Communication without Interception: Defense against Modulation Detection

Muhammad Zaid Hameed[1], András György[2], and Deniz Gündüz[1]
[1]Imperial College London, UK [2]DeepMind, London, UK

***Abstract*—We consider a communication scenario, in which an intruder tries to determine the modulation scheme of the intercepted signal. Our aim is to minimize the accuracy of the intruder, while guaranteeing that the intended receiver can still recover the underlying message with the highest reliability. This is achieved by constellation perturbation at the encoder, similarly to adversarial attacks against classifiers in machine learning. In image classification, the perturbation is limited to be imperceptible to a human observer, while in our case the perturbation is constrained so that the message can still be reliably decoded by a legitimate receiver that is oblivious to the perturbation. Simulation results demonstrate the viability of our approach to make wireless communication secure against both state-of-the-art deep-learning- and decision-tree-based intruders with minimal sacrifice in the communication performance.**

I. Introduction

Securing wireless communications is as essential for military, commercial as well as consumer communication systems. The standard approach is to encrypt the data; however, encryption may not always provide full security (e.g., side-channel attacks), or strong encryption may not be available due to complexity limitations (e.g., IoT devices). Encryption can be complemented with other techniques, preventing the adversary from even recovering the encrypted bits. As outlined in [1], an adversary implements its attacks in four steps: 1) tunes into the frequency of the transmitted signal; 2) detects whether there is signal or not; 3) intercepts the signal by extracting its features; and 4) demodulates the signal by exploiting the extracted features, and obtains a binary stream of data. Preventing any of these steps can strengthen the security. While encryption focuses on protecting the demodulated bit stream, physical layer security [2], [3] targets the fourth step. Recently, there has also been significant interest in preventing the second step through covert communications [4]. In this work, we instead focus on the third step, and aim at preventing the adversary from detecting the modulation scheme.

Modulation detection is the step between signal detection and demodulation, and thus plays an important role in data transmission, as well as in detection and jamming of unwanted signals in military communications and other sensitive applications [5]. Recently, deep learning techniques have led to significant progress in modulation detection accuracy, where convolutional neural networks (CNNs) and other deep neural networks (DNNs) were applied for modulation detection directly from the received symbols, without any explicit feature representation, surpassing the accuracy of traditional detectors based on likelihood function or feature-based representations [6]–[8].

Our goal is to prevent an intruder that employs a state-of-the-art modulation detector from successfully identifying the modulation scheme being used. We argue that, if the intruder is unable to identify the modulation scheme, it is unlikely to be able to decode the signal, or employ modulation-dependent jamming techniques. This would be trivial by sacrificing the performance of the intended receiver. The main challenge here is to guarantee that the intended receiver can continue to receive the underlying message at a reasonable probability of error. Here, we assume that the intended receiver is oblivious to the modifications employed by the transmitter to confuse the intruder; and therefore, the goal of the transmitter is to introduce as small modifications to the transmitted signal as possible, which are sufficient to fool the intruder but not larger than the error correction capabilities of the intended receiver.

Introducing small variations into the modulation scheme that can fool an intruder is similar to adversarial attacks on classifiers, in particular DNNs [9]–[11]. While the goal in these attacks is to expose the vulnerability of classifiers against small changes in the input, we exploit the same approach to defend a communication link against an intruder that employs DNNs or other standard classification methods for interception.

II. System Model

The transmitter maps a binary input sequence $\mathbf{w} \in \{0,1\}^m$ into a sequence of n complex symbols, $\mathbf{x} \in \mathbb{C}^n$, employing forward error correction. Formally, $\mathbf{x} = M_s(\mathbf{w})$, where $s \in \mathcal{S}$ is the modulation scheme, $\mathcal{S}$ the set of available modulation schemes. We assume that M_s satisfies the power constraint $(1/n)\|\mathbf{x}\|_2^2 \leq 1$, $\forall \mathbf{w}$. Signals $\mathbf{y}_1$ and $\mathbf{y}_2$, received by the receiver and the intruder, respectively, are given by

$$\mathbf{y}_i = M_s(\mathbf{w}) + \mathbf{z}_i = \mathbf{x} + \mathbf{z}_i,\ i = 1, 2,$$

