regularity conditions, and that the hypotheses identified by the different values of θ are distinguishable, namely $\mathcal{D}(\theta_*||\theta) > 0$, $\forall\,\theta_*, \theta \in \Theta$ with $\theta_* \neq \theta$. Under mild regularity conditions of the log-likelihood ratio the convergence holds also in mean square. At this point it is easy to recognize in (37) that R_n vanishes exponentially fast with n. The asymptotic LMGF of the first term $L^{(n)}_{\theta_*}$ is given by:

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\exp(n\,t\,L^{(n)}_{\theta_*})\right] = \varphi_{\theta_*}(t), \tag{62}$$

with $\varphi_{\theta_*}(t) = \log\mathbb{E}\left[\exp\left(t\,\log\frac{f(x_1|\theta_*)}{f(x_1|\theta_0)}\right)\right]$. The asymptotic LMGF of $L^{(n)}$ is given by:

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)(1+R_n)^t\right]. \tag{63}$$

We shall now multiply and divide by the expected value $\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)\right] = \exp(n\,\varphi_{\theta_*}(t))$, obtaining

$$\begin{aligned}&\frac{1}{n}\log\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)(1+R_n)^t\right]\\ &\quad= \varphi_{\theta_*}(t) + \frac{1}{n}\log\frac{\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)(1+R_n)^t\right]}{\exp\left(n\,\varphi_{\theta_*}(t)\right)}.\end{aligned} \tag{64}$$

Since $R_n > 0$, the second term in (64) is larger than $\frac{1}{n}\log\frac{\mathbb{E}[\exp(n\,t\,L^{(n)}_{\theta_*})]}{\exp(n\,\varphi_{\theta_*}(t))} = 0$ (the numerator and the denominator are equal). Then, we have

$$\frac{1}{n}\log\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)(1+R_n)^t\right] \geq \varphi_{\theta_*}(t). \tag{65}$$

We can now apply Hölder's inequality to the argument of the expectation in (63) and obtain

$$\begin{aligned}&\frac{1}{n}\log\mathbb{E}\left[\exp\left(n\,t\,L^{(n)}_{\theta_*}\right)(1+R_n)^t\right]\\ &\quad\leq \frac{1}{n\,p}\log\mathbb{E}\left[\exp\left(n\,p\,t\,L^{(n)}_{\theta_*}\right)\right] + \frac{1}{n\,q}\log\mathbb{E}\left[(1+R_n)^{q\,t}\right]\\ &\quad= \frac{1}{p}\varphi_{\theta_*}(p\,t) + \frac{1}{n\,q}\log\mathbb{E}\left[(1+R_n)^{q\,t}\right],\end{aligned} \tag{66}$$

where $p^{-1} + q^{-1} = 1$. The second term in the equation above vanishes if R_n (which is strictly positive) grows as n^r with $r < 1$. Now, given that R_n vanishes exponentially in probability thanks to the continuous mapping, the term $(1+R_n)^{q\,t} \to 0$

in probability. The convergence in mean holds if the function $(1+x)^{qt}$ is bounded, which is the case assuming a limited interval $t_{\min} < t < t_{\max}$, or if the convergence of R_n is a dominated convergence (see details in [112]). Consequently, we have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp\left(n\,t\,L_{\theta_*}^{(n)}\right)(1+R_n)^t\right] \leq \frac{1}{p}\varphi_{\theta_*}(p\,t),$$
$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp\left(n\,t\,L_{\theta_*}^{(n)}\right)(1+R_n)^t\right] \geq \varphi_{\theta_*}(t), \quad (67)$$

where in the second inequality we have exploited the continuity of $\varphi_{\theta_*}(t)$; the first inequality holds also in the limit of $p \to 1$. Considering both (65) and (67), we can conclude that the asymptotic scaled LMGF under $\mathcal{H}_1$ is $\varphi_{\theta_*}(t)$.

