Generalized Likelihood Ratio Test for Adversarially Robust Hypothesis Testing

Bhagyashree Puranik, *Student Member, IEEE*, Upamanyu Madhow, *Fellow, IEEE*, and Ramtin Pedarsani, *Senior Member, IEEE*

***Abstract*—Machine learning models are known to be susceptible to adversarial attacks, which can cause misclassification by introducing small but well designed perturbations. In this paper, we consider a classical hypothesis testing problem in order to develop fundamental insight into defending against such adversarial perturbations. We interpret an adversarial perturbation as a nuisance parameter, and propose a defense based on applying the generalized likelihood ratio test (GLRT) to the resulting composite hypothesis testing problem, jointly estimating the class of interest and the adversarial perturbation. While the GLRT approach is applicable to general multi-class hypothesis testing, we first evaluate it for binary hypothesis testing in white Gaussian noise under ℓ_∞ norm-bounded adversarial perturbations, for which a known minimax defense optimizing for the worst-case attack provides a benchmark. We derive the worst-case attack for the GLRT defense, and show that its asymptotic performance (as the dimension of the data increases) approaches that of the minimax defense. For non-asymptotic regimes, we show via simulations that the GLRT defense is competitive with the minimax approach under the worst-case attack, while yielding a better robustness-accuracy trade-off under weaker attacks. We also illustrate the GLRT approach for a multi-class hypothesis testing problem, for which a minimax strategy is not known, evaluating its performance under both noise-agnostic and noise-aware adversarial settings, by providing a method to find optimal noise-aware attacks, and ideas to find noise-agnostic attacks that are close to optimal in the high SNR regime. We show through experiments the application of the GLRT defense in colored Gaussian noise. We also demonstrate the use of GLRT defense beyond Gaussian settings by considering Laplacian noise and illustrating how our rule simplifies.**

***Index Terms*—Adversarial machine learning, hypothesis testing, robust classification.**

I. Introduction

WHILE discussion of security in machine learning predates deep learning [2], it becomes critical to address these concerns in view of the widespread adoption of deep neural networks in safety- and security-critical applications such as facial recognition for surveillance, autonomous driving and virtual assistants. In particular, it is known that deep neural networks are vulnerable to *adversarial attacks:* an adversary is often able to add small perturbations to data in an intelligent way to cause misclassification with high confidence [3], [4]. Studies have shown that adversarial examples exist even in real-world physical systems. For example, an adversarial attack can manipulate traffic signs to fool autonomous vehicles [5] or tamper with speech recognition systems [6], [7]. In applications that demand robustness, such adversarial attacks are fundamental threats, which motivates a rapidly growing body of research on both attacks and defenses. Some defenses are certifiably robust [8], [9], while others are empirical [10], [11]. Many suggested defenses have been broken by subsequent attacks [12], [13], [14]. The present state of the art defenses [10], [15], [16] are purely empirical, relying on *adversarial training,* wherein adversarial perturbations are applied while training the neural network. However, we do not yet have robustness guarantees or structural insights for such adversarially trained networks. Thus, existing defenses may be prone to new attacks that are conceived in the future, possibly taking advantage of the availability of increased computational power [17]. It is essential, therefore, to develop at least a statistical understanding of the robustness that can be provided by a classifier.

In this paper, we take a step back from deep neural networks, and attempt to develop fundamental insight into the impact of adversarial attacks on classification performance in the framework of classical hypothesis testing. Specifically, we investigate adversarial classification in the setting of composite hypothesis testing, in which the class-conditional distributions of the data are known, and the adversarial perturbation is treated as a nuisance parameter. We adopt a Generalized Likelihood Ratio Test (GLRT) formulation for defense against adversarial attacks, in which we jointly estimate the desired class and the action of the adversary.

Attack model: The observation to be classified is drawn from one of several class-conditional distributions, and the adversary has access to this observation when devising its attack. In the examples considered in this paper, the observation is modeled as a class-dependent signal plus noise. In this case, the adversary knows the class (and hence the signal) as well as the noise realization when formulating its attack. Our running example for attacks is an ℓ_∞-bounded additive perturbation, which can be tailored by the adversary to the specific realization of the signal and the noise. Such an adversary can be described as *noise-aware.* It

Manuscript received 14 November 2021; revised 9 June 2022; accepted 14 July 2022. Date of publication 11 August 2022; date of current version 24 August 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xue Jiang. This work was supported in part by the Army Research Office under Grant W911NF-19-1-0053, and in part by the National Science Foundation under Grants CIF-1909320 and CIF-2224263. This work was presented in part at the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing [DOI: 10.1109/ICASSP39728.2021]. *(Corresponding author: Bhagyashree Puranik.)*

The authors are with the Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: bpuranik@ucsb.edu; madhow@ece.ucsb.edu; ramtin@ece.ucsb.edu).

Digital Object Identifier 10.1109/TSP.2022.3198169

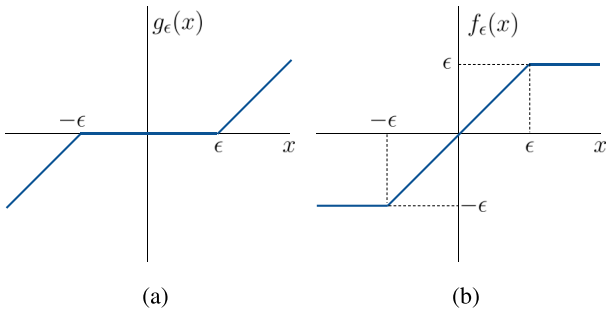

Fig. 1. The double-sided ReLU $g_\epsilon()$ and its "complement" $f_\epsilon()$.

is worth noting that this adversarial model is different from that in classical robust hypothesis testing [18], [19], [20], [21], in which the adversary (or nature) is constrained to a statistical attack; for example, for each hypothesis, the adversary may be allowed to choose the observation from a class-conditional distribution chosen from a set of distributions. In our running example of signal plus noise plus ℓ_∞-bounded adversarial perturbation, a *noise-agnostic* adversary whose perturbation depends only on the class-conditional signal can be interpreted within the classical robust hypothesis testing framework, even though it is not one of the typical models [21] considered in such settings.

GLRT versus minimax: The GLRT approach we consider *implicitly* defends against the adversary by jointly estimating its action along with the class. This approach applies to any composite hypothesis testing problem, although computation of the joint maximum likelihood estimate of the action and the class can be challenging. A minimax approach seeks to optimize worst-case classification performance, with the defender playing a game against the attacker. However, unlike the GLRT detector, the existence of a minimax solution is not guaranteed, unless the problem has a special structure. Furthermore, while the minimax solution is optimal against a worst-case attack, the GLRT approach potentially offers a better robustness/accuracy tradeoff by estimating and adapting to the attack.

We provide detailed insight into the comparison between the GLRT and minimax approaches via the example of binary Gaussian hypothesis testing, for which the minimax detector is known [22]. We briefly illustrate the differences between the two detectors here. It is shown in [22] that the minimax optimal classifier, which achieves the optimal adversarial risk (10), is a linear classifier with coefficients $\mathbf{w} = g_\epsilon(\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the class mean or the "signal template," ϵ is the attack budget, and $g_\epsilon(\cdot)$ is the "double-sided ReLU" function, as shown in Fig 1(a). Thus, the minimax classifier discards the signal coordinates of low strengths, specifically those whose signs could be flipped when the full attack budget is employed by the adversary. It retains the other coordinates after shrinking them by assuming the worst-case attack has been used, and provides a minimax optimal rule based only on these coordinates. In contrast, the proposed GLRT defense utilizes the signal strength in all the coordinates and applies the double-sided ReLU on a function of the received signal and template. Since GLRT estimates the perturbation, it adapts better when a weaker attack is employed, while minimax schemes are too pessimistic. This is the reason for better robustness-accuracy trade-off of our defense for different attack budgets.

Contributions: We summarize our contributions as follows:

- The well-known GLRT is proposed as a general approach to defense, in which the desired class and the perturbation are estimated jointly. The GLRT approach applies to any composite hypothesis testing problem [23], unlike minimax strategies optimizing for worst-case attacks, which are difficult to find.
- We compare the performance of the GLRT defense to a minimax strategy by considering binary Gaussian hypothesis testing with ℓ_∞ bounded attacks, for which the minimax strategy has been recently derived [22]. We demonstrate via an asymptotic analysis and by numerical evaluations that the GLRT approach provides competitive robustness when the attacker employs the full attack budget, while providing better robustness-accuracy trade-offs for weaker attacks.
- We illustrate via examples the application of the GLRT approach to settings for which minimax strategies are not known. The first example is multi-class Gaussian hypothesis testing, for which we provide an intuitively pleasing extension of the binary minimax classifier, which we term as the Pairwise Robust Linear classifier, to benchmark the performance of the GLRT. The second is binary hypothesis testing in Laplacian noise, for which we compare the GLRT against a non-robust maximum likelihood rule.
- We distinguish between noise-agnostic attacks (in which the attacker knows the correct hypothesis but not the noise realization) and noise-aware attacks (in which the attacker knows both the correct hypothesis and the noise realization). For the binary setting, we derive the worst-case attack for the GLRT defense, showing that the same attack is optimal for both noise-aware and noise-agnostic adversaries. For the multi-class setting, we provide an approach for finding the optimal noise-aware attack, and procedures to find a noise-agnostic attack which is close to the worst case at high SNR.

Notation: Throughout the paper, we represent vectors in boldface letters and scalars in regular letters. The norm $||\cdot||$ denotes ℓ_2 norm unless specified otherwise. We denote $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The symbols $\phi(\cdot)$, $\Phi(\cdot)$, and $Q(\cdot)$ represent the standard (zero-mean, unit variance) univariate Gaussian distribution, its cumulative distribution function (CDF) and the complementary CDF respectively.

II. Related Work

Certifiable robustness: There is a growing body of research on developing provable robustness guarantees against adversarial attacks. A provably robust defense was developed in [9], [24], which employs semidefinite programs and tight relaxations to train neural networks. The idea here is that although it is desirable to find defenses for all possible attacks, computation of worst case error is intractable, hence an upper bound is optimized as a regularizer during training. Another certifiable defense [8], [25] is based on linear programming and optimizing

over relaxed convex networks to bound the robustness. Other methods such as in [26] obtain guarantees under ℓ_2 attacks via a regularization functional, for small neural networks. Sparsity is exploited in [27] to provide a theoretical framework that guarantees robustness against ℓ_∞ attacks on linear classifiers, by introducing a front-end to the neural network that attenuates the impact of adversary. The above line of work on certifiable defenses considers tools from optimization to provide adversarially robust neural networks.

Insights from simplified models: Robustness in the presence of adversarial attacks has also been studied through signal processing tools, such as in [28], [29], [30]. [28] studies the robustness of subspace learning problems, such as principle component analysis, where data is modified by an adversary intentionally. The problem of minimax robust hypothesis testing with discrete-valued observations is considered in [29], which distinguishes between two settings: the attacker knows the true underlying hypothesis and the attacker is unaware of the true hypothesis. However, the adversary is noise-agnostic. Following the line of work in [20], they study the problem of minimizing the maximum probability of error and characterize optimal attack strategies in binary settings. The attack model is that after observing a sample, the adversary can change it to another sample probabilistically. However, the models considered in our work are more closely aligned with those in papers such as [22], [31], [32], [33], which seek fundamental insights into the problem of adversarial robustness by considering simplified continuous-valued observation models, ℓ_p norm bounded additive attack models and provide provable guarantees or design optimal classifiers. Limiting the capability of the attacker to norm-bounded perturbations is motivated by the rich literature on adversarial machine learning for image classification, where perturbations which are "imperceptible" to humans can be devised by imposing such bounds [13]. The paper [32] studies robust classification in the presence of ℓ_0 bounded adversaries under binary Gaussian setting. They propose a non-linear classifier that selects certain coordinates of the input and forms a test statistic based on a truncated inner product operation. They analyze the performance of the proposed classifier and derive bounds on the robust classification error. The problem of finding Bayes-optimal robust classifiers under binary and ternary settings with ℓ_2 and ℓ_∞ norm-bounded adversaries is addressed in [31], where class conditional distributions are Gaussian and possess symmetric means. Optimal robust classifiers are derived when the perturbations are ℓ_2 norm bounded. For the case when perturbations are ℓ_∞ norm bounded, they restrict attention to the class of linear classifiers and then obtain optimum robust linear classifiers among the restricted class. In general, finding robust optimal classifiers for ℓ_∞ norm bounded adversarial perturbations is not easily tractable. Analytical results for ℓ_∞ adversaries have been shown only for special cases, such as in [22], where minimax optimal robust classifiers are characterized in binary classification setting under Gaussian models with uniform priors, using ideas from optimal transport theory. We employ this minimax classifier as a benchmark throughout our paper, and discuss its results in Section III-C. Another recent paper [33] investigates optimal adversarial risk and classifiers by employing optimal transport theory. Comparable to [22], they characterize optimal adversarial risk for Gaussian models via optimal couplings, and apply similar strategies to find optimal classifiers for univariate uniform and triangular distributions.