where $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{C}^n$ are independent channel noise (also independent of $\mathbf{x}$) with independent zero-mean complex Gaussian components of variance σ_1^2 and σ_2^2, respectively. The intended receiver, upon receiving $\mathbf{y_1}$, decodes the underlying message with the goal of minimizing the bit error rate

$$e(\mathbf{w}, \mathbf{y}_1) = \textstyle\sum_{i=1}^m \mathbb{I}\{w_i \neq \hat{w}_i\}, \tag{1}$$

where $\hat{\mathbf{w}}$ is the decoded bit sequence from $\mathbf{y}_1$.[1]

The intruder aims to determine, for any $\mathbf{y}_2 \in \mathbb{C}^n$, the modulation scheme used by the transmitter, whereas the transmitter wants to communicate without its modulation scheme being detected by the intruder, while keeping the

[1]For any event E, $\mathbb{I}\{E\} = 1$ if E holds, and 0 otherwise. Furthermore, for any real or complex vector $\mathbf{v}$, v_i denotes its ith coordinate.

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

BER in an acceptable range. We consider that the intruder implements a score-based classifier and assigns to $\mathbf{y}_2$ the label $\hat{s} = \text{argmax}_{s' \in \mathcal{S}} f_\theta(\mathbf{y}_2, s')$, where $f_\theta : \mathbb{C}^n \times \mathcal{S} \to \mathbb{R}$ is a score function parametrized by $\theta \in \mathbb{R}^d$, which assigns a score (pseudo-likelihood) to each possible class $s' \in \mathcal{S}$ for every $\mathbf{y}_2$, and finally selects the class with the highest score. With a slight abuse of notation, we denote the resulting class label by $\hat{s} = f_\theta(\mathbf{y}_2)$. The goal of the intruder is to maximize its success probability $\Pr(s = \hat{s})$.[2] For the state-of-the-art modulation detection scheme of [7], f_θ is a CNN classifier, with θ being the weights of the neural network, while $f(\mathbf{y}_2, s')$ are the so-called logit values for the class labels $s' \in \mathcal{S}$.

III. Modulation perturbation to avoid detection

Our goal is to modify the encoder M_s such that, given $s \in \mathcal{S}$, the new encoding method M'_s ensures that the intruder's success probability is smaller, while the BER of the receiver (using the same decoding procedure for M_s) does not increase substantially. Our solution is motivated by adversarial attacks for image classification, where modifications imperceptible to a human observer can fool state-of-the-art image classifiers [9], [10]. Adversarial examples are particularly successful in fooling high-dimensional DNN classifiers. Applying the same idea to our problem, we aim to find modified modulation schemes M'_s such that $M'_s(\mathbf{w}) \approx M_s(\mathbf{w})$, but the intruder misclassifies the new received signal $\mathbf{y}'_2 = M'_s(\mathbf{w}) + \mathbf{z}_2$ with higher probability.

A. Adversarial attacks in an idealized scenario

Similarly to adversarial attacks on image classifiers [10], an idealized adversarial attack to classifier f_θ of the intruder would modify a correctly classified channel output sequence $\mathbf{y}_2$ (i.e., for which $s = f_\theta(\mathbf{y}_2)$) with a perturbation $\delta \in \mathbb{C}^n$ such that $f_\theta(\mathbf{y}_2 + \delta) \neq f_\theta(\mathbf{y}_2)$, the true label. In parallel, we require that the same modification to the input of the decoder does not hurt its performance, that is, the sequence $\mathbf{y}_1 + \delta$ is decoded at a similar accuracy as $\mathbf{y}_1$. To facilitate this, we impose $\|\delta\|_2 \leq \epsilon$ for some small positive constant ϵ. Thus, to mask the modulation scheme and keep the BER reasonable, we aim to find, for each correctly classified $\mathbf{y}_2$ separately, a perturbation δ that maximizes the zero-one loss:

$$\text{maximize } \mathbb{I}\{f_\theta(\mathbf{y}_2 + \delta) \neq s\} \text{ such that } \|\delta\|_2 \leq \epsilon \ , \quad (2)$$

where $s = f_\theta(\mathbf{y}_2)$ is the true modulation label.