Let us focus on $\mathcal{H}_0$, remembering that the data are generated according to $f(\cdot|\theta_0)$. We can rewrite the log-likelihood ratio (38) as follows:

$$L^{(n)} = \frac{1}{n}\log w_{\theta_m} + \underbrace{\frac{1}{n}\log\prod_{i=1}^{n}\frac{f(x_i|\theta_m)}{f(x_i|\theta_0)}}_{L_{\theta_m}^{(n)}}$$
$$+ \frac{1}{n}\log\left(1 + \underbrace{\sum\nolimits_{\theta\in\Theta,\theta\neq\theta_m} w_\theta \exp\left(-nD_n(\theta)\right)}_{R_n}\right), \quad (68)$$

where

$$\theta_m = \arg\min_{\theta\in\Theta} \mathcal{D}(\theta_0||\theta)$$

and

$$D_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_0)}{f(x_i|\theta)} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(x_i|\theta_0)}{f(x_i|\theta_m)}.$$

The term $D_n(\theta)$ converges a.s. to the difference between the divergences $\mathcal{D}(\theta_0||\theta) - \mathcal{D}(\theta_0||\theta_m)$, which is strictly positive if there is a single point of minimum θ_m (in the case of multiple minimum points the procedure is analogous). The structure of the log-likelihood ratio (68) is formally the same of (37), then with similar augmentations we can show that the asymptotic scaled LMGF under $\mathcal{H}_0$ is given by $\varphi_{\theta_0}(t) = \log\mathbb{E}\left[\exp\left(t\,\log\frac{f(x_1|\theta_m)}{f(x_1|\theta_0)}\right)\right]$, where the expectation is taken under $\mathcal{H}_0$ and the true parameter is θ_0.

APPENDIX B
WEAK LAW OF LARGE NUMBERS WHEN THE CHARACTERIZATION SET IS EQUAL TO THE TRAINING SET

It is possible to exploit the training set to compute the quantities of interest, such as the LMGF, as indicated in Section VI. This is the case when the characterization set and the training set are the same or overlapped. The goal is to show that the sample means, such as (54) and (55), converge to their expected means even if $\tau_j = t_{\boldsymbol{\omega}_\mathcal{Y}}(y_j), \forall j = 1, 2, \ldots, m_y$, where $\boldsymbol{\omega}_\mathcal{Y}$ is a stochastic function of the training data $\mathcal{Y} = (y_1, y_2, \ldots, y_{m_y})$, provided by the learning mechanism.[32] Therefore, we should consider a joint space probability between $\mathcal{Y}$ and $\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is referring to the aforementioned $\boldsymbol{\omega}_\mathcal{Y}$. Let us also assume that $\boldsymbol{\omega}$ takes values in a finite and discrete space Ω. The dependency of $\boldsymbol{\omega}$ from the size of the training set m_y is omitted for simplicity.

Let us consider the following event for any given $\epsilon > 0$

$$\mathcal{A}_\epsilon(\boldsymbol{\omega}, \mathcal{Y}) = \left\{\left|\frac{1}{m_y}\sum_{j=1}^{m_y} g(t_{\boldsymbol{\omega}}(y_j)) - \bar{g}_{\boldsymbol{\omega}}\right| > \epsilon\right\}, \quad (69)$$

where $\bar{g}_{\boldsymbol{\omega}} = \mathbb{E}[g(t_{\boldsymbol{\omega}}(y))|\boldsymbol{\omega}]$, where y is distributed accordingly to the marginal distribution of an element in $\mathcal{Y}$, which is independent of $\boldsymbol{\omega}$. Let us now evaluate the probability of the event of interest

