A Novel Method to Mitigate Adversarial Attacks on AI-driven Power Allocation in D-MIMO

Ömer Faruk Tuna
Ericsson Research
Istanbul, Turkey
omer.tuna@ericsson.com

Fehmi Emre Kadan
Ericsson Research
Istanbul, Turkey
fehmi.emre.kadan@ericsson.com

***Abstract*—Adversarial attacks have the potential to substantially compromise the security of AI-powered systems and posing high risks especially in the areas like telecommunication where security is a top priority. In this study, we focus on adversarial attacks targeting power allocation for the distributed multiple-input multiple-output networks. We propose a novel defense method to mitigate the effects of these attacks and help boosting the natural performance of the system. The detailed simulations show that the proposed method significantly increases the robustness of the system.**

***Index Terms*—Distributed MIMO, cell-free massive MIMO, power allocation, deep learning, trustworthy AI, 6G security.**

I. Introduction

Wireless networks must perform complicated tasks in a dynamic spectrum environment that is influenced by channel, interference, and traffic effects. Deep learning has emerged as a valuable tool for assisting with many wireless communication tasks. Deep Neural Networks (DNN) have been utilized to solve a variety of wireless network challenges such as power-allocation for multiple-input multiple-output (MIMO) systems, spectrum sensing, RF signal classification, signal authentication, and anti-jamming.

Despite having a track record of success in wireless applications, Machine Learning (ML) also poses some distinct security challenges. Recent research has revealed that numerous adversarial attacks can be deployed effectively against DNN-based wireless systems [1], [2]. Because of their small footprints, adversarial ML-based attacks are more covert and difficult to detect when compared to traditional wireless attacks such as jamming.

Distributed MIMO (D-MIMO) is a candidate radio access network technology for 6G and beyond where the radio units (RUs) are distributed over an area to increase the macro-diversity and decrease the shadowing effects. The distributed RUs are connected to a central processor (CP) via fronthaul links for coordinated joint transmission/reception of the signals. Power allocation together with precoding at RUs is performed to optimize user spectral efficiencies (SEs) and satisfy uniform and high quality-of-service to all user equipment (UE) connected to the network. The benefits of D-MIMO networks are described in detail in [3].

Adversarial attacks have the potential to substantially compromise the security of DNN-powered systems and posing high risks especially in the areas like telecommunication where security is a top priority. In this study, we focus on adversarial attacks on the power allocation functionality of the D-MIMO networks and propose a novel method which can be employed on the serving AI agent aiming to reduce the negative impact of such kind of attack threats to a reasonable level and help make future 6G power control methods in D-MIMO robust to such smart attacks.

II. System Model

In this study, we consider a D-MIMO network with M single antenna RUs and K single antenna UEs. All UEs are jointly served by all RUs in the same time/frequency resource block. All RUs are connected to a CP via fronthaul links. Fig. 1 shows an example D-MIMO network.

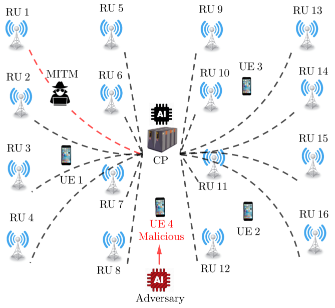

Fig. 1: An example D-MIMO network with 16 RUs and 4 UEs. Here, UE 4 is malicious and there is a man-in-the-middle (MITM) attack on the fronthaul link of the RU-1.

We focus on power allocation problem for downlink transmission to optimize the user SEs. We consider max-min fairness power control approach to maximize the minimum SE of UEs. This is a common approach satisfying uniform SEs for all UEs [4]. The complex baseband signal model for the transmitted signal from the m-th RU can be written as

$$x_m = \sum_{k=1}^{K} \sqrt{P_t \eta_{m,k}} w_{m,k} s_k, \quad \forall k, \tag{1}$$

where x_m is the transmitted signal from the m-th RU, P_t is the average transmit power limit of each RU, $\eta_{m,k}$ is the power control coefficient for the pair m-th RU and the k-th UE that controls the power allocation between UEs,