Such a δ results in a successful adversarial perturbation and a successful adversarial example $\mathbf{y}_2 + \delta$ (i.e., one for which the intruder makes a mistake), while the BER is likely still small. Thus, in practice we could achieve our goal if we could modify the encoder such that the channel output at the intruder is $\mathbf{y}_2 + \delta$, and $\mathbf{y}_1 + \delta$ at the receiver. However, in practice we can only control the channel input $\mathbf{x}$, and the channel outputs $\mathbf{y}_1$ and $\mathbf{y}_2$ depend not only on $\mathbf{x}$, but also on the channel noise. Therefore, we refer to the above mechanism, which was analyzed in [12], as an idealized and impractical scenario, and use it only as a baseline.

We note that the target function $\mathbb{I}\{f_\theta(\mathbf{y}_2 + \delta) \neq f_\theta(\mathbf{y}_2)\}$ in (2) is binary, hence flat, and thus, no gradient-based search is directly possible. To alleviate this, usually a surrogate loss function $L(\theta, \mathbf{y}_2, s)$ to the zero-one loss is used (which is also used in training the classifier f_θ, i.e., finding the parameter vector θ with the best classification performance over a training data set), which is amenable to gradient-based optimization. For classification problems, a standard choice is the cross-entropy loss, $L(\theta, \mathbf{y}_2, s) = -\log(1 + e^{-f_\theta(\mathbf{y}_2, s)})$, and one can search for adversarial perturbations by solving

$$\text{maximize } \ L(\theta, \mathbf{y}_2 + \delta, s) \text{ such that } \ \|\delta\|_2 \leq \epsilon. \quad (3)$$

Different methods are used in the literature to solve (3) approximately [10], [13]. In this paper we use the state-of-the-art projected (normalized) gradient descent (PGD) attack [14] to generate adversarial examples, which is an iterative method: starting from $\mathbf{y}^0 = \mathbf{y}_2$, in iteration t it calculates

$$\mathbf{y}^t = \Pi_{\mathcal{B}_\epsilon(\mathbf{y}_2)}\big(\mathbf{y}^{t-1} + \beta \ \text{sign}(\ \nabla_{\mathbf{y}} L(\theta, \mathbf{y}^{t-1}, s))\big), \quad (4)$$

where $\beta > 0$ denotes the step size, 'sign' denotes the sign operation, and $\Pi_{\mathcal{B}_\epsilon(\mathbf{y}_2)}$ denotes the Euclidean projection operator to the L_2-ball $\mathcal{B}_\epsilon(\mathbf{y}_2)$ of radius ϵ centered at $\mathbf{y}_2$.The attack is typically run for a given number of steps, depending on the computational resources; in practice $\mathbf{y}^t$ is more likely to be a successful adversarial example for larger values of t. We will refer to this *idealized* scheme as the *Oracle Defensive Modulation Scheme (ODMS)*.

Note that this formulation assumes that we have access to the logit function f_θ of the intruder; these methods are called *white-box* attacks. If f_θ is not known, one can create adversarial examples against another classifier $f_{\theta'}$, and hope that it will also work against the targeted model f_θ. Such methods are called *black-box* attacks, and are surprisingly successful against image classifiers [15].

B. Practical methods

The perturbation method in Section III-A is infeasible as the channel noise at the intruder is not known, and a practical scheme can only modify $\mathbf{x} = M_s(\mathbf{w})$. Thus, the new modulation scheme is defined as

$$M'_s(\mathbf{w}) = \alpha(M_s(\mathbf{w}) + \delta),$$

where we will consider different choices for $\delta \in \mathbb{C}^n$, and the multiplier $\alpha = \sqrt{n}/\|M_s(\mathbf{w}) + \delta\|_2$ is used to ensure that the new channel input $\bar{\mathbf{x}} = M'_s(\mathbf{w})$ satisfies the average power constraint $(1/n)\|\bar{\mathbf{x}}\|_2^2 \leq 1$. The signals received at the receiver and the intruder are $\bar{\mathbf{y}}_1 = \bar{\mathbf{x}} + \mathbf{z}_1$ and $\bar{\mathbf{y}}_2 = \bar{\mathbf{x}} + \mathbf{z}_2$, respectively. The difficulty in this scenario is that the effect of any carefully designed perturbation δ may (and, in fact, will) be at least partially masked by the channel noise. Furthermore, since now the perturbed signal is transmitted at the actual channel SNR, the effective SNR of the system is decreased, as the transmitted signal already includes the perturbation δ, which can be treated as noise from the intended receiver's point of view.