Connections with classical robust detection: The model for classical robust hypothesis testing is that the distribution of the observation conditioned on a class c is itself selected from a set of distributions $\mathcal{P}_c$. While the work of Huber [18] established the fundamentals of this framework for binary hypothesis testing decades ago, recent work continues to significantly extend these ideas to a richer class of models and algorithms [20], [21]. Minimum robust detectors in such settings, when they exist, are the optimal decision rules corresponding to "least favorable" distributions selected from among the allowable set of distributions in each class. For specific outlier models [18], [20], it turns out that the robust likelihood function corresponding to the optimal decision rule for the least favorable distributions is a censored version of the likelihood function for nominal distributions. Interestingly, this robust likelihood function is obtained by passing the nominal likelihood function through a nonlinearity which is a scaled version of the double-sided ReLU that appears in our running example. Under our model, an ℓ_p bounded additive adversarial attack is computed based on access to the observation drawn from a nominal class-conditional distribution. When this nominal model is class-dependent signal plus noise, such an adversary is noise-aware. A noise-agnostic adversary, on the other hand, can indeed be modeled within the framework of classical robust hypothesis testing, with set of distributions $\mathcal{P}_c$ for class c indexed by the possible values taken by the ℓ_p bounded adversarial perturbation. To the best of our knowledge, this particular uncertainty model has not been considered in the literature in robust hypothesis testing (e.g., see the recent comprehensive survey [21]). For the binary Gaussian hypothesis testing example that we study in detail in this paper, the noise-aware and noise-agnostic attacks are identical. Thus, as briefly noted in Section III-C, in this specific setting, the minimax rule derived under the noise-aware attack in [22] is actually also minimax robust for noise-agnostic attacks, and can be interpreted within a classical robust hypothesis testing framework under certain conditions on the uncertainty sets. Further pursuit of such connections is an interesting direction for future work, as noted in our concluding remarks.

Our prior work: The work reported here builds on preliminary results reported in our conference paper [1], where we introduce the GLRT approach for robustness against adversarial perturbations in the white Gaussian setting and analyze the performance under a fixed attack, which is the worst-case attack for the minimax classifier. In the current paper, we prove that the same attack is also the optimal attack from the adversary's point of view against the GLRT defense for binary classification problems, provide detailed performance analysis and further show that the analysis is asymptotically exact as the number of dimensions grows large. In addition, to characterize GLRT classifier's robustness in multi-class hypothesis testing problems, we provide a procedure to obtain a heuristic based attack that is close to optimal in high SNR regime and illustrate the performance through examples. We also propose a method to

identify an optimal noise-aware attack for multi-class settings. We generalize the formulation of the GLRT defense for colored Gaussian settings. To exemplify the application of GLRT beyond Gaussian settings, we show how the GLRT defense simplifies under Laplacian noise and provide empirical comparisons.

III. Adversarially Robust Classification

In this section, we propose the GLRT based defense against adversarial perturbations and illustrate its application in a Gaussian setting. We also describe a the minimax formulation of the adversarial hypothesis testing problem, developed in [22], and discuss the merits and limitations of the two approaches.

A. GLRT Based Defense

Consider the following standard classification or hypothesis testing problem:

$$\mathcal{H}_k : \mathbf{X} \sim p_k(\mathbf{x}).$$

The presence of an adversary increases the uncertainty about the class-conditional densities, which can be modeled as a composite hypothesis testing problem:

$$\mathcal{H}_k : \mathbf{X} \sim p_\theta(\mathbf{x}), \theta \in \Theta_k,$$

where the size of the uncertainty sets Θ_k depends on the constraints on the adversary. The GLRT defense consists of joint maximum likelihood estimation of the class and the adversary's parameter:

$$\hat{k} = \arg\max_k \max_{\theta \in \Theta_k} p_\theta(\mathbf{x}). \tag{1}$$

The GLRT defense is generic: it can be applied to any model as long as the optimization in (1) can be efficiently performed. In the paper, we focus on settings where the class-conditional densities are Gaussian. To describe the application of GLRT defense in settings beyond the Gaussian distribution, we show how our rule simplifies to an efficient and interesting form under the case where the class-conditional densities are zero-mean Laplace (or the two-sided exponential distribution) in Section VII.

We now apply this framework to Gaussian hypothesis testing with an adversary which can add a deterministic ℓ_∞-bounded perturbation $\mathbf{e}$: $||\mathbf{e}||_\infty \leq \epsilon$, where we term ϵ the "attack budget" or "adversarial budget".

$$\mathcal{H}_k : \mathbf{X} = \boldsymbol{\mu}_k + \mathbf{e} + \mathbf{N},$$

where $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We assume that the adversary has complete access: it knows the true hypothesis, class means, the covariance matrix and could also be aware of the noise realization.

Conditioned on the hypothesis k and the perturbation $\mathbf{e}$, the negative log likelihood is a standard quadratic expression. The GLRT rule for classification in (1) reduces to the following under the Gaussian setting:

$$\hat{k} = \arg\min_k \min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} (\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e}). \tag{2}$$

We view this optimization as a two-step process. We first estimate $\mathbf{e}$ under each hypothesis:

$$\hat{\mathbf{e}}_k = \arg\min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} (\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e}), \tag{3}$$

and then plug in to obtain the cost function to be minimized over k:

$$C_k = (\mathbf{X} - \boldsymbol{\mu}_k - \hat{\mathbf{e}}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}_k - \hat{\mathbf{e}}_k). \tag{4}$$

We remark that estimating the perturbation (3) is a quadratic program which can be solved efficiently, following which the GLRT rule involves only the comparison of the costs C_k under each hypothesis.

Now, let us consider the case where the noise is independent and identically distributed as $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. Under this setting, the estimation of the perturbation of the computation of costs take the form:

$$\hat{\mathbf{e}}_k = \arg\min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} ||\mathbf{X} - \boldsymbol{\mu}_k - \mathbf{e}||^2,$$

$$C_k = ||\mathbf{X} - \boldsymbol{\mu}_k - \hat{\mathbf{e}}_k||^2 \tag{5}$$

This yields illustratively pleasing answers in terms of the function $g_\epsilon(x) \triangleq \text{sign}(x)\max(0, |x| - \epsilon)$, which we term as the "double-sided ReLU" and its "complement," $f_\epsilon(x) = x - g_\epsilon(x)$, shown earlier in Fig. 1. The estimated perturbation under hypothesis k no longer requires solving an optimization problem, but is instead obtained in closed-form as

$$\hat{\mathbf{e}}_k = f_\epsilon\left(\mathbf{X} - \boldsymbol{\mu}_k\right),$$

where the non-linearity is applied coordinate-wise. Substituting into (5), we obtain

$$C_k = ||g_\epsilon\left(\mathbf{X} - \boldsymbol{\mu}_k\right)||^2 \tag{6}$$

where $g_\epsilon(.)$ is applied coordinate-wise. Thus, the GLRT detector

$$\hat{k} = \arg\min_k C_k \tag{7}$$

is a modified version of the standard minimum distance rule where the coordinate-wise differences between the observation and template are passed through the double-sided ReLU.

B. Nature of the Adversary

We consider both noise-aware and noise-agnostic adversarial settings in our paper. When the adversary knows the noise realization $\mathbf{N}$, given a classifier $\hat{\mathcal{H}}$ and the true hypothesis $\mathcal{H}$, the worst-case adversarial attack causes misclassification whenever possible, depending on the noise realization. The noise-aware formulation of the worst-case attack which is optimal from the adversary's point of view is as below:

$$\mathbf{e}^* = \arg\sup_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \mathbb{1}(\hat{\mathcal{H}}(\mathbf{X}) \neq \mathcal{H}(\mathbf{X})). \tag{8}$$

If the adversary does not have access to the noise realization, the optimal attack in the noise-agnostic regime is the maximizer of the class-conditional error, as described below:

$$\mathbf{e}^*_{agn} = \arg\max_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \mathbb{P}\mathrm{r}(\hat{\mathcal{H}}(\mathbf{X}) \neq \mathcal{H}(\mathbf{X})). \tag{9}$$

C. Minimax Approach to Adversarial Robustness

An alternate to the GLRT approach is the minimax formulation of the adversarially robust classification problem, which is game-theoretic. The attack model considered is an additive one, where the adversary observes the sample and adds an ℓ_∞ bounded perturbation. Let $\mathcal{H}$ denote the true hypothesis and $\hat{\mathcal{H}}$ be a classifier. The adversary attempts to cause misclassifications by choosing a suitable perturbation, while the defender tries to choose a classifier such that the expected loss is minimized. The optimum adversarial risk is:

$$R^* = \min_{\hat{\mathcal{H}}} \mathbb{E}\left[\sup_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \mathbb{1}(\hat{\mathcal{H}}(\mathbf{X}) \neq \mathcal{H}(\mathbf{X}))\right], \tag{10}$$

which is solved for the binary Gaussian setting with uniform priors in [22]. Unfortunately, the minimax approach may be overly conservative, unnecessarily compromising performance against attacks that are weaker than, or different from, the worst-case attack. In such scenarios, we expect the GLRT approach, which estimates the attack parameters, to provide an advantage. In an adversarial setting, a defender typically sets the attack budget conservatively, hence achieving good performance even at weaker attacks is of interest.

We now describe the binary Gaussian minimax classifier in [22]. Let the class means be denoted by $+\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ and the covariance matrix by $\boldsymbol{\Sigma}$. It is shown that the minimax optimal robust classifier is a linear classifier with coefficient vector $\mathbf{w} := \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{z}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}))$ that is based on the solution of the following convex program:

$$\boldsymbol{z}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}) = \arg \min_{\mathbf{z}:||\mathbf{z}||_\infty \leq \epsilon} (\boldsymbol{\mu} - \mathbf{z})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{z}), \tag{11}$$

where ϵ is termed the adversarial budget. A closed-form expression for the adversarial risk is provided as:

$$R^* = Q\left(\sqrt{(\boldsymbol{\mu} - \boldsymbol{z}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{z}_{\boldsymbol{\Sigma}}(\boldsymbol{\mu}))}\right). \tag{12}$$

Under white noise, the linear classifier can be interpreted as the inner product between the observation and the signal template after the application of the double-sided ReLU, i.e., $\mathbf{w} = g_\epsilon(\boldsymbol{\mu})$. The optimal noise-aware attack which achieves the optimal adversarial risk is $\mathbf{e} = \pm\epsilon \cdot \text{sign}(\boldsymbol{\mu})$.

While this formulation is noise-aware, a noise-agnostic counterpart that optimizes for the probability of error can be interpreted within the framework of classical robust hypothesis testing [21], albeit with an uncertainty model which does not appear to have been considered before in the literature. Specifically, the set of distributions under each class indexed by the allowable perturbation values that fall within the ℓ_∞ ball. In particular, the Gaussian class-conditional distributions come from an uncertainty set in which the mean is the sum of a nominal mean plus a norm-bounded shift. We note that (see Observation 1) the noise-aware and noise-agnostic worst-case attacks in the binary Gaussian setting and hence the corresponding minimax detectors are the same. It can be shown that the class-conditional distributions with means equal to $\pm g_\epsilon(\boldsymbol{\mu})$ are the least favorable distributions for these uncertainty classes. (We omit the derivation, since our focus here is on the GLRT detector and the noise-aware attack model.)