$$\begin{aligned}\mathbb{P}\left[\mathcal{A}_\epsilon(\boldsymbol{\omega}, \mathcal{Y})\right] &= \sum_{\bar{\boldsymbol{\omega}}\in\Omega} \mathbb{P}\left[\mathcal{A}_\epsilon(\boldsymbol{\omega}, \mathcal{Y}) \cap \{\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}\}\right] \\ &= \sum_{\bar{\boldsymbol{\omega}}\in\Omega} \mathbb{P}\left[\mathcal{A}_\epsilon(\bar{\boldsymbol{\omega}}, \mathcal{Y})\right]\mathbb{P}\left[\{\boldsymbol{\omega} = \bar{\boldsymbol{\omega}}\}\,|\mathcal{A}_\epsilon(\bar{\boldsymbol{\omega}}, \mathcal{Y})\right] \\ &\leq \sum_{\bar{\boldsymbol{\omega}}\in\Omega} \mathbb{P}\left[\mathcal{A}_\epsilon(\bar{\boldsymbol{\omega}}, \mathcal{Y})\right]. \quad (70)\end{aligned}$$

Note that by construction when we evaluate $\mathbb{P}[\mathcal{A}_\epsilon(\bar{\boldsymbol{\omega}}, \mathcal{Y})]$, $\boldsymbol{\omega}$ is fixed to the value $\bar{\boldsymbol{\omega}}$, and the evaluation of this probability is carried on according to the marginal of $\mathcal{Y}$. Now, being the observations y_i IID, by invoking the law of large numbers we have that for $m_y \to \infty$ for any $\bar{\boldsymbol{\omega}}$

$$\mathbb{P}\left[\mathcal{A}_\epsilon(\bar{\boldsymbol{\omega}}, \mathcal{Y})\right] \to 0. \quad (71)$$

From (70)–(71), it follows directly that $\mathbb{P}[\mathcal{A}_\epsilon(\boldsymbol{\omega}, \mathcal{Y})] \to 0$.

REFERENCES

[1] Y. K. Dwivedi et al., "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manage.*, vol. 57, 2021, Art. no. 101994.

[2] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[3] D. Bertsekas, *Rollout, Policy Iteration, and Distributed Reinforcement Learning*. Nashua, NH, USA: Athena Scientific, 2021.

[4] D. Bertsekas, *Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control*. Nashua, NH, USA: Athena Scientific, 2022.

[5] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[6] D. Silver et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[7] C. Lippert et al., "Identification of individuals by trait prediction using whole-genome sequencing data," *Proc. Nat. Acad. Sci.*, vol. 114, no. 38, pp. 10166–10171, 2017.

[8] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome: Clin. Res. Rev.*, vol. 14, no. 4, pp. 337–339, 2020.

[9] G. Soldi et al., "Quickest detection and forecast of pandemic outbreaks: Analysis of COVID-19 waves," *IEEE Commun. Mag.*, vol. 59, no. 9, pp. 16–22, Sep. 2021.

[32] Note that we omit the dependency from the hypothesis to simplify the notation.

[10] P. Braca, D. Gaglione, S. Marano, L. M. Millefiori, P. Willett, and K. R. Pattipati, "Decision support for the quickest detection of critical COVID-19 phases," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, 2021.

[11] S. Zhou et al., "Application of hidden Markov models to analyze, group and visualize spatio-temporal COVID-19 data," *IEEE Access*, vol. 9, pp. 134384–134401, 2021.

[12] S. B. Desai, A. Pareek, and M. P. Lungren, "Deep learning and its role in COVID-19 medical imaging," *Intell.-Based Med.*, vol. 3, 2020, Art. no. 100013.

[13] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[16] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature Neurosci.*, vol. 23, no. 4, pp. 575–582, 2020.

[17] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 33–47, Jan. 2022.

[18] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, 2020.

[19] S. Chen, H. Wang, F. Xu, and Y.-Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[20] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[21] S. A. Wagner, "SAR ATR by a combination of convolutional neural network and support vector machines," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2861–2872, Dec. 2016.

[22] A. H. Oveis, E. Giusti, S. Ghio, and M. Martorella, "Extended openmax approach for the classification of radar images with a rejection option," *IEEE Trans. Aerosp. Electron. Syst.*, early access, Jun. 17, 2022, doi: 10.1109/TAES.2022.3183953.

[23] G. Soldi et al., "Space-based global maritime surveillance Part I: Satellite technologies," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 9, pp. 8–28, Sep. 2021.