$w_{m,k}$ is the precoder of the m-th RU designed for k-th UE, and s_k is the information data of the k-th UE. We assume maximal ratio transmission (MRT) precoding that is commonly preferred by means of its local implementation at RUs. By MRT precoding, we have $w_{m,k} = h^*_{m,k}$ where $h_{m,k}$ is the instantaneous channel coefficient between the m-th RU and the k-th UE. Throughout the paper, we assume that $h_{m,k}$ is perfectly known by CP. The findings of this paper can be directly generalized to imperfect channel estimation case that we do not consider in this study for the sake of simplicity. The received signal by the k-th UE can be calculated as

$$y_k = \sum_{m=1}^{M} h_{m,k} x_m + z_k = \sum_{m=1}^{M} \sum_{\ell=1}^{K} \sqrt{P_t \eta_{m,\ell}} h_{m,k} h^*_{m,\ell} s_\ell + z_k, \tag{2}$$

where $z_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the thermal noise at the k-th UE receiver with an average power σ_k^2. The received signal y_k includes four parts which can be given as

$$y_k = y_{\text{desired},k} + y_{\text{mismatch},k} + y_{\text{interference},k} + z_k, \tag{3}$$

where

$$\begin{aligned} y_{\text{desired},k} &= \sum_{m=1}^{M} \sqrt{P_t \eta_{m,k}} \mathbb{E}[|h_{m,k}|^2] s_k \\ y_{\text{mismatch},k} &= \sum_{m=1}^{M} \sqrt{P_t \eta_{m,k}} (|h_{m,k}|^2 - \mathbb{E}[|h_{m,k}|^2]) s_k \\ y_{\text{interference},k} &= \sum_{\ell \neq k}^{K} \sum_{m=1}^{M} \sqrt{P_t \eta_{m,\ell}} h_{m,k} h^*_{m,\ell} s_\ell. \end{aligned} \tag{4}$$

Here we assume that each UE only knows the mean of its effective channel, i.e., $\mathbb{E}[|h_{m,k}|^2]$, and hence the desired signal only includes the mean part of the effective channel. Mismatch part includes the signal s_k with unknown coefficient and the signal related to other user data is considered under interference part. Using the information-theoretic approach in [5], achievable user SE for the k-th UE can be obtained as

$$\text{SE}_k = \log_2 \left(1 + \frac{|\mathbb{E}[y_{\text{desired},k}]|^2}{\mathbb{E}[|y_{\text{mismatch},k}|^2 + |y_{\text{interference},k}|^2] + \sigma_k^2} \right). \tag{5}$$

We maximize the minimum of SE_k's under the per-RU transmit power constraints $\mathbb{E}[|x_m|^2] \leq P_t$ which can be rewritten as $\sum_{k=1}^{K} \beta_{m,k} \eta_{m,k} \leq 1, \ \forall m$. Here $\beta_{m,k} = \mathbb{E}[|h_{m,k}|^2]$ is the large-scale fading coefficient that includes path-loss and shadowing effects. We consider a block fading model where we evaluate all expectations in (5) over Rayleigh small-scale fading with $h_{m,k} \sim \mathcal{CN}(0, \beta_{m,k})$. The expectations can be evaluated as given by [4] and we obtain the problem (P1)

$$\text{(P1)} \ \max_{\eta_{m,k}} \min_{k} \log_2(1 + \text{SINR}_k) \ \text{ such that } \forall m, k \tag{6}$$

$$\text{SINR}_k = \frac{\left(\sum_{m=1}^{M} \sqrt{\eta_{m,k}} \beta_{m,k} \right)^2}{\sum_{\ell=1}^{K} \sum_{m=1}^{M} \eta_{m,\ell} \beta_{m,\ell} \beta_{m,k} + \frac{\sigma_k^2}{P_t}}, \quad \sum_{k=1}^{K} \beta_{m,k} \eta_{m,k} \leq 1.$$

(P1) can be optimally solved using bisection search [4] with the asymptotic complexity $\mathcal{O}(N_{\text{iter}} \sqrt{K+M} M^3 K^4)$ [6] where N_{iter} is the number of iterations required in the bisection search. As the complexity is very high, AI based solutions are proposed in the literature [7].

The power allocation problem defined by (P1) has the input channel coefficient vector $\boldsymbol{\beta} = [\beta_{1,1} \ \beta_{1,2} \ \ldots \ \beta_{M,K}]^T \in \mathbb{R}^{MK}$ and the output power control coefficient vector $\boldsymbol{\eta} = [\eta_{1,1} \ \eta_{1,2} \ \ldots \ \eta_{M,K}]^T \in \mathbb{R}^{MK}$. Once we know $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, the operation, denoted by the function $q(\boldsymbol{\beta}, \boldsymbol{\eta})$, of finding the SE vector $\textbf{SE} = [\text{SE}_1 \quad \text{SE}_2 \quad \ldots \quad \text{SE}_K]^T$ is a straightforward analytical operation which can be done using the SINR formulas in (6). Fig. 2 describes the diagram of the analytical solution and the corresponding SEs of (P1).

$$\boldsymbol{\beta} \xrightarrow{f(\boldsymbol{\beta})} \boldsymbol{\eta} \qquad\qquad \boldsymbol{\beta}, \boldsymbol{\eta} \xrightarrow{q(\boldsymbol{\beta}, \boldsymbol{\eta})} \textbf{SE}$$

Fig. 2: Analytical solution. It finds $\boldsymbol{\eta}$ by analytically solving the problem (P1) defined in (6).