In order to compare the minimax and GLRT approaches, we next specialize to binary Gaussian hypothesis testing with symmetric means and equal priors to benchmark the performance of GLRT, in a setting where the worst-case attacks of both the classifiers are known. We remark that the analysis developed for the GLRT defense applies in a more general case of unequal priors as well, described subsequently.

IV. Binary Gaussian Hypothesis Testing

Consider the following binary hypothesis testing problem with symmetric means and equal priors for which the minimax rule is known [22]:

$$\mathcal{H}_0 : \mathbf{X} = \boldsymbol{\mu} + \mathbf{e} + \mathbf{N}$$
$$\mathcal{H}_1 : \mathbf{X} = -\boldsymbol{\mu} + \mathbf{e} + \mathbf{N}$$

where $\mathbf{e}$ is deterministic, chosen by an ℓ_∞ bounded adversary, with adversarial budget ϵ, who knows the true hypothesis and the noise realization, with $\mathbf{N} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_d)$. Moreover, since the worst-case attack for the minimax classifier is known under the IID setting, we employ it as a benchmark for the GLRT classifier in this section.

We first describe the GLRT, minimax and the naive minimum distance classifiers under this setting. Next, we analyze the performance of the GLRT defense and show that the analysis is asymptotically exact as the number of dimensions grows large. We then derive the worst-case attack of the GLRT detector and remark on the applicability of our results in settings with asymmetric means.

A. Relation Between Minimum Distance, Minimax and GLRT Rules

We now discuss how the structure of the optimal decision rule without attacks relates to the minimax and GLRT rules. In the absence of attacks, the optimal rule can be expressed as a minimum distance rule as follows:

$$||\mathbf{X} + \boldsymbol{\mu}||^2 \underset{H_1}{\overset{H_0}{\gtrless}} ||\mathbf{X} - \boldsymbol{\mu}||^2. \tag{13}$$

This minimum distance rule can alternatively be expressed as a linear detector which correlates the "signal template" $\boldsymbol{\mu}$ with the observation:

$$\boldsymbol{\mu}^T \mathbf{X} \underset{H_1}{\overset{H_0}{\gtrless}} 0.$$

As shown in [22], the minimax decision rule turns out to be a linear detector of the form

$$g_\epsilon(\boldsymbol{\mu})^T \mathbf{X} \underset{H_1}{\overset{H_0}{\gtrless}} 0.$$

That is, the minimax rule applies the double-sided ReLU to the "signal template" $\boldsymbol{\mu}$, and then performs the correlation. Thus, it simply ignores signal coordinates which are small enough such that their signs could be flipped using the worst-case attack budget of ϵ, and shrinks the remaining coordinates to provide an optimal rule *assuming that the worst-case attack has been applied.*

On the other hand, the GLRT rule in this setting simplifies to a simple modification of the minimum distance rule as following:

$$C_1 = ||g_\epsilon(\mathbf{X}+\boldsymbol{\mu})||^2 \underset{H_1}{\overset{H_0}{\gtrless}} C_0 = ||g_\epsilon(\mathbf{X}-\boldsymbol{\mu})||^2. \tag{14}$$

Comparing (13) and (14), we see that the GLRT rule applies a coordinate-wise double-sided ReLU to the minimum distance (squared) formulation. Since GLRT applies the double-sided ReLU to the difference between the actual observation and signal templates, we expect that, in contrast to the minimax detector, it should be able to adapt if the attack level is lower than the worst-case attack employing the full budget ϵ.

One of the possible worst-case attacks for the minimax classifier is: $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu})$ under $\mathcal{H}_0$ and $\mathbf{e} = \epsilon \cdot \text{sign}(\boldsymbol{\mu})$ under $\mathcal{H}_1$. We prove in Section IV-E that the same attack is indeed the worst-case attack for our GLRT defense under binary classification with Gaussian class-conditionals, for both noise-aware and noise-agnostic adversarial settings.

Under this attack, it is easy to see that the "defenseless" minimum distance detector makes errors with probability at least half whenever the attack budget satisfies $\epsilon > ||\boldsymbol{\mu}||^2/||\boldsymbol{\mu}||_1$. Thus, the system is less vulnerable (i.e., the adversary needs a large attack budget) when the ℓ_1 norm of $\boldsymbol{\mu}$ is small relative to the ℓ_2 norm. That is, signal sparsity helps in robustness, as has been observed before [27], [34].

B. Performance Analysis of the GLRT Defense

Since the GLRT rule is nonlinear, its performance is more difficult to characterize than that of a linear detector. However, we are able to provide insight via a central limit theorem (CLT) based approximation (which holds for large number of dimensions d). Let the priors for the two classes be π_0 and π_1. We perform the following analysis with the attack fixed to $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu})$ (or $\mathbf{e} = +\epsilon \cdot \text{sign}(\boldsymbol{\mu})$), which is the worst-case attack or the optimal perturbation from the attacker's point of view under $\mathcal{H}_0$ (or $\mathcal{H}_1$ respectively), proven in Section IV-E. By the symmetry of the observation model and the resulting symmetry induced on the attack model, the class-conditional errors are equal under both the hypotheses. Hence, we may condition on $\mathcal{H}_0$ and the corresponding worst-case attack $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu})$, and consider $\mathbf{X} = \boldsymbol{\mu} - \epsilon\text{sign}(\boldsymbol{\mu}) + \mathbf{N}$. The costs are

$$C_0 = \sum_{i=1}^{d} (g_\epsilon(-\epsilon\text{sign}(\boldsymbol{\mu}[i]) + \mathbf{N}[i]))^2$$

$$C_1 = \sum_{i=1}^{d} (g_\epsilon(2\boldsymbol{\mu}[i] - \epsilon\text{sign}(\boldsymbol{\mu}[i]) + \mathbf{N}[i]))^2.$$

and the error probability of interest is

$$\begin{aligned} P_e &= \pi_0 P_{e|0} + \pi_1 P_{e|1} = P_{e|0} \\ &= P[C = C_1 - C_0 < 0 | \mathcal{H}_0]. \end{aligned} \tag{15}$$

We now perform a coordinate-wise analysis of the cost difference $C[i] = C_1[i] - C_0[i]$, where $C_k[i]$ indicates the contribution in cost C_k from coordinate i. Let the mean and variance of $C[i]$ be denoted by $m_{C[i]}$ and $\rho^2_{C[i]}$ respectively. Applying CLT on the sum across coordinates, the error probability can be estimated as:

$$P_e = \quad P\left(\sum_{i=1}^{d} C[i] < 0\right) \approx Q\left(\frac{\sum_{i=1}^{d} m_{C[i]}}{\sqrt{\sum_{i=1}^{d} \rho^2_{C[i]}}}\right). \tag{16}$$

The error probability analysis can be made exact in the limit as $d \to \infty$, if the Lindeberg's condition is satisfied for CLT to hold for independent, but not necessarily identically distributed random variables. We show in Section IV-C that Lindeberg's condition is indeed satisfied in our setting.

Asymptotic equivalence with minimax classifier: Consider a particular coordinate i, set $C = C[i]$, and let $\boldsymbol{\mu}[i] = \mu$. Assume $\mu > 0$ without loss of generality: we simply replace μ by $|\mu|$ after performing our analysis, since the analysis is entirely analogous for $\mu < 0$, given the symmetry of the noise and the attack. We can numerically compute the mean and variance of the cost difference for the coordinate, $C = (g_\epsilon(2\mu + N - \epsilon))^2 - (g_\epsilon(N-\epsilon))^2$, but the following lower bound yields insight:

$$C \geq Y \triangleq \mathbb{1}_{\{N \geq -t\}}(t+N)^2 - N^2, \tag{17}$$

where $t = 2(\mu - \epsilon)$. Note that $t > 0$ ($|\mu| > \epsilon$) corresponds to coordinates that the minimax detector would retain. The high-SNR (t/σ large) behavior is interesting. For $t > 0$, we can show that $Y \approx t^2 + 2Nt$; these coordinates exhibit behavior similar to the minimax detector. On the other hand, for $t < 0, Y \approx -N^2$; these coordinates, which would have been deleted by the minimax detector, contribute noise in favor of the incorrect hypothesis (this becomes negligible at high SNR). These observations indicate that, at high SNR, the performance of the GLRT detector approaches that of the minimax detector under worst-case attack.

Without loss of generality, let us redefine $t = 2(|\mu| - \epsilon)$. The mean and variance of Y, irrespective of sign(μ), can be computed in closed form as follows:

$$m_Y = Q\left(\frac{-t}{\sigma}\right)(t^2 + \sigma^2) - \sigma^2 + \sigma t \phi\left(\frac{t}{\sigma}\right) \tag{18}$$

$$\begin{aligned} \rho_Y^2 = {} & 3\sigma^4 + Q\left(\frac{-t}{\sigma}\right)(t^4 + 4t^2\sigma^2 - 3\sigma^4) \\ & + \sigma t \phi(t/\sigma)(t^2 + 3\sigma^2) - m_Y^2. \end{aligned} \tag{19}$$

Fig. 2 shows the empirical mean and empirical variance of $C[i]$, i.e., m_i and ρ_i^2, in comparison with m_Y and ρ_Y^2 obtained through (18) and (19). Here, the adversarial budget is set to $\epsilon = 1$ and noise variance $\sigma^2 = 1$.

GLRT under low noise limit: The error probability in (16) can also be bounded by applying CLT on the lower bounding

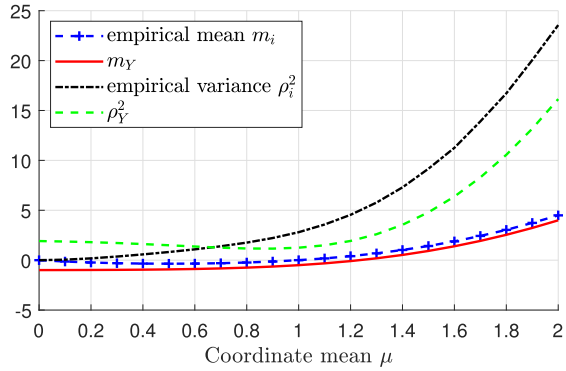

Fig. 2. Comparison of empirical mean and variance of $C[i]$ with the mean and variance of lower bounding variable Y_i.

terms $Y_i \leq C[i]$ as follows:

$$P\left(\sum_{i=1}^{d} C[i] < 0\right) \leq P\left(\sum_{i=1}^{d} Y_i < 0\right) \approx Q\left(\frac{\sum_{i=1}^{d} m_{Y_i}}{\sqrt{\sum_{i=1}^{d} \rho_{Y_i}^2}}\right). \quad (20)$$

Bounding the probability of error in this fashion helps in yielding the following insight. Under low noise limit ($\sigma^2 \to 0$), the variance $\rho_{Y_i}^2 = 0, \forall i$; and the mean is given by $m_{Y_i} = t^2$, if $|\boldsymbol{\mu}[i]| > \epsilon$, otherwise it is zero. Thus as long as $\exists i$ such that $|\boldsymbol{\mu}[i]| > \epsilon$, or equivalently $\epsilon < ||\boldsymbol{\mu}||_\infty$, we have $P_e \to 0$. Interestingly, for the error of the naive minimum distance detector to approach zero under low noise limit, $\epsilon < ||\boldsymbol{\mu}||^2/||\boldsymbol{\mu}||_1$ should hold, which is a more stringent condition than that required by the GLRT detector.

Also note that since each of the means and variances are $\mathcal{O}(1)$ terms, we have $P_e \leq k_1 e^{-k_2 d}$, where k_1, k_2 are positive constants, irrespective of the SNR requirements.

C. Asymptotic Exactness Through Lindeberg's Condition

The random variables Y_k, for $1 \leq k \leq d$, are independent, but not identically distributed. For brevity, let the mean and variance of Y_k be denoted by m_k and ρ_k^2 respectively. The sum of variances of all the d random variables is given by $s_d^2 = \sum_{k=1}^{d} \rho_k^2$. A sufficient condition for the central limit theorem (CLT) to hold in the case of independent but not necessarily identically distributed random variables is the Lindeberg's condition.