[24] G. Soldi et al., "Space-based global maritime surveillance Part II: Artificial intelligence and data fusion techniques," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 9, pp. 30–42, Sep. 2021.

[25] N. Forti, E. d'Afflisio, P. Braca, L. M. Millefiori, S. Carniel, and P. Willett, "Next-gen intelligent situational awareness systems for maritime surveillance and autonomous navigation," *Proc. IEEE*, vol. 110, no. 10, pp. 1532–1537, Oct. 2022.

[26] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 4329–4346, Dec. 2021.

[27] S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Uncertainty-aware recurrent encoder-decoder networks for vessel trajectory prediction," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2021, pp. 1–5.

[28] B. Murray and L. P. Perera, "An AIS-based deep learning framework for regional ship behavior prediction," *Rel. Eng. Syst. Saf.*, vol. 215, 2021, Art. no. 107819.

[29] S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Recurrent encoder-decoder networks for vessel trajectory prediction with uncertainty estimation," *IEEE Trans. Aerosp. Electron. Syst.*, early access, Nov. 11, 2022, doi: 10.1109/TAES.2022.3216823.

[30] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[31] P. Bergmann, T. Meinhardt, and L. Leal-Taixé, "Tracking without bells and whistles," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 941–951.

[32] N. Samuel, T. Diskin, and A. Wiesel, "Learning to detect," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2554–2564, May 2019.

[33] F. Sohrabi, T. Jiang, W. Cui, and W. Yu, "Active sensing for communications by learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1780–1794, Jun. 2022.

[34] L. Wang, J. Tang, and Q. Liao, "A study on radar target detection based on deep neural networks," *IEEE Sens. Lett.*, vol. 3, no. 3, pp. 1–4, Mar. 2019.

[35] G. Soldi et al., "Underwater tracking based on the sum-product algorithm enhanced by a neural network detections classifier," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 5460–5464.

[36] D. Gaglione, G. Soldi, P. Braca, G. De Magistris, F. Meyer, and F. Hlawatsch, "Classification-aided multitarget tracking using the sum-product algorithm," *IEEE Signal Process. Lett.*, vol. 27, pp. 1710–1714, 2020.

[37] Z. Geng, H. Yan, J. Zhang, and D. Zhu, "Deep-learning for radar: A survey," *IEEE Access*, vol. 9, pp. 141800–141818, 2021.

[38] Z. Baird, M. K. McDonald, S. Rajan, and S. Lee, "A Neyman-Pearson criterion-based neural network detector for maritime radar," in *Proc. IEEE Int. Conf. Inf. Fusion*, 2021, pp. 1–8.

[39] S. M. D. Rizvi, S. Ahmad, K. Khan, A. Hasan, and A. Masood, "Deep learning approach for fixed and rotary-wing target detection and classification in radars," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 3, pp. 32–42, Mar. 2022.

[40] E. Blasch et al., "Machine learning/artificial intelligence for sensor data fusion–opportunities and challenges," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 7, pp. 80–93, Jul. 2021.

[41] W. Koch, "On digital ethics for artificial intelligence and information fusion in the defense domain," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 7, pp. 94–111, Jul. 2021.

[42] S. Yu and J. C. Principe, "Understanding autoencoders with information theoretic concepts," *Neural Netw.*, vol. 117, pp. 104–123, 2019.

[43] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–13.

[44] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, Aug. 2014.

[45] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop*, 2015, pp. 1–5.

[46] S. Mallat, "Understanding deep convolutional networks," *Philos. Trans. Roy. Soc. A: Math., Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, Art. no. 20150203.

[47] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," 2017, *arXiv:1703.00810*.

[48] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, "Deep neural networks as gaussian processes," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–17.

[49] R. Novak et al., "Bayesian deep convolutional networks with many channels are gaussian processes," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–35.

[50] R. M. Neal, *Priors for Infinite Networks*. New York, NY, USA: Springer, 1996, pp. 29–53.