Since the analytical solution $f(\boldsymbol{\beta})$ of finding the optimum power allocation coefficients $\boldsymbol{\eta}$ is a highly complex operation, the exact solution is generally approximated by a well-trained AI model as shown in Fig. 3. In this option, using a training data which is composed of $\boldsymbol{\beta}$ and the corresponding optimum $\boldsymbol{\eta}$ vector, we train a regression type AI model and use this model at the inference time to speed up the decision-making process and avoid the extensive computations.

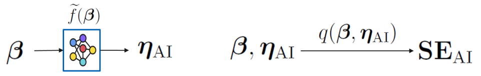
$\widetilde{f}(\boldsymbol{\beta})$
$\boldsymbol{\beta}$
$\boldsymbol{\eta}_{\text{AI}}$
$\boldsymbol{\beta}, \boldsymbol{\eta}_{\text{AI}}$
$q(\boldsymbol{\beta}, \boldsymbol{\eta}_{\text{AI}})$
$\textbf{SE}_{\text{AI}}$

Fig. 3: AI-based solution. The AI solution is denoted by $\widetilde{f}$ and it approximates f by maintaining $\textbf{SE} \approx \textbf{SE}_{\text{AI}}$.

III. Adversarial Attacks on AI Models

It is known that AI models are highly vulnerable to adversarial attack threats where carefully crafted perturbations can create significant errors in the expected outputs. Previous research studies have investigated potential adversarial attack threats for AI-driven power control implementations in massive MIMO systems [1]. To give another example, the authors in [2] have shown that threats against target AI model in MIMO systems which might be originated from malicious UEs can substantially decrease the SE performance by applying a successful adversarial sample under different scenarios. And, it has been shown that the risk associated with these kinds of adversarial attacks are bigger than the standard attack threats. In a real world scenario, the success of an adversarial attack in a D-MIMO network is constrained by three key factors from the adversary's perspective. Firstly, the adversary mostly cannot use the original AI model as in the case of whitebox setting to craft adversarial samples due to lack of access to target model's architecture and weights. Secondly, the attacker cannot have complete knowledge of the input features of the AI model, as it is almost impossible to know the channel information of each UE. Last but not least, the adversary will not be able to introduce perturbations to all features of the input vector, even if the channel information is known beforehand. However, despite all these constraints, there are established strategies in literature that boost the attacker's success [8]. In this study, to examine and show the efficacy of our proposed method, we chose the worst-case scenario with the most disastrous possible outcome, in which the adversary has access to (read/modify) each element of the input vector fed to the AI model in CP together with the details of the target model.

During normal operation, we feed the channel information $\boldsymbol{\beta}$ obtained from UEs to the AI model, and this model outputs a nearly optimal $\boldsymbol{\eta}_{\text{AI}}$ value which maximizes the sum of the SEs of all UEs as previously shown in Fig. 3. However, during a malicious activity, an adversary can add a well-crafted perturbation to the channel input $\boldsymbol{\beta}$ to generate the perturbed input $\boldsymbol{\beta} + \boldsymbol{\Delta}$ yielding a power allocation $\boldsymbol{\eta}'_{\text{AI}}$ resulting in decreased SEs. This malicious scenario is depicted in Fig. 4.

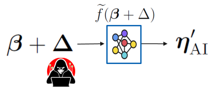
$\widetilde{f}(\boldsymbol{\beta} + \Delta)$
$\boldsymbol{\beta} + \boldsymbol{\Delta}$
$\boldsymbol{\eta}'_{\text{AI}}$

Fig. 4: Adversarial attack on the AI solution.

Theoretically, the attacker can produce any kind of perturbation $\boldsymbol{\Delta}$ including randomly generated simple solutions. In the adversarial attack scenario, the objective of the attacker is to use $\partial \text{SE}_{\text{AI, sum}} / \partial \boldsymbol{\beta}$ to minimize SEs, where $\text{SE}_{\text{AI, sum}}$ is the sum of elements of $\textbf{SE}_{\text{AI}}$. If the adversary has access to the target AI model architecture and weights, it can use this original AI model to find the gradient of the output sum rate with respect to $\boldsymbol{\beta}$ values. Then, the adversary adds a perturbation in the opposite direction of the gradient information to minimize SEs which is widely known as Fast Gradient Sign Method (FGSM) [9] in literature. If the adversary does not have the original AI model, it can use a surrogate AI model to craft the adversarial perturbation in a similar manner. Later, Kurakin et al. [10] proposed a small but effective improvement to the FGSM, known as Basic Iterative Method (BIM). In this approach, rather than taking only one step of size ϵ in the gradient sign's direction, the attacker takes several but smaller steps, and use the given ϵ value to clip the result. For our case, the BIM algorithm can be given by Algorithm 1.