Proposition 1: If the independent, non-identically distributed, random variables $Y_k, k \in [d]$, $\forall \delta > 0$ satisfy the following, then the central limit theorem holds.

$$\lim_{d\to\infty} \frac{1}{s_d^2} \sum_{k=1}^{d} \mathbb{E}\left[(Y_k - m_k)^2 \mathbb{1}_{\{|Y_k - m_k| \geq \delta s_d\}}\right] = 0. \quad (21)$$

The proof of this proposition is deferred to the Appendix A. It can further be shown that the Lindeberg's condition is also satisfied by the sum of per coordinate cost differences $C[k]$. Since showing this is analogous, we do not give the detailed case-by-case calculation, but only provide a sketch in Appendix A. Thus, the approximate equalities in (16) and (20) are indeed exact.

D. Application to Unequal Means

The analysis in this paper for the GLRT scheme applies, without loss of generality, to asymmetric means (say $\boldsymbol{\mu_0}$ and $\boldsymbol{\mu_1}$), by shifting of coordinates equivalently, leading to the worst case attack of $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu_0} - \boldsymbol{\mu_1})$ under $\mathcal{H}_0$. The performance analysis also applies in the general case by converting an asymmetric means setting to one with symmetric means by the shift of coordinates, and then due to the symmetry of the attack model and the observation model, the error probabilities are as in (15). We note that the minimax classifier also applies in a setting with generic means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. By change of coordinates, we can arrive at a symmetric mean setting, where if the attacker employs $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu_0} - \boldsymbol{\mu_1})$ under $\mathcal{H}_0$ as the worst-case attack against a linear classifier with $\mathbf{w} = g_\epsilon(\frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{2})$, and from the defender's point of view, fixing the classifier to the minimax scheme is still optimal given such an attack. In general, the minimax classifier takes the form:

$$g_\epsilon\left(\frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{2}\right)^T \left(\boldsymbol{X} - \frac{\boldsymbol{\mu_0} + \boldsymbol{\mu_1}}{2}\right) \underset{H_1}{\overset{H_0}{\gtrless}} 0. \quad (22)$$

E. Worst-Case Attack Under Binary Setting

In this section, we find the optimal attack from the adversary's point of view, also termed the worst-case attack, given a classifier. Firstly, we note that the worst-case attack for the GLRT defense is not unique. In the following proposition, we show that an attack that is oblivious to the noise realization is also a worst-case attack in the noise-aware setting for binary hypothesis testing under the GLRT classifier.

Proposition 2: A worst-case attack for the GLRT defense in a binary Gaussian classification problem with class means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, under both noise-aware and noise-agnostic adversarial settings, is given by

$$\mathbf{e}^* = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1), \text{ under } \mathcal{H}_0 \quad (23)$$

$$\mathbf{e}^* = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0), \text{ under } \mathcal{H}_1. \quad (24)$$

Proof: Without loss of generality, let us first consider the symmetric mean case. Following the notation in Section IV-B, we first show that for all coordinates where $\mu \geq 0$, the per coordinate cost difference under $\mathcal{H}_0$, given by $C[i]$, is non-decreasing in $\mathbf{e}[i]$, and where $\mu < 0$, $C[i]$ is decreasing in $\mathbf{e}[i]$. The proof is deferred to Appendix B. Let $\mathbf{e}_1$ and $\mathbf{e}_2$ be two attacks and $\mathbf{N}$ a noise realization. Denoting $\sum_i C[i] = \underline{C}$ for brevity, and assuming that $\mu > 0$ for all the coordinates, it follows from the monotonicity of per-coordinate cost difference, that for any fixed $\mathbf{N}$, if $\mathbf{e}_1 \succcurlyeq \mathbf{e}_2$, then $\underline{C}(\mathbf{e}_1, \mathbf{N}) \geq \underline{C}(\mathbf{e}_2, \mathbf{N})$. Let $\mathbf{e}_2 = -\epsilon \cdot \underline{\mathbf{1}}$. For any other attack $\mathbf{e}_1$, and $\forall \mathbf{N}$,

$$\underline{C}(\mathbf{e}_1, \mathbf{N}) \geq \underline{C}(-\epsilon \cdot \underline{\mathbf{1}}, \mathbf{N}). \quad (25)$$

Relaxing the assumption on μ and utilizing the result that the per coordinate cost difference for indices where $\mu < 0$ is decreasing

in corresponding $\mathbf{e}[i]$, it follows that

$$\underline{C}(\mathbf{e}, \mathbf{N}) \geq \underline{C}(-\epsilon \cdot \text{sign}(\boldsymbol{\mu}), \mathbf{N}) \tag{26}$$

for any $\mathbf{e}$ and $\forall \mathbf{N}$. For binary classification, it suffices that the per-coordinate cost difference is negative to cause misclassification. Under $\hat{\mathcal{H}}_0$, we can see from (26) that the attack $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu})$ is sufficient to cause misclassification, whenever possible, for any noise realization. Extending to generic means by shift of coordinates, the worst case attack under $\mathcal{H}_0$ is thus given by $\mathbf{e}^* = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$. Since the attack does not utilize the noise realization, it is also the best a noise-agnostic adversary can do. ■

Observation 1: The same attack $\mathbf{e}^* = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ is the worst-case attack in the presence of both noise-agnostic and noise-aware adversaries under binary settings for minimax and also the naive minimum distance based classifier.

For minimum distance classifier, the cost of choosing hypothesis $\mathcal{H}_i$ is $C_i = ||\mathbf{X} - \boldsymbol{\mu}_i||^2$. For binary problems, if a noise-aware adversary wants to cause misclassification, it requires to pick a perturbation such that under $\mathcal{H}_0$, costs are such that $C_1 < C_0$, which reduces to finding a perturbation such that given noise $\mathbf{N}$,

$$\min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T (\mathbf{e} + \mathbf{N}).$$

However, irrespective of noise, the perturbation $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ minimizes the above, due to which it is an optimal attack when the minimum distance classifier is used by a defender. This is also optimal for an agnostic adversary as the attack does not require the knowledge of noise. It is also shown in detail in *Observation* 2.

Similarly, for the minimax classifier, from (22) it can be deduced that the adversary attempts to perform

$$\min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} g_\epsilon \left(\frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{2} \right)^T (\mathbf{e} + \mathbf{N})$$

under $\mathcal{H}_0$, leading to $\mathbf{e} = -\epsilon \cdot \text{sign}(g_\epsilon(\frac{\boldsymbol{\mu_0} - \boldsymbol{\mu_1}}{2}))$, which is equivalent to $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$. Thus the same attack is optimal for the three classifiers under binary setting.

V. Binary Examples and Discussion

Focusing on binary classification problems with symmetric means in this section, we consider settings where a fraction p of the coordinates have means $\mu = a\epsilon$ and a fraction $(1-p)$ have $\mu = b\epsilon$, where $a > 1$ and $0 \leq b \leq 1$. We consider such class means for ease of representation and remark that this is typical of data employed in learning applications, which is sparse, leading to a small fraction of high-valued coordinates and a large fraction of coordinates closer to zero. Let the designed adversarial budget be ϵ and the actual attack be $\mathbf{e} = \mp\kappa \text{sign}(\boldsymbol{\mu})$, where $0 \leq \kappa \leq \epsilon$. Under this setting, it is simple to approximate (or describe exactly in the case of minimax classifier) the error through computation of the effective signal-to-noise ratio (SNR).

For the minimax scheme, note that only the fraction p of the coordinates contribute to signal energy. The decision statistic is $(g_\epsilon(\boldsymbol{\mu}))^T \boldsymbol{X}$, from which it follows that the effective SNR is:

$$\text{SNR}_{\text{minimax}} = (a-k)^2 dp \left(\frac{\epsilon}{\sigma}\right)^2.$$

For the GLRT scheme, the SNR can be obtained directly from (16) as

$$\text{SNR}_{\text{GLRT}} \approxeq d \frac{(pm_a + (1-p)m_b)^2}{p\rho_a^2 + (1-p)\rho_b^2},$$

where m_a and m_b are means, ρ_a^2 and ρ_b^2 are variances of a single component of $C[i]$ contributed by terms with component means $a\epsilon$ and $b\epsilon$ respectively. The probability of error for both the classifiers is given by $Q(\sqrt{\text{SNR}})$. Note that for the GLRT detector, it is only an approximation as convergence is slow at high SNR, and we need to rely on simulations for more accurate error probabilities. With the above setting in mind, we will now illustrate the performance of the GLRT defense. We also remark that in the binary IID setting, since the worst-case noise-agnostic and noise-aware attacks are identical for GLRT, minimax and the minimum distance rules, we do not provide a distinct treatment of the aware and agnostic scenarios in this section.

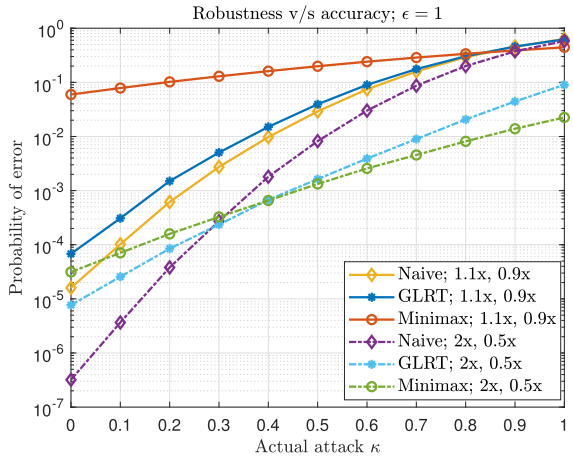

Fig. 3. Robustness v/s accuracy trade-off as the actual attack is varied, while the designed adversarial budget is fixed to $\epsilon = 1$.

A. Trade-Off Between Robustness and Accuracy

We consider binary classification problems with symmetric means and uniform priors to draw a comparison with the minimax optimal scheme, and also a naive minimum distance classifier that is optimal under zero attack. The GLRT detector performs better than minimax for weaker attacks, and it has a significant advantage over minimax in settings where the class mean $\boldsymbol{\mu}$ has components which are smaller than ϵ, but larger than the actual attack. GLRT utilizes signal energy from these components while for minimax, such components are nulled. Fig. 3 depicts the performance advantage of GLRT under weaker attacks, for a problem with parameters $\epsilon = 1$, $d = 20$, $p = 0.1$, $a = 1.1$, $b = 0.9$ and noise variance $\sigma^2 = 1$.

The naive minimum distance classifier does poorly under a large attack, specifically in settings where $\boldsymbol{\mu}$ has a large number of small components. Under strong attacks, these smaller

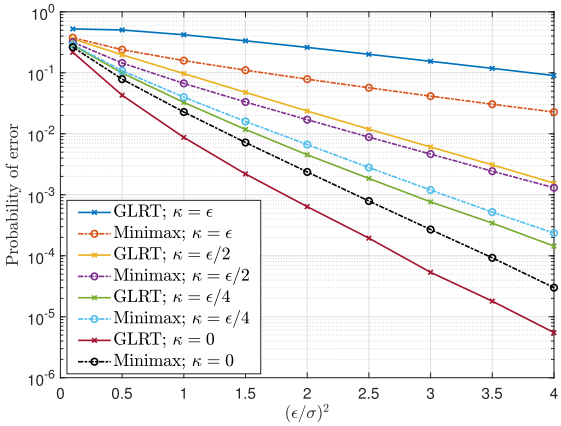

Fig. 4. Probability of error as a function of $(\epsilon/\sigma)^2$ for different values κ of actual attack (with $\epsilon = 1$, $a = 2$, $b = 0.5$).