[51] R. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inf. Theory*, vol. 20, no. 4, pp. 405–417, Jul. 1974.

[52] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–408, Mar. 1989.

[53] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal?," *IEEE Trans. Inf. Theory*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.

[54] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura et al., "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4381–4396, Sep. 2011.

[55] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime," *IEEE Trans. Inf. Theory*, vol. 62, no. 8, pp. 4710–4732, Aug. 2016.

[56] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Distributed detection over adaptive networks: Refined asymptotics and the role of connectivity," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 4, pp. 442–460, Dec. 2016.

[57] S. Marano and P. K. Willett, "Algorithms and fundamental limits for unlabeled detection using types," *IEEE Trans. Signal Process.*, vol. 67, no. 8, pp. 2022–2035, Apr. 2019.

[58] S. Marano and P. Willett, "Making decisions by unlabeled bits," *IEEE Trans. Signal Process.*, vol. 68, pp. 2935–2947, 2020.

[59] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[60] Y. Xu and R. Goodacre, "On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *J. Anal. Testing*, vol. 2, no. 3, pp. 249–262, 2018.

[61] C. M. Bishop, "Model-based machine learning," *Philos. Trans. Roy. Soc. A: Math., Phys. Eng. Sci.*, vol. 371, no. 1984, 2013, Art. no. 20120222.

[62] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. New York, NY, USA: Springer-Verlag, 1998.

[63] F. Den Hollander, *Large Deviations*, vol. 14. Ann Arbor, MI, USA: Amer. Math. Soc., 2008.

[64] E. Lehmann, *Testing Statistical Hypotheses*. Berlin, Germany: Springer, 2005.

[65] H. Touchette, "The large deviation approach to statistical mechanics," *Phys. Rep.*, vol. 478, no. 1–3, pp. 1–69, 2009.

[66] H. Touchette, "A basic introduction to large deviations: Theory, applications, simulations," 2011, *arXiv:1106.4146*.

[67] N. Reid, "Saddlepoint methods and statistical inference," *Stat. Sci.*, vol. 3, no. 2, pp. 213–227, 1988.

[68] W. Bryc, "A remark on the connection between the large deviation principle and the central limit theorem," *Statist. Probability Lett.*, vol. 18, no. 4, pp. 253–256, 1993.

[69] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I–Fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.

[70] R. S. Blum, S. A. Kassam, and H. V. Poor, "Distributed detection with multiple sensors: Part II–Advanced topics," *Proc. IEEE*, vol. 85, no. 1, pp. 64–79, Jan. 1997.

[71] P. Braca, S. Marano, V. Matta, and P. Willett, "Asymptotic optimality of running consensus in testing statistical hypotheses," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 814–825, Feb. 2010.

[72] G. Gennarelli, G. Vivone, P. Braca, F. Soldovieri, and M. G. Amin, "Multiple extended target tracking for through-wall radars," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6482–6494, Dec. 2015.

[73] K. Granström, A. Natale, P. Braca, G. Ludeno, and F. Serafino, "Gamma Gaussian inverse Wishart probability hypothesis density for extended target tracking using X-band marine radar data," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6617–6631, Dec. 2015.

[74] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality," *IEEE Trans. Inf. Theory*, vol. 68, no. 7, pp. 4790–4817, Jul. 2022.

[75] P. Braca, L. M. Millefiori, A. Aubry, S. Marano, A. De Maio, and P. Willett, "Large deviations for classification performance analysis of machine learning systems," 2023, *arXiv:2301.07104*.

[76] N. Shlezinger, N. Farsad, Y. C. Eldar, and A. J. Goldsmith, "Model-based machine learning for communications," 2021, *arXiv:2101.04726*.

[77] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, "Model-based deep learning: On the intersection of deep learning and optimization," in *IEEE Access*, vol. 10, pp. 115384–115398, 2022, doi: 10.1109/ACCESS.2022.3218802.

[78] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[79] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[80] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2794–2802.