In Algorithm 1, $\boldsymbol{\Delta}_i$ is the crafted adversarial sample (to be added to the input of the victim) at the i^{th} iteration, $i_{\text{max},1}$ is the maximum number of iterations, ϵ is a tunable parameter, limiting maximum level of perturbation for L_∞ norm, α_1 is the step size, and $\text{clip}_\epsilon\{\cdot\}$ is the clipping operator that clips entries of the argument larger than ϵ to ϵ and less than $-\epsilon$ to $-\epsilon$, and $\text{sign}(\cdot)$ is the sign operator mapping positive entries to 1, negative entries to -1, and zero entries to 0.

Algorithm 1: BIM attack algorithm.

```
Input: $\boldsymbol{\beta}, \widetilde{f}(\cdot), i_{\text{max},1}, \epsilon, \alpha_1$
Output: $\boldsymbol{\Delta}$
$\boldsymbol{\Delta}_0 = \mathbf{0},\ i = 0$.
while $i < i_{max,1}$ do
    Compute the AI output: $\boldsymbol{\eta}'_{\text{AI}} = \widetilde{f}(\boldsymbol{\beta} + \boldsymbol{\Delta}_i)$.
    Compute the SE vector: $\text{SE}'_{\text{AI}} = q(\boldsymbol{\beta}, \boldsymbol{\eta}'_{\text{AI}})$.
    Compute the sum of SEs: $\text{SE}'_{\text{AI, sum}}$ is equal to the
      sum of the elements of $\text{SE}'_{\text{AI}}$.
    Update the perturbation vector:
      $\boldsymbol{\Delta}_{i+1} = \text{clip}_\epsilon \left( \boldsymbol{\Delta}_i - \alpha_1 \cdot \text{sign}(\partial \text{SE}'_{\text{AI, sum}} / \partial \boldsymbol{\beta}) \right)$.
    Increase $i$ by 1.
return $\boldsymbol{\Delta} = \boldsymbol{\Delta}_i$.
```

IV. Defense Against Adversarial Attacks

The adversarial attacks substantially degrades the performance of AI-powered solutions. In Fig. 5, we present the cumulative distribution function (CDF) of per-user SEs for adversarial attack (BIM) and no-attack cases where AI solution is applied in both cases. We observe that the SE performance under adversarial attack significantly degrades. We observe a performance loss about 0.24 bps/Hz in the median, and 0.46 bps/Hz in the 5th percentile[1] compared to no attack case. The results show that smart defense methods are required to mitigate the effects of these threats.

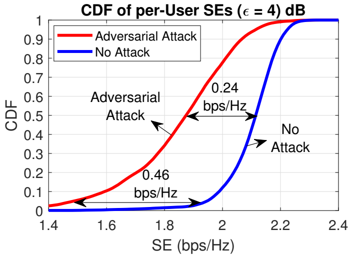

Fig. 5: Effect of adversarial attack on the AI solution of (P1). $M = 16, K = 4, \epsilon = 4$ dB.

A. Problems With Existing Defense Solutions

To mitigate adversarial attacks, several defense solutions have been proposed in the literature, with adversarial training being one of the most dominant approach. In this technique, training of the AI model is done by augmenting training data with adversarial samples in an aim to make the model robust to such kind of inputs during inference time. However, adversarial training type of defense approaches cannot work for most of the regression tasks/models.

In Fig. 6, we provide the CDF plots of per-user SEs for the analytical solution. Under the adversarial attack scenario, the performance of the analytical solution (ground truth) also degrades, showing that adversarial training is not suitable for this regression task. Because applying perturbation to channel

[1] The SE values corresponding to CDF $= 0.05$.

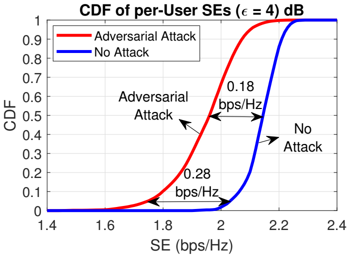

Fig. 6: Effect of adversarial attack on the analytical solution of (P1). $M = 16, K = 4, \epsilon = 4$ dB.

information vector results into a completely different optimum power allocation values from the ones in no attack case. This observation proves that applying adversarial training will degrade the natural (clean) performance of the system for these kinds of complex regression models.