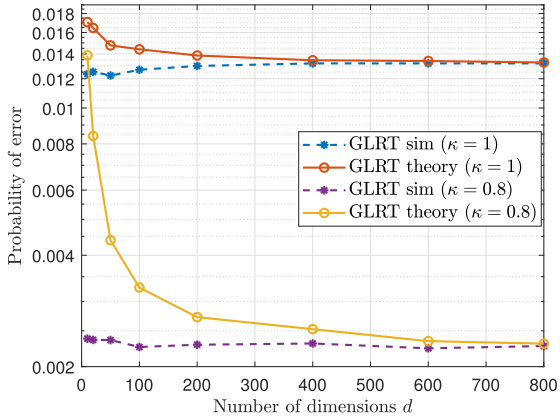

Fig. 5. Asymptotic convergence of error predicted from CLT approximation to the simulation performance.

components contribute to costs in such a way that the wrong class is favored by the naive detector. Consider a problem with parameters $d = 10$, $p = 0.1$, $\epsilon = 1$, $a = 2$, $b = 0.5$ and $\sigma^2 = 0.25$. The comparison of all three detectors under this setting is plotted in Fig. 3, which clearly indicates the failure of naive scheme at high attacks, emphasizing the need for a robust detector. Fig. 4 shows the variation of the error probability as a function of $(\epsilon/\sigma)^2$, under four different values of actual attack, for the same problem setting.

In summary, the GLRT defense serves as a graceful intermediate between the naive minimum distance rule which fails under a large attack, and the minimax classifier which is too pessimistic if the actual attack is weaker than the one which the classifier is designed for. The margin of improvement which the GLRT defense provides over the other classifiers depends on the parameters of the setting, as demonstrated by the two set of parameters that we consider in Fig. 3. In situations where the defender does not possess an exact knowledge of the adversary's budget, the defender would make an informed guess about it and design the classifier pessimistically. Hence, considering the performance of the designed classifier at weaker attacks becomes important.

B. Speed of Convergence

Using CLT to approximate the error probability of the GLRT defense holds only in the limiting case of large d. We observe that the distributions of per-coordinate cost differences for each of the coordinates, specifically under low noise, could have narrow asymmetric tails, due to which convergence is slow. We consider a setting with parameters $a = 1.1$, $b = 0.9$, $p = 0.3$, $\epsilon = 1$ and compare the error probabilities as indicated by simulation and those calculated from (16), where the means and variances of each coordinate of $C[i]$ are computed empirically. We consider two attacks of the form $\mathbf{e} = -\kappa \cdot \text{sign}(\boldsymbol{\mu})$, one with the full strength of attack $\kappa = \epsilon = 1$, and another, a weaker attack $\kappa = 0.8$. For both these settings, the true error is fixed to two different values $P_{err_1} \approx Q(\sqrt{5})$ and $P_{err_2} \approx Q(\sqrt{8})$ respectively. Since error is a smooth function of noise variance, the value of σ^2 for a particular d and the fixed P_{err} is found through grid search. As expected, Fig. 5 shows that the theoretical performance approaches that of the simulation as the number of dimensions grows.

C. Colored Noise Setting

In this subsection, we empirically compare the GLRT and minimax classifiers under non-spherical covariance matrix settings. Previously, we have illustrated the benefit of GLRT defense under weak attacks in the white Gaussian settings where its worst-case attack is known. Since, the worst-case attack for the GLRT classifier under colored Gaussian noise is unknown, comparing the performance of minimax and GLRT for a fixed attack direction with variation in the strength of the attack, as done in Fig. 3 is not meaningful. Instead, we compare the adversarial risks (10) of the two classifiers.

Given a binary setting with equal priors, symmetric means and covariance matrix $\boldsymbol{\Sigma}$, it is shown in [22] that the adversarial risk can be calculated in closed-form from (12). Given hypothesis $\mathcal{H}$, the adversarial risk of the GLRT defense is:

$$R^{GLRT} = \mathbb{E}\left[\sup_{\mathbf{e}:\|\mathbf{e}\|_\infty \le \epsilon} \mathbb{1}(\hat{\mathcal{H}}_{GLRT}(\mathbf{X}) \neq \mathcal{H}(\mathbf{X}))\right]. \quad (27)$$

Since the worst-case attack is not characterized, the adversarial risk can only be computed empirically. One needs to check, for each noise realization seen, if there exists a perturbation $\mathbf{e}$ within the ℓ_∞ ball of radius ϵ that can cause a misclassification. For the GLRT rule, this implies checking for a perturbation that results in a smaller cost for an incorrect hypothesis. For the binary setting, under $\mathcal{H}_0$, we have $\mathbf{X} = \boldsymbol{\mu}_0 + \mathbf{e} + \mathbf{N}$, and the problem of finding a feasible attack reduces to the following optimization

$$\min_{\mathbf{e}:\|\mathbf{e}\|_\infty \le \epsilon} C_1 - C_0, \quad (28)$$

and checking if the minimizer indeed makes $C_1 < C_0$. However, note that in the above objective function, cost C_k depends on parameter $\hat{\mathbf{e}}_k$, which in turn depends on the given attack $\mathbf{e}$ through the observation $\mathbf{X}$ (see (3)). This is potentially a non-convex problem, which we solve by numerical optimization in our experiments. This leads to a lower bound on the adversarial risk of the GLRT classifier since the numerical search does not

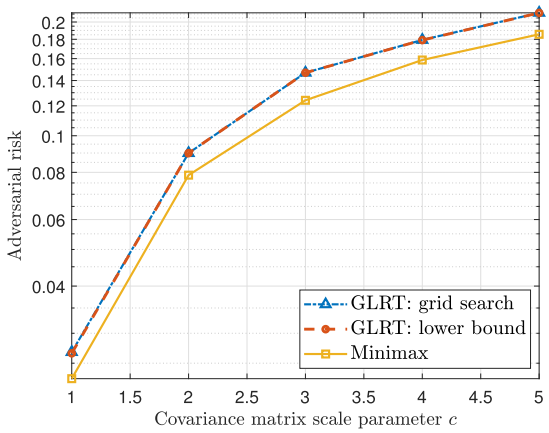

Fig. 6. Adversarial risk of the GLRT and minimax classifiers under colored noise setting.

guarantee arriving at a feasible attack if it exists. We observe in our simulations that the lower bound obtained as such is close to the exact adversarial risk in the lower dimensional settings. We also empirically compute the adversarial risk of GLRT through a grid search and compare with the lower bound obtained. Fig. 6 shows the bound and exact adversarial risk of the GLRT classifier and the optimal adversarial risk of the minimax classifier, for a binary setting with parameters $d = 2$, $\boldsymbol{\mu}_0 = [2, 0.5]$, $\boldsymbol{\mu}_1 = [-2, -0.5]$, designed adversarial budget of $\epsilon = 1$, and covariance matrix

$$\boldsymbol{\Sigma} = c \cdot \begin{bmatrix} 0.25 & 0.35 \\ 0.35 & 0.75 \end{bmatrix},$$

where c is a scalar multiplier which is varied. Note that the minimax classifier, by definition, achieves the minimum adversarial risk and that of the GLRT classifier is comparable with the optimal. In addition, the GLRT is expected to possess a better robustness-accuracy trade-off when a weaker attack is employed, as previously demonstrated.

VI. Multi-Class Gaussian Hypothesis Testing

The GLRT defense applies to multi-class setting with generic means and priors naturally, as described in (7). In order to benchmark the performance of GLRT, we do the following:

1) We first note that deriving a minimax optimal classifier in multi-class setting is a difficult problem, even with the assumption of uniform priors. We consider a heuristic-based extension of the binary minimax classifier, termed the *Pairwise Robust Linear* (PRL) classifier, which we employ to benchmark the performance of GLRT, along with comparing it with the minimum distance classifier.
2) We illustrate that finding an optimal noise-agnostic attack in the multi-class setting is a difficult problem, and provide a heuristic attack, that is close to the optimal noise-agnostic attack in the high SNR regime, by obtaining a procedure to identify the neighboring class which contributes the most to errors.
3) We provide a simple method to identify the optimal noise-aware attack in multi-class problems by extending our knowledge about the optimal noise-aware attack in binary setting. This also gives a lower bound on the classifier's performance.

We begin by first describing the Pairwise Robust Linear classifier that we propose to provide a benchmark for the performance of GLRT defense in multi-class settings. Given an M-ary classification problem, we can form $\binom{M}{2}$ pairs of binary minimax classifiers. The observation is classified as belonging to class k if k is a clear winner in all $M-1$ binary tests $\mathcal{H}_k$ v/s $\mathcal{H}_i$, $i \neq k$, else it is considered an error. We term this as the PRL classifier, since the binary minimax classifier is linear. Note that the PRL classifier need not be minimax optimal. Instead of requiring that a particular class wins against all others, one could also make a decision based on the majority winner among all classes, but for simplicity, we restrict ourselves to requiring a clear winner against all other hypotheses.

Next, we will describe the procedure to find a near-optimal attack when the adversary is agnostic to noise, and later provide a way to find an optimal noise-aware attack in multi-class settings.

A. Noise-Agnostic Attacks

The optimal noise-agnostic attack, described in (9), is difficult to obtain in closed-form for non-binary settings due to the complicated geometry of the problem. However, our approach to find a near-optimal noise-agnostic attack is the following. Given M classes, the class conditional error is upper bounded by the sum of errors of pairwise binary hypothesis tests, and at high SNR, we can assume that there is a single competing class that dominates the error calculations. Thus we are interested in finding this competing *nearest neighbor* class. Given M classes, under $\mathcal{H}_j$, one can think of $M-1$ binary classification problems $\mathcal{H}_j$ v/s $\mathcal{H}_i$, where $i \neq j$, and find which of these binary hypothesis tests yields the worst probability of error. At high SNR, the class conditional error for M-ary hypothesis testing depends primarily on the worst of the $M-1$ binary hypothesis tests. We term the competing class which yields this worst probability of error as the *nearest neighbor* class. As a proxy for the true worst-case attack in the multi-class setting, one can use the worst-case attack of the binary hypothesis test against the nearest neighbor (NN) class. Therefore, we want to find the NN class, under every hypothesis.

We provide procedures to identify the NN class under the minimum distance, GLRT and PRL classifiers through the below series of observations, which we substantiate in the rest of this section, culminating in Observation 4 which identifies the NN class for the GLRT classifier in the most generic setting.

Observation 2: The NN class under hypothesis $\mathcal{H}_j$ for the minimum distance classifier is:

$$\hat{k}(j) = \arg\min_k ||\boldsymbol{\mu}_{jk}|| - \epsilon \frac{||\boldsymbol{\mu}_{jk}||_1}{||\boldsymbol{\mu}_{jk}||}, \tag{29}$$

where $\boldsymbol{\mu}_{jk} = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_k)/2$.

We substantiate the above observation as follows. Let us first consider the minimum distance classifier under binary,

symmetric means setting. Under $\mathcal{H}_0$, we have $\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{e} + \boldsymbol{N}$, and the linear classifier of the form $\mathbf{w}_{clean} = \boldsymbol{\mu}$ as discussed earlier. The class conditional error simplifies as follows:

$$\begin{aligned} P_{e|H_0} &= P(\boldsymbol{\mu}^T \boldsymbol{X} < 0) \\ &= P(||\boldsymbol{\mu}||^2 + \boldsymbol{\mu}^T \mathbf{e} + \boldsymbol{\mu}^T \boldsymbol{N} < 0) \\ &= Q\left(\frac{\boldsymbol{\mu}^T \mathbf{e}/||\boldsymbol{\mu}|| + ||\boldsymbol{\mu}||}{\sigma}\right) \end{aligned}$$

and the worst case noise-agnostic attack is:

$$\mathbf{e}^*_{agn} = \arg \min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \boldsymbol{\mu}^T \mathbf{e}/||\boldsymbol{\mu}|| + ||\boldsymbol{\mu}||.$$

Through Holder's inequality, we have $\boldsymbol{\mu}^T \mathbf{e} \geq -||\mathbf{e}||_\infty ||\boldsymbol{\mu}||_1 \geq -\epsilon ||\boldsymbol{\mu}||_1$, and equality is achieved when $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu})$. Thus, the error corresponding to the worst attack is of the form

$$P_{e|H_0} = Q\left(\frac{||\boldsymbol{\mu}|| - \epsilon ||\boldsymbol{\mu}||_1/||\boldsymbol{\mu}||}{\sigma}\right).$$

In the case of asymmetric means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, it follows that the worst-case error is:

$$\begin{aligned} P_{e|H_0} &= Q\left(\frac{||\frac{\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1}{2}|| - \epsilon ||\frac{\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1}{2}||_1/||\frac{\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1}{2}||}{\sigma}\right) \\ &= Q\left(\frac{||\boldsymbol{\mu}_{01}|| - \epsilon ||\boldsymbol{\mu}_{01}||_1/||\boldsymbol{\mu}_{01}||}{\sigma}\right), \end{aligned} \tag{30}$$

where we denote $(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)/2$ as $\boldsymbol{\mu}_{01}$. The optimal noise-agnostic attack in binary setting is hence $\mathbf{e}^*_{agn} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$, under $\mathcal{H}_0$. Generalizing to a multi-class setting under hypothesis $\mathcal{H}_j$, under high SNR the error probability would be dominated by the binary hypothesis test between class j and another *closest neighbor* class. A heuristic way of proposing an agnostic attack that is close to the optimal agnostic attack is to simply attack that class which contributes the most error. Using (30) and the fact that $Q(.)$ is a monotonically decreasing function, the binary test that contributes the largest error is against the class determined in (29), which is termed as the nearest neighbor class.