[81] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[82] G. V. Moustakides and K. Basioti, "Training neural networks for likelihood/density ratio estimation," 2019, *arXiv:1911.00405*.

[83] R. Veldhuis and D. Zeng, "On Neyman-Pearson optimality of binary neural net classifiers," Tech. Rep., Jun. 2021, doi: 10.36227/techrxiv.14709930.v1.

[84] E. Lehmann, *Elements of Large Sample Theory*. New York, NY, USA: Springer, 2004.

[85] S. S. Varadhan, *Large Deviations and Applications*. Philadelphia, PA, USA: SIAM, 1984.

[86] P. Braca, S. Marano, V. Matta, and P. Willett, "Consensus-based Page's test in sensor networks," *Signal Process.*, vol. 91, no. 4, pp. 919–930, 2011.

[87] D. Jakovetić, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4306–4320, Aug. 2012.

[88] D. Li, S. Kar, J. M. F. Moura, H. V. Poor, and S. Cui, "Distributed Kalman filtering over massive data sets: Analysis through large deviations of random riccati equations," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1351–1372, Mar. 2015.

[89] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[90] A. T. Wood, J. G. Booth, and R. W. Butler, "Saddlepoint approximations to the CDF of some statistics with nonnormal limit distributions," *J. Amer. Stat. Assoc.*, vol. 88, no. 422, pp. 680–686, 1993.

[91] V. Y. Tan et al., "Asymptotic estimates in information theory with non-vanishing error probabilities," *Found. Trends Commun. Inf. Theory*, vol. 11, no. 1/2, pp. 1–184, 2014.

[92] H. E. Daniels, "Saddlepoint approximations in statistics," *Ann. Math. Statist.*, vol. 25, pp. 631–650, 1954.

[93] A. R. Barron, "Are Bayes rules consistent in information?," in *Open Problems in Communication and Computation*. New York, NY, USA: Springer, 1987, pp. 85–91.

[94] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[95] E. L. Lehmann and G. Casella, *Theory of Point Estimation* (Springer Texts in Statistics). Berlin, Germany: Springer, 1998.

[96] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[97] E. Tuncel, "On error exponents in hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2945–2950, Aug. 2005.

[98] J. W. Koch, "Bayesian approach to extended object and cluster tracking using random matrices," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, no. 3, pp. 1042–1059, Jul. 2008.

[99] J. Melo, D. Gaglione, G. Soldi, P. Braca, and S. Dugelay, "A belief propagation based algorithm for autonomous mapping of underwater targets," in *Proc. IEEE Glob. Ocean.: Singap.–USA Gulf Coast*, 2020, pp. 1–7.

[100] D. Williams and A. España, "Toward explainable convolutional neural network classifiers with acoustic-color sonar data," *J. Acoust. Soc. Amer.*, vol. 148, no. 4, pp. 2661–2661, 2020.

[101] D. P. Williams, "On the use of tiny convolutional neural networks for human-expert-level classification performance in sonar imagery," *IEEE J. Ocean. Eng.*, vol. 46, no. 1, pp. 236–260, Jan. 2021.

[102] G. Soldi, F. Meyer, P. Braca, and F. Hlawatsch, "Self-tuning algorithms for multisensor-multitarget tracking using belief propagation," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3922–3937, Aug. 2019.

[103] G. Vivone, P. Braca, K. Granstrom, and P. Willett, "Multistatic Bayesian extended target tracking," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 6, pp. 2626–2643, Dec. 2016.

[104] S. Asmussen and P. W. Glynn, *Stochastic Simulation: Algorithms and Analysis*, vol. 57. New York, NY, USA: Springer, 2007.

[105] J. Bucklew, *Introduction to Rare Event Simulation*. New York, NY, USA: Springer, 2004.

[106] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *Amer. Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[107] B. Errasti-Alcala, W. Fuscaldo, P. Braca, and G. Vivone, "Realistic extended target model for track before detect in maritime surveillance," in *Proc. IEEE OCEANS 2015-Genova*, 2015, pp. 1–9.