Adversarial training generally works in image classification tasks where the input of the AI model has some form of semantic integrity. For example, the right side image in Fig. 7 is misclassified as a 'sports car' by a state-of-the-art image classifier model. However, for us as human beings, the left and right images are all interpreted as dog. We do not actually change the ground truth label of this image when we apply an adversarial perturbation. In other words, the exact solution is the same for both the clean image and adversarial image. Therefore, we could actually use the perturbed image sample in the right side of Fig. 7 for adversarial training.

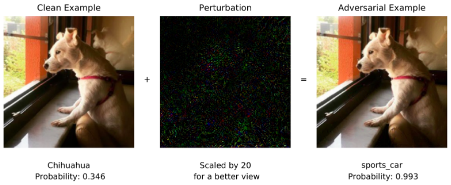

Fig. 7: An example adversarial attack on image domain.

However, in our regression task modeled by the analytical function $f(.)$ that maps input $\boldsymbol{\beta}$ to output $\boldsymbol{\eta}$, when we apply a perturbation to input $\boldsymbol{\beta}$ and get $\boldsymbol{\beta}'$, $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\beta}')$ will not be equal and hence we cannot train our model with $(\boldsymbol{\beta}', \boldsymbol{\eta})$ pairs anymore.

B. Proposed Defense Method

To mitigate the effects of adversarial attacks, we suggest a proactive defense approach. We propose a method in which we try to find an imaginary perturbation $\boldsymbol{\delta}$ which yields a better power allocation output by maximizing the user SEs.

During the life-cycle of the solution of (P1), a well-trained AI model might be used to find the near-optimum power allocation for a given channel input. The AI model is not aware of any possible malicious activity and cannot understand whether any adversarial perturbation is added to the input $\boldsymbol{\beta}$. Therefore, without any defense mechanism it finds the output power coefficients using the input $\boldsymbol{\beta}' = \boldsymbol{\beta} + \boldsymbol{\Delta}$ as shown in Fig. 3. When there is no attack, $\boldsymbol{\Delta} = \mathbf{0}$, and if there is an attack on the input, then $\boldsymbol{\Delta}$ becomes a non-zero vector.

We suggest finding a virtual perturbation $\boldsymbol{\delta}$ which yields a better power allocation. By injecting a virtual perturbation, the method improves the output user SEs of the AI solution by optimizing the sum of the user SEs, $\text{SE}_{\text{AI, sum}}$. Instead of using $\boldsymbol{\beta}'$ to predict the output, we propose to iteratively compute the derivative of our objective function $\text{SE}_{\text{AI, sum}}$ with respect to $\boldsymbol{\beta}'$ and add the resulting accumulated gradient vector $\boldsymbol{\delta}$ to $\boldsymbol{\beta}'$ to predict the output. We compute the virtual perturbation vector $\boldsymbol{\delta}$ iteratively. At each iteration, we compute the gradient $\partial \text{SE}_{\text{AI, sum}} / \partial \boldsymbol{\beta}'$ via the AI model and update the virtual perturbation vector considering the direction of the gradient so that the $\text{SE}_{\text{AI, sum}}$ increases. After a finite number of steps, the algorithm converges, and we find the final $\boldsymbol{\delta}$ vector. The details of our proposed method is given in Algorithm 2.

Algorithm 2: The proposed defense algorithm.

```
Input: $\boldsymbol{\beta}', \widetilde{f}(\cdot), i_{\max,2}, \epsilon, \alpha_2$
Output: $\boldsymbol{\delta}$
$\boldsymbol{\delta}_0 = \mathbf{0},\ i = 0$.
while $i < i_{max,2}$ do
    Compute the AI output: $\boldsymbol{\eta}''_{\text{AI}} = \widetilde{f}(\boldsymbol{\beta}' + \boldsymbol{\delta}_i)$.
    Compute the SE vector: $\text{SE}''_{\text{AI}} = q(\boldsymbol{\beta}', \boldsymbol{\eta}''_{\text{AI}})$.
    Compute the sum of SEs: $\text{SE}''_{\text{AI, sum}}$ is equal to the
      sum of the elements of $\text{SE}''_{\text{AI}}$.
    Update the virtual perturbation vector:
      $\boldsymbol{\delta}_{i+1} = \text{clip}_{\epsilon}\left(\boldsymbol{\delta}_i + \alpha_2 \cdot \text{sign}(\partial \text{SE}''_{\text{AI, sum}} / \partial \boldsymbol{\beta}')\right)$.
    Check if the algorithm is converged: If the
      difference between $\boldsymbol{\delta}_{i+1}$ and $\boldsymbol{\delta}_i$ is low enough,
      terminate. Otherwise increase $i$ by 1.
return $\boldsymbol{\delta} = \boldsymbol{\delta}_i$.
```