This equation also captures that sparsity improves robustness of the naive classifier. For a system with fixed ℓ_2 norm of the pairwise separation between means $\boldsymbol{\mu}_{jk}$, the class with greater ℓ_1 norm corresponds to the NN class.

Observation 3: A procedure to identify the NN class for GLRT under hypothesis $\mathcal{H}_j$ is:

$$\hat{k}(j) = \arg \min_k \sum_{i:|\boldsymbol{\mu}_{jk}[i]| \geq \epsilon} (|\boldsymbol{\mu}_{jk}[i]| - \epsilon)^2 \tag{31}$$

and the attack under hypothesis $\mathcal{H}_j$, that is close to the optimal noise-agnostic attack is $\mathbf{e}_{agn} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_{j\hat{k}(j)})$. As SNR increases, $\mathbf{e}_{agn}$ approaches $\mathbf{e}^*_{agn}$.

The above is demonstrated as follows. From (20), the class conditional error can be upper bounded by using CLT on the bounding variables $Y[i]$. Observe that in the high-SNR regime, the bound in (17) is close to equality. Thus the true probability of error can be approximated as the error found through CLT on the bounding variables. Suppose that a fraction p of the coordinates are such that $|\boldsymbol{\mu}_{0\,k}[i]| \geq \epsilon$. From (17) and following the notation from Section IV-B, the mean and variance of $Y[i]$ can be verified to be the following, based on $t_i = 2(\boldsymbol{\mu}_{0\,k}[i] - \epsilon)$ being positive or negative.

$$m_{Y_i} = \begin{cases} t_i^2 & \text{if } t > 0 \\ -\sigma^2 & \text{if } t < 0 \end{cases}$$

$$\rho^2_{Y_i} = \begin{cases} 4\sigma^2 t_i^2 & \text{if } t > 0 \\ 2\sigma^4 & \text{if } t < 0 \end{cases}$$

The error is estimated as the following under high SNR:

$$\begin{aligned} P_{e|H_0} &\approx Q\left(\frac{\sum_{i=1}^d m_{Y_i}}{\sqrt{\sum_{i=1}^d \rho^2_{Y_i}}}\right) \\ &= Q\left(\frac{-(1-p)d\sigma^2 + \sum_{i=1}^{pd} 4(|\boldsymbol{\mu}_{0\,k}[i]| - \epsilon)^2}{\sqrt{2(1-p)d\sigma^4 + 4\sigma^2 \sum_{i=1}^{pd} 4(|\boldsymbol{\mu}_{0\,k}[i]| - \epsilon)^2}}\right) \\ &\approx Q\left(\frac{1}{2\sigma}\sqrt{\sum_{i:|\boldsymbol{\mu}_{0\,k}[i]| \geq \epsilon} 4(|\boldsymbol{\mu}_{0\,k}[i]| - \epsilon)^2}\right). \end{aligned}$$

Thus, it follows that we can identify the NN class for GLRT under hypothesis $\mathcal{H}_j$ as in Observation 3, and further simplify as:

$$\begin{aligned} \hat{k}(j) &= \arg \min_k \sum_{i:|\boldsymbol{\mu}_{jk}[i]| \geq \epsilon} (|\boldsymbol{\mu}_{jk}[i]| - \epsilon)^2 \\ &= \arg \min_k \sum_{i:|\boldsymbol{\mu}_{jk}[i]| \geq \epsilon} (\boldsymbol{\mu}_{jk}[i])^2 - 2\epsilon(|\boldsymbol{\mu}_{jk}[i]|) \\ &= \arg \min_k ||g_\epsilon(\boldsymbol{\mu}_{jk})||^2 \end{aligned} \tag{32}$$

and the attack is $\mathbf{e}_{agn} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_{j\hat{k}(j)})$.

It is interesting to note that the same coordinates would have been retained by the PRL classifier under the high SNR regime, leading to the same NN class. Note that the analysis above implicitly assumes that the attack utilizes the entire adversarial budget. If the actual attack is a weaker attack of the form $\mathbf{e} = \mp\kappa \cdot \text{sign}(\boldsymbol{\mu})$, where $\kappa < \epsilon$, it follows, analogous to (17), that

$$\begin{aligned} C[i] &= (g_\epsilon(2\boldsymbol{\mu}[i] + \boldsymbol{N}[i] - \kappa))^2 - (g_\epsilon(\boldsymbol{N}[i] - \kappa))^2 \\ &\geq \mathbb{1}_{\{\boldsymbol{N}[i] \geq -t_i\}}(t_i + \boldsymbol{N}[i])^2 - (\boldsymbol{N}[i])^2 \\ &\triangleq Y_i \end{aligned} \tag{33}$$

where we redefine $t_i = 2\boldsymbol{\mu}[i] - \kappa - \epsilon$.

Observation 4: Under attacks that are weaker than the designed budget, the NN class for GLRT is given by

$$\hat{k}(j) = \arg \min_k \sum_{i:2|\boldsymbol{\mu}_{jk}[i]| \geq \kappa + \epsilon} (2|\boldsymbol{\mu}_{jk}[i]| - \kappa - \epsilon)^2. \tag{34}$$

It is interesting to note that method for determining the NN class derived here implicitly depends on the sparsity of separation between the pairwise means $\boldsymbol{\mu}_{jk}$, but measured only over the surviving coordinates as expressed by (32).

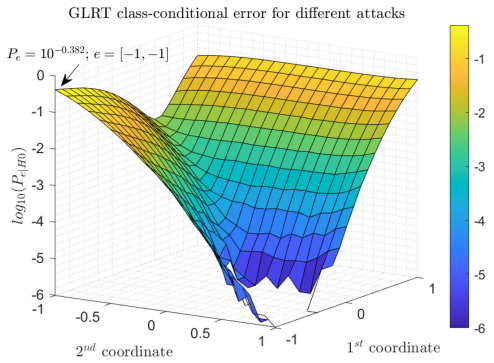

Fig. 7. Error surface for the ternary GLRT classifier and its worst case attack.

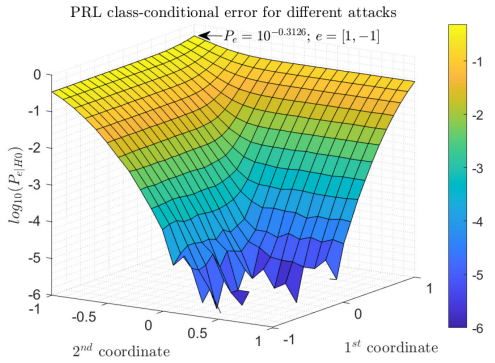

Fig. 8. Error surface for the ternary pairwise robust linear classifier and its worst case attack.

B. Noise-Aware Attacks

When the true class and the noise realization is known, the attacker aims to employ the worst attack that causes misclassification, when possible. Note that given the noise realization and true label, it is not computationally hard to find the worst-case attack. The adversary checks for feasibility, i.e., whether there exists a perturbation that can cause misclassification. However, rather than compute such an attack numerically, we provide a simple procedure to identify the optimal noise-aware attack in multi-class settings, that builds upon the optimal noise-aware attack in binary settings, identified in *Observation* 1.

We will now address the question of finding optimal noise-aware attack for $M-$ary classification problems. Recall that the adversary needs to find a perturbation $\mathbf{e} : ||\mathbf{e}||_\infty \leq \epsilon$ such that under true hypothesis $\mathcal{H}_i$, the cost for the classifier under some incorrect hypothesis $\mathcal{H}_j$, (given by (6) for the GLRT classifier), is smaller than the cost under the true hypothesis $\mathcal{H}_i$. Since the adversary knows the true class and the noise realization, it can compute and check if by employing any of the $M-1$ binary optimal noise-aware attacks for $\mathcal{H}_i$ v/s $\mathcal{H}_j$, $j \neq i$, if the resulting costs are such that $C_j < C_i$, for some j. If there exists such a class j, then $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ is sufficient to cause misclassification. If such j is not found, then it implies that none of the incorrect class costs can be made small enough, which means it is not possible to cause misclassification in the multiclass setting. Thus using this procedure for every realization of noise seen, the adversary can behave optimally in the noise-aware multiclass setting. The costs under hypotheses are clear for the minimum distance classifier. Since the PRL is an extension of binary minimax classifier, adversary can attack so as to cause at least one of the binary minimax hypothesis tests to fail. Thus the costs for minimax classifier can be used in this procedure to find an optimal noise-aware attack for PRL classifier.

C. Ternary Classification Examples

Let us consider a simple two-dimensional ternary classification problem with parameters $\boldsymbol{\mu}_0 = [0, 0]$, $\boldsymbol{\mu}_1 = [2.5, 0.25]$, $\boldsymbol{\mu}_2 = [-1.75, -2.25]$ and $\sigma^2 = 0.1$, and empirically explore the variation of class-conditional error as a function of the attack, for all valid attacks. Figs. 7 and 8 illustrate the class-conditional error under true class is $\mathcal{H}_0$, for GLRT and PRL classifier respectively. The error surface and the direction of the optimal noise-agnostic attack is different for these classifiers, as indicated in the figures. We also observe that the error surface for GLRT drops faster, for the considered example, in comparison to the error of PRL classifier. The NN class under $\mathcal{H}_0$ as suggested by (31) is $\mathcal{H}_2$, and the corresponding attack $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_2) = [-1, -1]$, which agrees with $\mathbf{e}^*_{agn}$ as seen in Fig. 7. We also checked empirically that the same attack leads to the worst class-conditional error for the PRL classifier, albeit not at the noise level considered in Fig. 8, but at a higher SNR, for the same problem. Though it is simple in a two-dimensional setting to empirically verify that the optimal noise agnostic attack at a particular SNR concurs with the attack suggested by employing NN class calculations, for a large dimensional problem, it is hard to know if the noise variance is low enough for the heuristics to hold, and the optimal noise-agnostic attacks for GLRT, PRL and minimum distance classifiers could be different.