[108] B. Errasti-Alcala and P. Braca, "Track before detect algorithm for tracking extended targets applied to real-world data of X-band marine radar," in *Proc. IEEE 17th Int. Conf. Inf. Fusion*, 2014, pp. 1–8.

[109] G. Vivone and P. Braca, "Joint probabilistic data association tracker for extended target tracking applied to X-band marine radar data," *IEEE J. Ocean. Eng.*, vol. 41, no. 4, pp. 1007–1019, Oct. 2016.
[110] G. Vivone, L. M. Millefiori, P. Braca, and P. Willett, "Performance assessment of vessel dynamic models for long-term prediction using heterogeneous data," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6533–6546, Nov. 2017.
[111] S. Maresca, P. Braca, J. Horstmann, and R. Grasso, "Maritime surveillance using multiple high-frequency surface-wave radars," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5056–5071, Aug. 2014.
[112] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," in *Handbook Econometrics*, vol. 4. Amsterdam, The Netherlands: Elsevier, 1994, pp. 2111–2245.

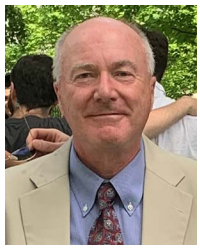
PAOLO BRACA (Senior Member, IEEE) received the Laurea degree (*summa cum laude*) in electronic engineering and the Ph.D. degree (with Hons.) in information engineering from the University of Salerno, Fisciano, Italy, in 2006 and 2010, respectively. In 2009, he was a Visiting Scholar with the Electrical and Computer Engineering Department, University of Connecticut, Storrs, CT, USA. From 2010 to 2011, he was a Postdoctoral Associate with the University of Salerno. In 2011, he joined the NATO Science and Technology Organization (STO) Centre for Maritime Research and Experimentation, La Spezia, Italy, where he is currently a Senior Scientist and a Project Manager. Furthermore, he led a number of research projects funded by the European Commission, by the U.S. Office of Naval Research (ONR), and other national and international institutions. In 2019, he was an Adjunct Professor with the University of Cassino, Italy. He is on the technical committee of the major international conferences in the field of signal processing and data fusion. He has coauthored more than 200 publications in international scientific journals, conference proceedings, and NATO technical reports. His research interests include the general area of machine learning and statistical signal processing with emphasis on detection and estimation theory, wireless sensor networks, multiagent algorithms, target tracking and data fusion, adaptation and learning over graphs, radar (sonar) signal processing, and machine learning. He was awarded with the National Scientific Qualification to function as an Associate Professor and a Full Professor in Italian universities, in 2017 and 2018, respectively. Dr. Braca was the recipient of the Best Student Paper Award (first runner-up) at the International Conference on Information Fusion (FUSION) in 2009, and NATO Science and Technology Organization (STO) Scientific Achievement Award, in 2017, for its contribution to the Development and Demonstration of Networked Autonomous ASW. He coauthored the paper and was the recipient of the Best Paper Award (first runner-up) at the International Conference in Sensor Signal Processing for Defence (SSPD) in 2019. He was also the recipient the IET 2019 Premium Award for the best paper published on the *IET Radar, Sonar & Navigation*, and NATO STO SAA, in 2020, as a Team Leader for the Advances in Artificial Intelligence and Information Fusion for Maritime Situational Awareness. He coauthored the article was the recipient of the Young Scientist Contest Award at the Signal Processing Symposium 2021. He is an Associate Editor for IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, *ISIF Journal of Advances in Information Fusion*, and *IET Radar, Sonar & Navigation*. In 2017, he was the Lead Guest Editor for the Special Issue on Sonar Multisensor Applications and Techniques of *IET Radar, Sonar & Navigation*. He was an Associate Editor for IEEE SIGNAL PROCESSING MAGAZINE, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and *EURASIP Journal Advances in Signal Processing*.