In Algorithm 2, $\boldsymbol{\beta}'$ is the input (possible perturbed by the attacker) channel information vector, $\widetilde{f}(\cdot)$ is the AI model used to solve (P1), $i_{\max,2}$ is the maximum number of iterations, α_2 is the step-size, and $\boldsymbol{\delta}$ is the output virtual perturbation.

It is important to note that the attacker uses the gradient of the sum of SEs with respect to the actual channel vector $\boldsymbol{\beta}$ and tries to manipulate the target AI model output via providing $\boldsymbol{\beta}' = \boldsymbol{\beta} + \boldsymbol{\Delta}$. On the other hand, the original unperturbed input $\boldsymbol{\beta}$ is not available for the defense method and hence it uses the potentially perturbed channel vector $\boldsymbol{\beta}'$ as the input and hence the gradient is evaluated with respect to $\boldsymbol{\beta}'$. When there is no attack, both gradients coincide; however, when there is an attack, they differ. Our numerical results show that although we might not use the actual channel vector when calculating the gradient, we observe a significant benefit under adversarial attack scenario. It should be noted that the gradient $\partial \text{SE}_{\text{AI, sum}} / \partial \boldsymbol{\beta}'$ depends on the AI model as $\text{SE}_{\text{AI, sum}}$ is a function of $\boldsymbol{\eta}''_{\text{AI}}$ which is the output of the AI model. Therefore, the gradient cannot be directly evaluated using an analytical function. We evaluate it via the AI model using the structure of the model. The calculation is efficient because the calculation of SEs using the input and output of the AI model by the function q is a low-complexity operation.

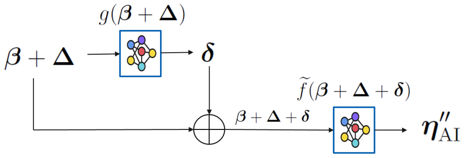
$g(\boldsymbol{\beta}+\boldsymbol{\Delta})$
$\boldsymbol{\beta}+\boldsymbol{\Delta}$
$\boldsymbol{\delta}$
$\widetilde{f}(\boldsymbol{\beta}+\boldsymbol{\Delta}+\boldsymbol{\delta})$
$\boldsymbol{\beta}+\boldsymbol{\Delta}+\boldsymbol{\delta}$
$\boldsymbol{\eta}''_{\text{AI}}$

Fig. 8: The operation for proposed defense technique

Fig. 8 shows the diagram of the proposed defense method. The function $g(\cdot)$ shows the defense method that evaluates the virtual perturbation using the perturbed input as described in Algorithm 1. After finding the virtual perturbation, we add it to the perturbed input and run the AI model to generate the power allocation output. The algorithm aims to eliminate the negative effect of the unknown perturbation $\boldsymbol{\Delta}$ by designing a virtual perturbation $\boldsymbol{\delta}$ so that $\boldsymbol{\Delta}+\boldsymbol{\delta}\approx\mathbf{0}$.

V. Simulation Results

In this section, we present the effectiveness of the proposed defense technique by detailed simulations. For this purpose, we uniformly distribute $M = 16$ RUs and $K = 4$ UEs in a 500 m x 500 m square region. The simulation parameters are chosen as given in Table I.

TABLE I: Simulation Parameters

Parameter	Model and/or Value
Carrier frequency, bandwidth	1.9 GHz, 20 MHz
Path-loss	$[\text{PL}_{m,k}]_{\text{dB}}$ is chosen according to the 3-slope model in [4]
Shadowing	$[\beta_{m,k,\text{sh}}]_{\text{dB}} \sim \mathcal{N}(0,8^2)$
Channel model	$h_{m,k} \sim \mathcal{CN}(0,\beta_{m,k})$ where $[\beta_{m,k}]_{\text{dB}} = [\text{PL}_{m,k}]_{\text{dB}} + [\beta_{m,k,\text{sh}}]_{\text{dB}}$
Average RU transmit power, UE receiver noise power	$P_t = 0.2$ W, $\sigma_k^2 = -92$ dBm, $\forall k$

We use a DNN-based regression model to learn the mapping between the large-scale fading coefficient vector $\boldsymbol{\beta} = [\beta_{1,1}\, \beta_{1,2}\, \dots\, \beta_{M,K}]^T \in \mathbb{R}^{MK}$ and the power control coefficient vector $\boldsymbol{\eta} = [\eta_{1,1}\, \eta_{1,2}\, \dots\, \eta_{M,K}]^T \in \mathbb{R}^{MK}$. We train a model which we denote as $\widetilde{f}(\cdot)$ representing the AI model used in D-MIMO system and we assume the adversary has whitebox access to the AI model to craft adversarial samples. The detailed model architecture is given in Fig. 9.