We now consider a ternary setting with equi-probable classes, and parameters $d = 20$, $\epsilon = 1$, noise variance $\sigma^2 = 0.1$, class mean $\boldsymbol{\mu}_0$ such that the first $p_0 = 0.15$ fraction of the coordinates are at 0, and the rest at 1, i.e., $\boldsymbol{\mu}_0 = [0, 0, 0, 1, \ldots, 1]$, $\boldsymbol{\mu}_1$ such that the first $p_1 = 0.1$ fraction of the coordinates at -2.1 and rest at 0.9 ($\boldsymbol{\mu}_1 = [-2.1, -2.1, 0.9, \ldots, 0.9]$), and $\boldsymbol{\mu}_2$ such that the first $p_2 = 0.2$ fraction of the coordinates at -1.8 and rest at 1.75 ($\boldsymbol{\mu}_2 = [-1.8, -1.8, -1.8, -1.8, 1.75, \ldots, 1.75]$). The strength of the attack is is varied such that $0 \leq \kappa \leq \epsilon = 1$. These mean parameters, though seemingly arbitrary, have been chosen such that the pairwise difference between the means possess (i) large number of small components (ii) some components that are smaller the designed budget, but larger than actual attack strength employed. Recall from Section V that settings with these properties showcase simultaneously the quick deterioration of minimum distance classifier as attack strength increases and superiority of GLRT over minimax for weak attacks. This can be observed in Fig. 9, which shows the error probabilities (or error frequencies) of GLRT, PRL and minimum distance classifiers, for both noise-agnostic and noise-aware adversaries. For each of the classifiers, their respective noise-aware optimal

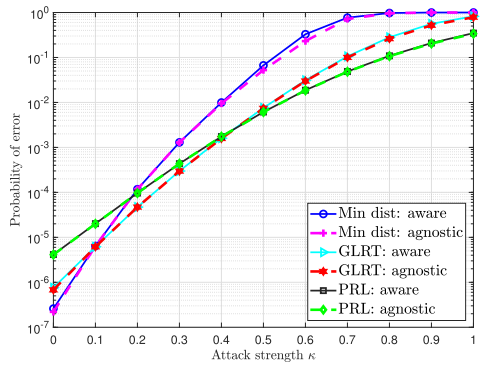

Fig. 9. Performance of GLRT, PRL and minimum distance classifiers for the ternary classification problem considered.

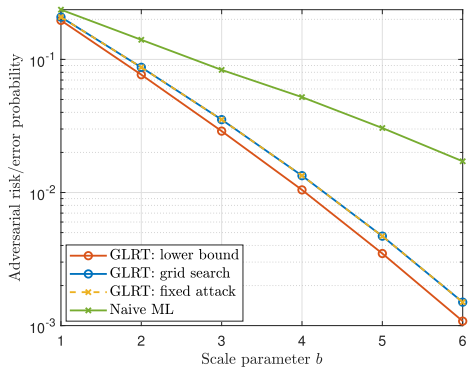

Fig. 10. Performance of GLRT defense under Laplacian noise in comparison with a non-robust ML classifier.

attacks are employed to obtain the performance of the classifiers under noise-aware settings. In the case of performance under noise-agnostic adversaries, at each value of the attack strength, the direction of attack is chosen based on the heuristic noise-agnostic attack for each of the classifiers by identifying their respective NN classes (for GLRT, PRL as per (34), and for minimum distance classifier as per (29)). Due to the complex geometry of the problem in multi-class settings, it is not easy to provide insights on how better a noise-aware adversary could do in comparison with an agnostic one, but the example considered here serves the purpose of showing a case where the performance under aware and agnostic adversaries is not too different. In general the gap between the performance depends on the separation between class mean parameters and noise variance. The plots for optimal noise-aware attacks also give a lower bound on the performance of the respective classifiers for adversarial hypothesis testing.

Note that the noise variance is small ($\sigma^2 = 0.1$). At such relatively high SNR, we expect that each class-conditional error will be dominated by misclassifications occurring due to its nearest neighbor class. Hence the optimal attack in the multi-class setting would essentially be *close* to the binary worst-case attack corresponding to the nearest neighbor class. The aware and agnostic attacks are the same in binary setting, due to which at high SNR, the performance of aware of agnostic adversaries in multi-class setting is close, and the empirical results in Fig. 9 support the same. As the noise variance grows larger, one will observe a larger gap in the performance of noise-aware and noise-agnostic adversaries.

VII. GLRT Defense Beyond Gaussianity

While we have focused on the Gaussian setting for which an adversarially robust minimax rule is known, in this section, we illustrate the application of GLRT defense beyond the Gaussian setting. Consider that the noise is distributed as an independent and identically distributed zero-mean Laplace random variable $\mathbf{N}[i] \sim Laplace(0, b)$, where the zero-mean Laplace density is given by

$$p(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

Under this setting, the GLRT rule in (1) simplifies to

$$\hat{k} = \arg\min_{k} \min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \sum_{i=1}^{d} |\mathbf{X}[\mathbf{i}] - \boldsymbol{\mu}_k[i] - \mathbf{e}[i]|. \tag{35}$$

Breaking the above optimization, the estimation of the perturbation is essentially

$$\hat{\mathbf{e}}_k = \arg \min_{\mathbf{e}:||\mathbf{e}||_\infty \leq \epsilon} \sum_{i=1}^{d} |\mathbf{X}[\mathbf{i}] - \boldsymbol{\mu}_k[i] - \mathbf{e}[i]|,$$

which reduces to $\hat{\mathbf{e}}_k = f_\epsilon(\mathbf{X} - \boldsymbol{\mu}_k)$, just as in the Gaussian setting, and the cost under each hypothesis now takes the form $C_k = ||g_\epsilon(\mathbf{X} - \boldsymbol{\mu}_k)||_1$, analogous to (6).

In the absence of any perturbation, the maximum likelihood detection rule under the Laplacian noise would result in the minimization of the cost $C_k^{ML} = ||\mathbf{X} - \boldsymbol{\mu}_k||_1$. In comparison, the adversarially robust GLRT classifier requires only an additional step of the simple coordinate-wise application of the double-sided ReLU, like in the Gaussian setting.

The adversarial risk can be obtained by a feasibility check for causing misclassifications. Given an adversarial budget of ϵ, for each noise realization seen, one can check if there exists a perturbation within the ℓ_∞ ball of radius ϵ, that can result in a smaller cost for an incorrect hypothesis for the GLRT classifier, thus causing a misclassification. For the binary setting, under $\mathcal{H}_0$, the problem of finding a feasible attack as in (28) is solved by numerical optimization to obtain a lower bound on the adversarial risk of the GLRT classifier under Laplacian noise.

Alternately, the adversarial risk can be inferred by finding the optimal attack from the adversary's point of view, and the error probability under that worst-case attack is the adversarial risk. We conjecture that for this setting as well, the attack $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ is indeed the worst-case attack. To support the same, we find the maximum adversarial risk in a smaller dimensional setting by a grid search for a perturbation that causes misclassification, under each noise realization seen, and we observe that the error probability obtained through grid search coincides with that when the attack is fixed to $\mathbf{e} = -\epsilon \cdot \text{sign}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$. For comprehensive comparison, we also evaluate through simulations, the performance of a non-robust detector in this setting, such as the maximum-likelihood rule

which minimizes the ℓ_1 distance of the observation from the mean.

All of the above comparisons are plotted in Fig. 10, for a binary Laplacian setting with $d = 2$, the designed adversarial budget of $\epsilon = 1$, $\boldsymbol{\mu}_0 = [2, 0.5]$ and $\boldsymbol{\mu}_1 = [-2, -0.5]$. We remark that the GLRT defense thus applies naturally to a wide range of such settings where adversarially robust minimax classifiers are not known or are hard to derive.

VIII. Conclusion

Our study of binary Gaussian hypothesis testing under ℓ_∞ bounded adversarial attacks shows that the GLRT approach is competitive with the known minimax benchmark. The GLRT detector has the same asymptotic performance as the minimax detector at high SNR for perturbations at the attack level that the minimax detector is designed for. For lower attack levels, the GLRT detector can provide better performance, depending on the specific values of the signal components relative to the attack budget. In general, the minimax detector is difficult to find (and may not exist), but the GLRT is a generic multi-class detector that can work with any priors and for multiple classes, as illustrated by our examples for multi-class hypothesis testing in Gaussian noise and binary hypothesis testing in Laplacian noise.

In principle, the GLRT is applicable to any classification problem in which the data distributions are known or can be estimated, but obtaining such generative models for real-world datasets is intractable in high dimensions. Even if such models were known, the computational complexity of joint estimation of the adversarial perturbation and the class can be excessive. Thus, while our results and those of related papers such as [22] indicate the potential of detection-theoretic approaches to adversarial machine learning, adapting these ideas to more complex data sets is a wide open problem.

Another interesting connection worth exploring is the connection with ideas from classical robust hypothesis testing [21], which considers a weaker adversary without access to the signal realization. As discussed in Section III-C, for binary Gaussian hypothesis testing, minimax robust detection with a noise-aware adversary (as in the adversarial machine learning literature) is identical to that with a noise-agnostic adversary (as in classical robust hypothesis testing). Thus, another interesting area for further research is the investigation of uncertainty models within the classical framework which are consistent with the norm-bounded additive perturbation models considered in the adversarial machine learning literature.

Appendix A
Lindeberg's Condition for Convergence

Denoting $\boldsymbol{\mu}[k] = \mu_k$, recall that the difference of costs under the two classes for coordinate k is given by:

$$\begin{aligned} C[k] =& \left(g_\epsilon(2\mu_k + N - \epsilon)\right)^2 - \left(g_\epsilon(N - \epsilon)\right)^2 \\ &\geq \mathbb{1}_{\{N \geq -t_k\}}(t_k + N)^2 - N^2 \triangleq Y_k, \end{aligned}$$

where $t_k = 2(|\mu_k| - \epsilon)$.

From (18) and (19) we note that the per-coordinate means and variances are finite constants. Let us define the unfavorable event as $B_k = \{|Y_k - m_k| > \delta s_d\}$. Note that the probability of this event is small. By Chernoff bounding, it can be shown, for constants $k_1 > 0$ and $k_2 > 0$ that

$$P(B_k) \leq k_1 e^{-k_2 \delta s_d}. \tag{36}$$

In the above, we used the fact that in piecewise intervals, Y_k obeys the distribution of polynomials in a Gaussian random variable, and noted that its moment-generating function exists.

The expectation term in (21) can be split as follows by conditioning on the event B_k, and observing that under B_k^c, the expectation is zero. Thus, we have

$$\begin{aligned} &\mathbb{E}\left[(Y_k - m_k)^2 \mathbb{1}_{\{|Y_k - m_k| \geq \delta s_d\}}\right] \\ &= \mathbb{E}\left[(Y_k - m_k)^2 | B_k\right] P(B_k) \\ &= \left(\mathbb{E}\left[Y_k^2 | B_k\right] + m_k^2 - 2m_k \mathbb{E}\left[Y_k | B_k\right]\right) P(B_k) \end{aligned} \tag{37}$$

Consider the computation of $\mathbb{E}[Y_k|B_k]$. Further conditioning on the event $A_k = \{N \leq -t_k\}$, it simplifies as

$$\begin{aligned} \mathbb{E}\left[Y_k | B_k\right] =& \mathbb{E}\left[Y_k | A_k, B_k\right] P(A_k | B_k) \\ &+ \mathbb{E}\left[Y_k | A_k^c, B_k\right] P(A_k^c | B_k). \end{aligned} \tag{38}$$

It can be checked from (36) the definitions of the events that under A_k and B_k, $Y_k = -N^2$, governed by the conditions $N \leq -t_k$ and $N < -\sqrt{\delta s_d - m_k}$ for large d. Further, for any $\delta > 0$, d can be chosen to be sufficiently large, so that the stricter condition turns out to be the latter. Thus, we have

$$\begin{aligned} &\lim_{d\to\infty} \mathbb{E}\left[Y_k | A_k, B_k\right] P(B_k) \\ &= \lim_{d\to\infty} \mathbb{E}\left[-N^2 | N < -\sqrt{\delta s_d - m_k}\right] P(B_k) \\ &= \lim_{d\to\infty} \left(\frac{-\sigma^2 \alpha \phi(-\alpha/\sigma) - \sigma^2 \Phi(-\alpha/\sigma)}{\Phi(-\alpha/\sigma)}\right) P(B_k) \\ &= \lim_{d,\alpha\to\infty} \left(\frac{k_3 \alpha}{R(\alpha/\sigma)} + k_4\right) P(B_k), \end{aligned} \tag{39}$$

where $\alpha = \sqrt{\delta s_d - m_k}$ and k_3 and k_4 are finite constants. The quantity $R(\alpha) = \frac{1-\Phi(\alpha)}{\phi(\alpha)}$ is called the Mills' ratio. For $\alpha > 0$, it has been shown in [35] that $\lim_{\alpha\to\infty} \alpha R(\alpha) = 1$. Using this fact and (36), we can write

$$\lim_{d\to\infty} \mathbb{E}\left[Y_k | A_k, B_k\right] P(A_k | B_k) P(B_k) = 0 \tag{40}$$