[2]Here we assume an underlying probabilistic model about how the bit sequence $\mathbf{w}$ and the modulation scheme are selected.

We present three practical methods to find a perturbation δ. The first and simplest one disregards the effects of the channel noise and the resulting BER at the receiver:

1) Defensive modulation scheme without BER control (DMS): In this method we aim to solve the optimization problem (3) with $\mathbf{x}$ in place of $\mathbf{y}_2$, via (4) initialized at $\mathbf{y}^0 = \mathbf{x}$, and with projection to $\mathcal{B}_\epsilon(\mathbf{x})$ (for a given number of iterations t and perturbation size ϵ).

Next, we consider methods that also take into account the BER, $e(\bar{\mathbf{y}}_1, \mathbf{w})$ at the receiver (see Eqn. 1): that is, instead of enforcing the perturbation δ to be small and hoping for only a slight increase in the BER, we also explicitly optimize for the latter. There is an inherent trade-off between these two targets: a larger δ results in a bigger reduction in the detection accuracy of the intruder, but will also increase the BER at the receiver. We consider two methods to handle this trade-off:

2) BER-aware defensive modulation scheme (BDMS): Consider a (signed) linear combination of the two target functions

$$L_\lambda(\theta, \bar{\mathbf{x}}, s, \mathbf{z}_1, \mathbf{z}_2) = L(\theta, \bar{\mathbf{x}} + \mathbf{z}_2, \delta) - \lambda e(\bar{\mathbf{x}} + \mathbf{z}_1, \mathbf{w})$$

for some $\lambda > 0$, where $\bar{\mathbf{y}}_i = \bar{\mathbf{x}} + \mathbf{z}_i, i = 1, 2$, and aim to find a perturbation δ, or, equivalently, a modulated signal $\bar{\mathbf{x}} = \mathbf{x} + \delta$, that maximizes the expectation

$$\mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[L_\lambda(\theta, \mathbf{x}, s, \mathbf{z}_1, \mathbf{z}_2)] \tag{5}$$

with respect to the channel noise $\mathbf{z}_1, \mathbf{z}_2$. Here we can use stochastic gradient descent (ascent) to compute an approximate local optimum, but in practice we find that enforcing δ to be small during the iterations improves the performance; hence, we use a stochastic version of PGD (4): starting at $\mathbf{x}^0 = \mathbf{x}$, our candidate for $\bar{\mathbf{x}}$ is iteratively updated as

$$\mathbf{x}^t = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}\left(\mathbf{x}^{t-1} + \beta \ \mathrm{sign}(\nabla_{\mathbf{x}} L(\theta, \mathbf{x}^{t-1}, s, \mathbf{z}_1^t, \mathbf{z}_2^t))\right), \tag{6}$$

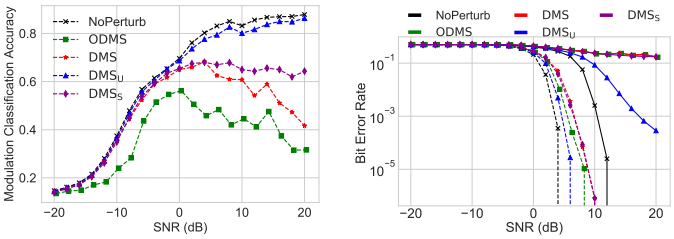
where $\mathbf{z}_i^t$ are independent copies of $\mathbf{z}_i$, for $i = 1, 2$, and $t = 1, 2, \ldots$. Although $\mathbb{E}_{\mathbf{z}_1}[e(\bar{\mathbf{x}} + \mathbf{z}_1, \mathbf{w})]$ is differentiable, $e(\mathbf{y}, \mathbf{w})$ for a given fixed $\mathbf{y}$ is not (since it takes values from the finite set $\{0, 1/n, \ldots, 1\}$). Similarly to [16], we approximate the gradient of the expected error using SPSA [17] as