In Fig. 10, we present the per-user CDFs for the AI model (Fig. 3) together with the result of analytical solution (Fig. 2). We observe 0.03 bps/Hz and 0.08 bps/Hz gaps for the median and 5th percentile SEs, respectively. The results show that the AI model can accurately approximate the analytical solution.

In the simulations, we compare 3 different methods under 3 different scenarios which are listed in Table II.

Notice that we always evaluate the SEs using the actual channel vectors even if the input is perturbed by the attacker. This is because the real performance of the system depends on the actual channel. When there is an attack on the input, the victim system cannot evaluate the actual SEs as the original channel information will not be available. On the other hand, the computation of the actual SEs is not required to find power allocation. We evaluate the actual SEs in simulations to see the performance of the system with and without attack.

TABLE II: Scenarios and Methods

Scenario 1	No attack where the system has the true channel vectors
Scenario 2	White Gaussian Noise (WGN) attack where the attacker injects zero-mean, ϵ variance WGN to the original channel samples
Scenario 3	Adversarial attack where the attacker applies BIM
Method 1	AI solution without any defense
Method 2	Analytical solution without any defense
Method 3	AI solution with the proposed defense method

Fig. 11 involves CDF of UEs for all 9 cases obtained by all combinations of Methods 1-3 and Scenarios 1-3 for $\epsilon = 4$ dB. We observe that the proposed defense technique significantly enhances the performance for UEs with low SE values. With the proposed technique, 0.28 bps/Hz enhancement on 5th percentile per-user SEs is obtained. Considering that the remaining gap to AI solution without any attack is equal to 0.19 bps/Hz, it can be concluded that more than half of the performance loss due to adversarial attack can be regained by the proposed defense method. When there is no attack or there exists a WGN attack, a slight improvement is obtained over AI solution without any defense. This shows that the proposed method can be used without any performance loss regardless of the attack situation. The proposed defense method is effective against adversarial attacks which are the most disruptive ones for AI-based power allocation methods.

To measure the robustness of the proposed defense method, we define a metric $r_{\text{robustness}}$ as

$$r_{\text{robustness}} = \frac{\text{SE}_{\text{defense}} - \text{SE}_{\text{attack}}}{\text{SE}_{\text{no-attack}} - \text{SE}_{\text{attack}}}, \tag{7}$$

where $\text{SE}_{\text{no-attack}}, \text{SE}_{\text{attack}}, \text{SE}_{\text{defense}}$ indicates 5th percentile per-user SEs of AI solution without attack and defense, under adversarial attack without defense, and under adversarial attack with the proposed defense method, respectively. $r_{\text{robustness}}$ shows the ratio of performance that can be regained by the defense method for UEs with low SE values.

In Fig. 12, we present the robustness ratios for various ϵ values. We see that the robustness ratio values are larger than 40 percent for all $\epsilon \leq 8$ dB. The robustness ratio is a decreasing function of ϵ as it becomes harder to compensate the effect of adversarial attacks as the perturbation magnitude increases. Notice that for ϵ values larger than 8 dB, which is the shadowing standard deviation, the system might detect the existence of an attack by observing unnatural changes in the channel coefficients, and can perform different actions.

Fig. 13 shows average runtimes of AI, proposed defense, and analytical solutions under adversarial attack. As expected, the proposed defense method is slower than the AI solution as it runs the AI function $\widetilde{f}$ several times to find a virtual perturbation. On the other hand, it is an efficient defense technique as it is roughly 6 times faster than the analytical solution.