LEONARDO M. MILLEFIORI (Member, IEEE) received the B.Sc. degree in aerospace information engineering and the M.Sc. degree (*summa cum laude*) in telecommunication engineering with a focus on radar systems and remote sensing from Sapienza University of Rome, Rome, Italy, in 2010 and 2013, respectively, and the Ph.D. degree (with Hons.) in information engineering from the University of Pisa, Pisa, Italy, in 2022. He was a Visiting Researcher with the NATO Science and Technology Organization Centre for Maritime Research and Experimentation, La Spezia, Italy, where he joined the Research Department, as a Research Scientist in 2014. His research interests include target motion modeling, statistical learning and signal processing, target tracking, data fusion, and machine learning. Since December 2020, Dr. Millefiori has been an Associate Member of the NATO Sensors and Electronics Technology (SET) Panel. In 2020, he was the recipient of the NATO STO Scientific Achievement Award as one of the leaders of the Data Knowledge and Operational Effectiveness Research Group, Centre for Maritime Research and Experimentation, for Advances in Artificial Intelligence and Information Fusion for Maritime Situational Awareness.

AUGUSTO AUBRY (Senior Member, IEEE) received the Dr. Eng. degree in telecommunication engineering (with Hons.) and the Ph.D. degree in electronic and telecommunication engineering from the University of Naples Federico II, Naples, Italy, in 2007 and 2011, respectively. From February to April 2012, he was a Visiting Researcher with Hong Kong Baptist University, Hong Kong. He is currently an Assistant Professor with the University of Naples Federico II. His research interests include statistical signal processing and optimization theory, with emphasis on MIMO communications and radar signal processing. He was the co-recipient of the 2013 Best Paper Award (entitled to B. Carlton) of the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS with the contribution Knowledge-Aided (Potentially Cognitive) Transmit Signal and Receive Filter Design in Signal-Dependent Clutter. Dr. Aubry was the Recipient of the 2022 IEEE Fred Nathanson Memorial Award as the young (less than 40 years of age) AESS Radar Engineer 2022, with the following citation For outstanding contributions to the application of modern optimization theory to radar waveform design and adaptive signal processing.

STEFANO MARANO (Senior Member, IEEE) received the Laurea (*summa cum laude*) degree in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Naples, Naples, Italy, in 1993 and 1997, respectively. He is currently a Professor with Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, Fisciano, Italy, and Research Affiliate with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. He held visiting positions with the University of Wales, College of Cardiff, Cardiff, U.K., and ECE Department, University of California San Diego, La Jolla, CA, USA, in 1996 and 2013, respectively. His research interests include decision theory, distributed inference, information theory, and quantum information science. He was the recipient of the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION 1999 Best Paper Award. He was with the Organizing Committee of FUSION 2006 and RADARCON 2008. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING during 2010–2014, and an Associate Editor and the Technical Editor of the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS during 2009–2016.

ANTONIO DE MAIO (Fellow, IEEE) was born in Sorrento, Italy, on June 20, 1974. He received the Dr. Eng. degree (with Hons.) and the Ph.D. degree in information engineering from the University of Naples Federico II, Naples, Italy, in 1998 and 2002, respectively. He is currently a Professor with the University of Naples Federico II. His research interest include statistical signal processing, with emphasis on radar detection and optimization theory applied to radar signal processing. Dr. De Maio was the recipient of the 2010 IEEE Fred Nathanson Memorial Award as the young (less than 40 years of age) AESS Radar Engineer 2010 whose performance is particularly noteworthy as evidenced by contributions to the radar art over a period of several years, with the following citation for robust CFAR detection, knowledge-based radar signal processing, and waveform design and diversity.



PETER WILLETT (Fellow, IEEE) received the B.A.Sc. degree in engineering science from the University of Toronto, Toronto, ON, Canada, in 1982, and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 1986. Since 1986, he has been a Faculty Member of the Electrical and Computer Engineering Department, University of Connecticut, Storrs, CT, USA. Since 1998, he has been a Professor. Since 2003, he has been an IEEE Fellow. His research interests include statistical signal processing, detection, machine learning, communications, data fusion, and tracking.