$$\bar{\nabla}_{\mathbf{y}}\, e(\mathbf{y}, \mathbf{w}) = \frac{1}{K} \sum_{k=1}^{K} \frac{e(\mathbf{y} + \eta \mathbf{r}_k, \mathbf{w}) - e(\mathbf{y} - \eta \mathbf{r}_k, \mathbf{w})}{2\eta} \mathbf{r}_k^\top,$$

where $\mathbf{r}_1, \ldots, \mathbf{r}_K$ are random vectors selected independently and uniformly from $\{-1, 1\}^n$.

3) BER-aware orthogonal defensive modulation scheme (BODMS): An alternative method is that instead of maximizing the combined target (5), we try to maximize the cross-entropy loss $L(\theta, \bar{\mathbf{y}}_2, s)$, without increasing (substantially) the BER $e(\bar{\mathbf{y}}_1, \mathbf{w})$. In order to do so, we maximize $L(\theta, \bar{\mathbf{y}}_2, s)$ using stochastic PGD (again, in every step we choose independent noise realizations), but we restrict the steps to directions in which the bit error rate does not change. Thus, in every step we update $\mathbf{x}^{t-1}$ in a direction *orthogonal* to the gradient of the BER defined as

$$\nabla_{\mathbf{x}} L(\theta, \mathbf{x}^{t-1} + \mathbf{z}_2^t, s) - \left\langle \nabla_{\mathbf{x}} L(\theta, \mathbf{x}^{t-1} + \mathbf{z}_2^t, s), d_e \right\rangle d_e,$$



(a) Modulation classification accuracy (b) Bit error rate

Fig. 1: Modulation classification accuracy of the intruder and bit error rates for PSK8 (dashed lines) and QAM64 (solid lines) as a function of the SNR for different defensive modulation schemes.

where $d_e = \bar{\nabla}_{\mathbf{x}} e(\mathbf{x}^{t-1} + \mathbf{z}_1^t, \mathbf{w}) / \|\bar{\nabla}_{\mathbf{x}} e(\mathbf{x}^{t-1} + \mathbf{z}_1^t, \mathbf{w})\|_2$ is the (approximate) gradient direction of the BER.

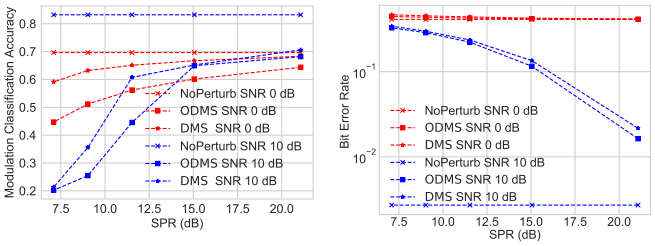
IV. Experimental evaluation

We assume that the binary source data is generated independently, uniformly at random, and is encoded using a rate 2/3 convolutional code. Eight modulation schemes are considered: 'GFSK', 'CPFSK', 'PSK8', 'BPSK', 'QPSK', 'PAM4', 'QAM16', 'QAM64', and the channel signal-to-noise ratio (SNR) varies between -20dB and 20dB. After demodulation, the receiver uses Viterbi decoding to estimate the original source data. Modulation detection should be completed based on a short sequence of intercepted complex I/Q (in-phase /quadrature) channel symbols; as in [7], we set the sequence length to $n = 128$. As the classifier, we first consider the same CNN architecture in [7] for the intruder, which operates on the aforementioned 256-dimensional data.

For each modulation scheme, we generate data resulting in approximately 245000 I/Q channel symbols, split into blocks of 128 I/Q symbols. The last 300 blocks for each modulation scheme are reserved for testing the performance (tests are repeated 20 times), while we train a separate classifier for each SNR value based on the data in the preceding blocks. As shown in Fig. 1a (see the graph with label 'NoPerturb'), for high SNR values the accuracy is close to 90%.