Under the events A_k^c and B_k, we have $Y_k = t_k^2 + 2t_k N$, with condition $N > (\delta s_d + m_k - t_k^2)/2t_k$. Thus, we have the following equations:

$$\begin{aligned} &\lim_{d\to\infty} \mathbb{E}\left[Y_k | A_k^c, B_k\right] P(B_k) \\ &= \lim_{d\to\infty} \mathbb{E}\left[t_k^2 + 2t_k N | N > \frac{\delta s_d + m_k - t_k^2}{2t_k}\right] P(B_k) \\ &= \lim_{d,\alpha\to\infty} \left(k_5 + \frac{k_6}{R(\alpha/\sigma)}\right) P(B_k) \\ &= 0, \end{aligned} \tag{41, 42}$$

where $\alpha = \frac{\delta s_d + m_k - t_k^2}{2t_k}$ and k_5 and k_6 are finite constants and we again used the limiting value of the Mills' ratio and the exponential bound on $P(B_k)$. Therefore, we have,

$$\lim_{d\to\infty} \mathbb{E}\left[Y_k | A_k^c, B_k\right] P(A_k^c|B_k)P(B_k) = 0 \tag{43}$$

Similarly, it can be seen that $\lim_{d\to\infty} \mathbb{E}[Y_k^2|A_k, B_k]P(B_k) = 0$ as shown below:

$$\begin{aligned}
&\lim_{d\to\infty} \mathbb{E}\left[Y_k^2|A_k, B_k\right] P(B_k) \\
&= \lim_{d\to\infty} \mathbb{E}\left[N^4 | N < -\sqrt{\delta s_d - m_k}\right] P(B_k) \\
&= \lim_{d,\alpha\to\infty} \frac{\sigma^2\alpha^3\phi(\frac{-\alpha}{\sigma}) + 3\sigma^4(\alpha\phi(\frac{-\alpha}{\sigma}) + \Phi(-\alpha/\sigma))}{\Phi(\frac{-\alpha}{\sigma})} \cdot P(B_k) \\
&= 0,
\end{aligned} \tag{44}$$

where $\alpha = \sqrt{\delta s_d - m_k}$. Along similar lines, it can be checked that $\lim_{d\to\infty} \mathbb{E}[Y_k^2|A_k^c, B_k]P(B_k) = 0$. Thus from (37), (40), (43), and (44) the Lindeberg's condition for CLT holds.

It can further be shown that the Lindeberg's condition is also satisfied by the sum of per coordinate cost differences $C[k]$. Assuming $\mu_k > \epsilon$, the expressions for $C[k]$ are obtained as

$$C[k] = \begin{cases} (2\mu_k + N - 2\epsilon)^2 - (N - 2\epsilon)^2 & N \geq 2\epsilon \\ (2\mu_k + N - 2\epsilon)^2 & 0 \leq N \leq 2\epsilon \\ (2\mu_k + N - 2\epsilon)^2 - N^2 & 2\epsilon - 2\mu_k \leq N \leq 0 \\ -N^2 & -2\mu_k \leq N \leq 2\epsilon - 2\mu_k \\ (2\mu_k + N)^2 - N^2 & N \leq -2\mu_k \end{cases} \tag{45}$$

The mean and variance of $C[k]$ are finite, as they involve conditional expectations of Gaussian powers. Following through the steps in the previous proof, computing $\mathbb{E}[C[k]|B_k]$ requires conditioning on the events A_k^i, $i \in \{1, 2 \ldots, 5\}$, considered in the branches of (45). These events partition the sample space of N. It can be shown that $\lim_{d\to\infty} \mathbb{E}[C[k]|A_k^i, B_k]P(B_k) = 0$ and $\lim_{d\to\infty} \mathbb{E}[C[k]^2|A_k^i, B_k]P(B_k) = 0$ analogous to (40), and the proof follows.

APPENDIX B
MONOTONICITY OF PER-COORDINATE COST DIFFERENCE

The per-coordinate cost difference under $\mathcal{H}_0$, restated below, is given by the following:

$$C = C_1 - C_0 = \left(g_\epsilon(2\mu + N + e)\right)^2 - \left(g_\epsilon(N + e)\right)^2 \tag{46}$$

The above expression can take one of the nine possible values based on the relative values of mean, attack and noise, that determine in which region of the double-sided ReLU their arguments lie. We show that for cases that are valid, the derivative of cost difference with respect to attack is non-negative when $\mu \geq 0$.

1) Let us first consider the case when parameters are such that the arguments of double-sided ReLU terms in both C_1 and C_0 lie in the *negative linear region*, i.e., $2\mu + e + N \leq -\epsilon$ and $e + N \leq -\epsilon$. We have,
$$C = (2\mu + e + N + \epsilon)^2 - (e + N + \epsilon)^2,$$
and $\partial C/\partial e = 4\mu$ is non-negative.
2) C_1 in negative linear region and C_0 in the *null region* ($2\mu + e + N \leq -\epsilon$ and $-\epsilon \leq e + N \leq \epsilon$) : these conditions are contradictory and the value of C defined by these regions is not legitimate.
3) Note that similar contradictions result when C_1 is in the null region and C_0 in *positive linear region* ($e + N \geq \epsilon$).
4) C_1 in positive linear and C_0 in negative linear region: we have $\partial C/\partial e = 4(\mu - \epsilon)$. If $\mu \geq \epsilon$, it is clear that the derivative is non-negative. If $\mu < \epsilon$, the conditions are not simultaneously satisfied, resulting in a contradiction.
5) For the other five cases not explicitly shown, it follows from a simple substitution of the conditions on the arguments of the double-sided ReLU and evaluating the expression for C that $\partial C/\partial e \geq 0$. Thus for any N, the per coordinate cost difference C is monotonically non-decreasing in e.

Following similar steps, when $\mu < 0$, it can be shown that C is monotonically decreasing in e. When C_1 is in negative linear and C_0 in positive linear regions, $\partial C/\partial e = 4\mu + 4\epsilon$, which is negative when $|\mu| > \epsilon$. Otherwise, we note that the inequalities $2\mu + e + N < -\epsilon$ and $e + N > \epsilon$ are not simultaneously satisfied and this case cannot occur. Similar contradictions occur when i) C_1 in null and C_0 in negative linear region; ii) C_1 in positive linear and C_0 in null region of the double-sided ReLU. For all other cases, it is easy to verify that C is decreasing in e, if $\mu < 0$.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the anonymous reviewers for their valuable feedback which helped in improving the paper.

REFERENCES

[1] B. Puranik, U. Madhow, and R. Pedarsani, "Adversarially robust classification based on GLRT," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3785–3789.
[2] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, no. 2, pp. 121–148, 2010.
[3] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.
[4] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognit.*, vol. 84, pp. 317–331, 2018.
[5] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2018, pp. 99–112.
[6] N. Carlini et al., "Hidden voice commands," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 513–530.
[7] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6989–6993.
[8] E. Wong and J. Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5286–5295. [Online]. Available: http://proceedings.mlr.press/v80/wong18a.html
[9] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," 2018, *arXiv:1801.09344*.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. 6th Int. Conf. Learn. Representations*, 2018.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[12] F. Tramér, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1633–1645.
[13] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.
[14] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 274–283.
[15] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.
[16] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. Liang, "Unlabeled data improves adversarial robustness," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11192–11203.
[17] S. Bubeck, Y. T. Lee, E. Price, and I. P. Razenshteyn, "Adversarial examples from computational constraints," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 831–840.
[18] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, pp. 1753–1758, 1965.
[19] R. Martin and S. Schwartz, "Robust detection of a known signal in nearly gaussian noise," *IEEE Trans. Inf. Theory*, vol. 17, no. 1, pp. 50–56, Jan. 1971.
[20] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5572–5587, Sep. 2017.
[21] M. Fauß, A. M. Zoubir, and H. V. Poor, "Minimax robust detection: Classic results and recent advances," *IEEE Trans. Signal Process.*, vol. 69, pp. 2252–2283, 2021.
[22] A. N. Bhagoji, D. Cullina, and P. Mittal, "Lower bounds on adversarial robustness from optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7496–7508.
[23] H. V. Poor, *An Introduction to Signal Detection and Estimation*. Berlin, Germany: Springer, 2013.
[24] A. Raghunathan, J. Steinhardt, and P. Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10900–10910.
[25] E. Wong, F. R. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8410–8419.
[26] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2266–2276.
[27] Z. Marzi, S. Gopalakrishnan, U. Madhow, and R. Pedarsani, "Sparsity-based defense against adversarial attacks on linear classifiers," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 31–35.
[28] F. Li, L. Lai, and S. Cui, "On the adversarial robustness of subspace learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 1470–1483, 2020.
[29] Y. Jin and L. Lai, "On the adversarial robustness of hypothesis testing," *IEEE Trans. Signal Process.*, vol. 69, pp. 515–530, 2021.
[30] G. Liu and L. Lai, "Action-manipulation attacks against stochastic bandits: Attacks and defense," *IEEE Trans. Signal Process.*, vol. 68, pp. 5152–5165, 2020.
[31] E. Dobriban, H. Hassani, D. Hong, and A. Robey, "Provable tradeoffs in adversarially robust classification," 2020, *arXiv:2006.05161*.
[32] P. Delgosha, H. Hassani, and R. Pedarsani, "Robust classification under ℓ_0 attack for the Gaussian mixture model," *SIAM J. Math. Data Sci.*, vol. 4, no. 1, pp. 362–385, 2022.
[33] M. S. Pydi and V. Jog, "Adversarial risk via optimal transport and optimal couplings," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 7814–7823.
[34] C. Bakiskan, S. Gopalakrishnan, M. Cekic, U. Madhow, and R. Pedarsani, "Polarizing front ends for robust CNNs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 4257–4261.
[35] R. D. Gordon, "Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument," *Ann. Math. Statist.*, vol. 12, no. 3, pp. 364–366, 1941.

Bhagyashree Puranik (Student Member, IEEE) received the B.E. degree in electronics and communication engineering from PES University, Bengaluru, India, in 2015, and the M.E. degree in electrical communication engineering from the Indian Institute of Science, Bengaluru, India, in 2017. She is currently working toward the Ph.D. degree in electrical and computer engineering with the University of California, Santa Barbara (UCSB), CA, USA. She was a Communication Systems Engineer with MaxLinear for two years. She was the recipient of the Regents Fellowship at UCSB. Her research interests include fairness and robustness in machine learning.

Upamanyu Madhow (Fellow, IEEE) received the bachelor's degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1985, and the Ph. D. degree in electrical engineering from the University of Illinois Urbana-Champaign, Champaign, IL, USA, in 1990. He is currently a Professor of Electrical and Computer Engineering with the University of California, Santa Barbara, CA, USA. He was a Research Scientist with Bell Communications Research, Morristown, NJ, and as a Faculty with the University of Illinois Urbana-Champaign. He is the author of two textbooks published by Cambridge University Press, *Fundamentals of Digital Communication* in 2008, and *Introduction to Communication Systems* in 2014. His current research interests focus on next generation communication, sensing and inference infrastructures centered around millimeter wave systems, and on robust machine learning. Dr. Madhow is the recipient of the 1996 NSF CAREER award, and co-recipient of the 2012 IEEE Marconi prize paper award in wireless communications. He was an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

Ramtin Pedarsani (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, in 2009, the M.Sc. degree in communication systems from the Swiss Federal Institute of Technology, Lausanne, Switzerland, in 2011, and the Ph.D. degree from the University of California, Berkeley, CA, USA, in 2015. He is currently an Associate Professor in ECE Department with the University of California, Santa Barbara, CA, USA. His research interests include machine learning, information and coding theory, networks, and transportation systems. He was the recipient of the Communications Society and Information Theory Society Joint Paper Award in 2020, the Best Paper Award in the IEEE International Conference on Communications (ICC) in 2014, and the NSF CRII Award in 2017.