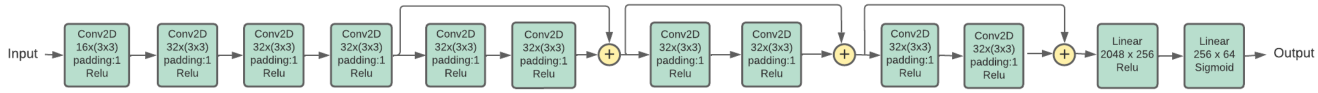
Input
Conv2D 16x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
+
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
+
Conv2D 32x(3x3) padding:1 Relu
Conv2D 32x(3x3) padding:1 Relu
+
Linear 2048 x 256 Relu
Linear 256 x 64 Sigmoid
Output

Fig. 9: AI Model

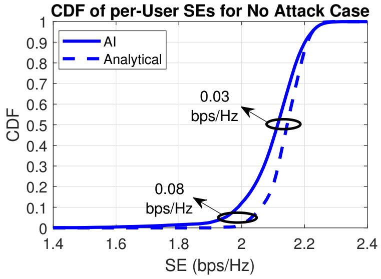

Fig. 10: Comparison of AI and analytical solutions of (P1) for $M = 16, K = 4$.

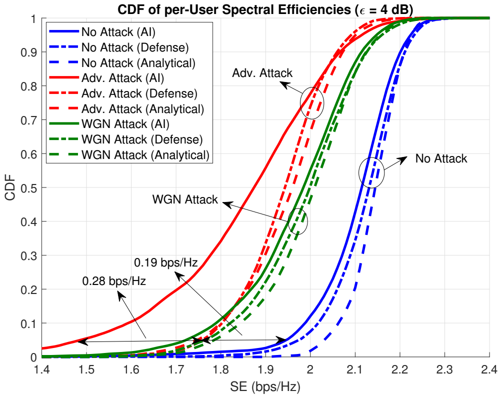

Fig. 11: CDF of per-user SEs for various cases ($\epsilon = 4$ dB).

VI. Conclusion

In this paper, we proposed a novel method to increase the robustness of the AI-driven power allocation in D-MIMO against adversarial attacks. We empirically shown the effectiveness of the proposed approach and verified that it significantly increases the spectral efficiency of the UEs in the presence of adversarial inputs from malicious parties. The important thing to note is that the proposed method does not have any negative effect on the natural (clean) SE performance of the system when there is no attack and yields superior performance than AI solution by increasing AI model's SE performance towards the optimum solution. Furthermore, its complexity is much lower than that of the analytical solution. These results demonstrates that the proposed solution can be safely and efficiently implemented to increase robustness of the system without sacrificing normal performance. As a future work, we plan to elaborate on whether our proposed method is applicable to other AI-driven tasks in D-MIMO such as codebook-based beamforming or RU selection.

Acknowledgment

This work was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) through the 1515 Frontier Research and Development Laboratories Support Program under Project 5169902, and has been partly funded by the European Commission through the H2020 project Hexa-X (Grant Agreement no. 101015956).

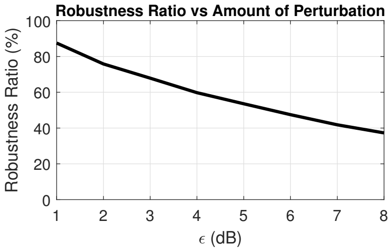

Fig. 12: Robustness ratios for various ϵ values.

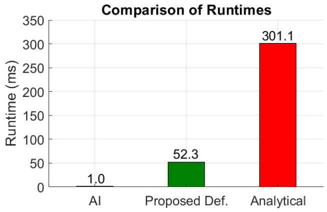

Fig. 13: Comparison of runtimes of the three solutions.

References

[1] P. M. Santos, B. R. Manoj, M. Sadeghi, and E. G. Larsson, "Universal adversarial attacks on neural networks for power allocation in a massive mimo system," *IEEE Wireless Commun. Lett.*, vol. 11, no. 1, pp. 67–71, 2022.

[2] B. Kim, Y. Shi, Y. E. Sagduyu, T. Erpek, and S. Ulukus, "Adversarial attacks against deep learning based power control in wireless communications," in *2021 IEEE Globecom Workshops*, 2021, pp. 1–6.

[3] Ö. T. Demir, E. Björnson, L. Sanguinetti *et al.*, "Foundations of user-centric cell-free massive mimo," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021.

[4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive mimo versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017.

[5] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive mimo systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, 2017.

[6] F. E. Kadan and O. Haliloğlu, "A performance bound for maximal ratio transmission in distributed mimo," *IEEE Wireless Communications Letters*, vol. 12, no. 4, pp. 585–589, 2023.

[7] N. Rajapaksha, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-Aho, "Deep learning-based power control for cell-free massive mimo networks," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–7.

[8] O. F. Tuna, F. E. Kadan, and L. Karaçay, "Practical adversarial attacks against ai-driven power allocation in a distributed mimo network," 2023. [Online]. Available: https://arxiv.org/abs/2301.09305

[9] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017. [Online]. Available: https://arxiv.org/abs/1607.02533