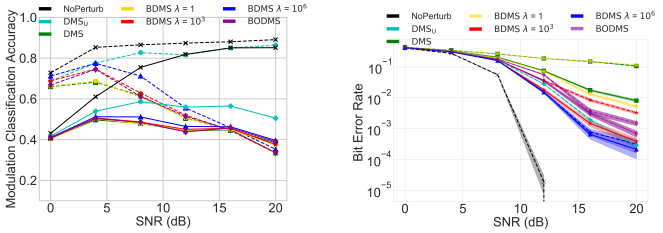
We compare (i) our three defensive modulation schemes, *DMS*, *BDMS*, and *BODMS*; (ii) the oracle defensive modulation scheme *ODMS*; (iii) adding uniform random noise of L_2-norm ϵ to a block, called *DMS-uniform* (*DMS*$_U$); (iv) a black-box mechanism that calculates *DMS* against a classifier that has the same architecture as the intruder, but is trained separately (assuming no channel noise); we call this *substitute DMS* (*DMS*$_S$).

All the above schemes, except for *DMS*$_U$, are implemented using the PGD method (6) from the CleverHans Library [18], with 20 iterations, $\beta = 0.2$ and $\epsilon = 3$. *DMS*$_U$ uses the same ϵ. Note that a perturbation of this size accounts for about 7% of the total energy of a block (which is 128 due to our normalization to the energy constraint). *ODMS* serves as an upper bound on the achievable defensive performance in practice, while the role of *DMS*$_U$ is to analyze the effect of carefully crafted perturbations instead of selecting them



(a) Modulation classification accuracy (b) Bit error rate

Fig. 2: Effect of SPR on the modulation classification accuracy and the bit error rate (QAM64).



(a) Modulation classification accuracy (b) Bit error rate

Fig. 3: Classification accuracy and BER (QAM64) for BER-aware schemes ($\epsilon = 3$), against the DNN-based detector (dashed lines) and for the decision tree (solid lines).

randomly. DMS_S explores the more practical situation where the exact classifier of the intruder is not known, but its training method and/or a similar classifier is available.

Fig. 1a shows the modulation classification accuracy for the different methods. It can be seen that adding random noise (DMS_U) helps very little compared to no defense at all (*NoPerturb*). The basic defense mechanism *DMS* and its black-box version (DMS_S) become effective from about -5 dB SNR, and—as expected—*DMS* outperforms DMS_S. For smaller SNR values the classification accuracy is relatively small (the channel noise already makes classification hard), and only the oracle defense *ODMS* could gives noticeable improvement. As expected, the performance of *DMS* gets closer to its lower bound, *ODMS*, as SNR increases (note that the two methods coincide at the limit of infinite SNR). Similar performance of DMS_S and *DMS* at medium SNR illustrates transferability of adversarial perturbations in our model, as was observed in other machine learning problems, e.g., in image classification [15], although *DMS* becomes more effective as SNR increases. Observe that the classification accuracy of *DMS* increases up to 0 dB SNR, where the noise is the main cause of the performance limitation of the intruder, while the accuracy decreases for larger SNR where the defense mechanism starts working.

The reduced classification accuracy of the intruder for *DMS* and DMS_S are countered by the increased BER at the receiver. To illustrate this effect, Fig. 1b shows the BER for PSK8 and QAM64; for other modulation schemes the BER behaves similarly to the case of PSK8, with up to 5dB difference of where it starts to drop sharply. On the other hand, the price of using any defense mechanism on QAM64 is severe, causing orders of magnitude larger BER in the high SNR regime. This can be suppressed if the perturbation size is decreased, which—at the same time—results in increased detection accuracy. This is shown in Fig. 2, as a function of the signal-to-perturbation ratio, $\mathrm{SPR} \triangleq n/\|\delta\|_2^2$ (recall $n = 128$, and $\mathrm{SPR} \approx 11.5$dB corresponds to $\epsilon = 3$). In every case, *DMS* trades off increased BER for reduced detection accuracy compared to no-defense.

In addition to DNN-based detectors, we also examine defense against one of the best "standard" modulation detection schemes in the literature, a multi-class decision tree trained with expert features obtained from [19], [20].

Fig. 3 shows the modulation classification accuracy and the BER for *BDMS* and *BODMS*, also compared with *DMS*, DMS_U and without any defense mechanism, for both DNN- and decision-tree-based intruders. For *BDMS*, multiple λ values are considered. Due to space constraints, the BER is only shown for QAM64 as this is the modulation scheme most affected by our perturbations (similarly to the previous experiments, the error rate for other modulation schemes becomes very small for all defense mechanisms for larger SNR values, but the relative performance of the schemes is similar to that of QAM64).

For each modulation detector, it can be seen that for large SNR ($\geq$ 12dB), all defensive schemes achieve roughly the same accuracy (much smaller than for the no-defense case and DMS_U), while *BODMS* and *BDMS* for large λ provide significant improvements in the BER (shown for QAM64). Note, however, that the errors are still significantly higher than for the standard QAM64 modulation. The BER values for the DNN- and tree-based classifiers are approximately the same for *BDMS* and *BODMS*, while the accuracy of the DNN classifier is consistently higher, except for some cases for high SNR, when they are approximately the same. For larger λ values, the BER of *BDMS* is smaller than or approximately the same as for DMS_U, which adds uniform random noise of the same perturbation size, while it significantly outperforms DMS_U in classification accuracy. Note that *BODMS* approaches the performance of *BDMS* with a large λ ($10^3 - 10^6$), without the need to tune the hyperparameter λ, and these methods provide a good compromise between the effectiveness of the defense and the increase in the BER.

V. Conclusions

We proposed a novel secure communication scheme against an intruder whose goal is to detect the modulation scheme (which is typically the first step of a more advanced attack). In the proposed scheme, the constellation at the transmitter is perturbed using an adversarial perturbation derived against intruder's modulation classifier. Experimental results on synthetic problems verify the viability of our approach by showing that our method is able to substantially reduce the modulation classification accuracy of the intruder with minimal sacrifice in the communication performance.

REFERENCES

[1] G. E. Prescott, "Performance metrics for low probability of intercept-communication system," in *Air Force Off. of Sci. Res., Tech. Rep.*, 1993.

[2] A. D. Wyner, "The wire-tap channel," *The Bell Sys. Tech. Journal*, vol. 54, no. 8, pp. 1355–1387, Oct 1975.

[3] D. Gunduz, D. R. Brown, and H. V. Poor, "Secret communication with feedback," in *2008 Int'l Symp. on Inform. Theory and Its Apps.*, Dec 2008, pp. 1–6.

[4] B. Bash *et al.*, "Square root law for communication with low probability of detection on AWGN channels," *CoRR*, vol. abs/1202.6423, 2012.

[5] O. A. Dobre *et al.*, "Survey of automatic modulation classification techniques: classical approaches and new trends," *IET Comm.*, vol. 1, no. 2, pp. 137–156, 2007.

[6] G. J. Mendis *et al.* , "Deep learning-based automated modulation classification for cognitive radio," in *IEEE Int. Conf. on Comm. Sys. (ICCS)*. IEEE, 2016, pp. 1–6.

[7] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Comm.&Net.*, vol. 3, no. 4, 2017.

[8] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *IEEE DySPAN*, 2017, pp. 1–6.

[9] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[10] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] S. Kokalj-Filipovic, R. Miller, and J. Morman, "Targeted adversarial examples against RF deep classifiers," in *Proceedings of the ACM Workshop on Wireless Security and Machine Learning*. ACM, 2019, pp. 6–11.

[12] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Comm. Letters*, 2018.

[13] A. Kurakin *et al.*, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[14] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[15] N. Papernot *et al.*, "Practical black-box attacks against machine learning," in *ACM Asia Conf. Comp. and Comm. Security*, 2017, pp. 506–519.

[16] J. Uesato *et al.*, "Adversarial risk and the dangers of evaluating against weak attacks," *arXiv preprint arXiv:1802.05666*, 2018.

[17] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. on Automatic Ctrl.*, vol. 37, no. 3, pp. 332–341, 1992.

[18] N. Papernot *et al.*, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint arXiv:1610.00768*, 2018.

[19] A.-V. Rosti, "Statistical methods in modulation classification," 1998.

[20] A. Abdelmutalab *et al.*, "Automatic modulation classification based on high order cumulants and hierarchical polynomial classifiers," *Physical Comm.*, vol. 21, pp. 10–18, 2016.