Adversarially Robust Hypothesis Testing

Yulu Jin and Lifeng Lai
Department of Electrical and Computer Engineering
University of California, Davis
{yuljin,lflai}@ucdavis.edu

***Abstract*—In this paper, we investigate the adversarial robustness of classification problems. In the considered model, after a sample is generated, it will be modified by an adversary before being observed by the classifier. The classifier needs to decide the underlying hypothesis that generates the sample from the adversarially modified data. We formulate this problem as a minimax hypothesis testing problem, in which the goal of the adversary is to design attack strategy to maximize the error probability while the decision maker aims to design decision rule so as to minimize the error probability. We solve this minimax problem and characterize the corresponding optimal strategies.**

***Index Terms*—minimax problem, hypothesis testing, adversarial robustness**

I. Introduction

Even though neural networks have many applications, they are not robust to adversarial attacks [1]. By adding hardly perceptible perturbations on the input data, the decision of a deep network can be easily manuplicated. Many follow up works design attack algorithms to find adversarial examples more efficiently [2]–[5]. At the same time, there are significant amount of research works that focus on developing defense strategies with the goal of constructing robust classifiers that can work well in the presence of adversarial perturbations [6]–[8]. Unfortunately, most of these defense strategies are quickly defeated by new attack methods. This phenomenon motivates many studies aiming to establish the fundamental limits on the robustness of classifiers [9] [10]. Most of these works rely on tools from concentration of measure [11] and provide interesting results when the dimension of data is high and the distribution of data satisfies certain conditions.

The goal of this paper is to understand the fundamental limits of classification under adversarial attacks from decision theoretical perspective, regardless the dimension and distribution of data. In particular, we formulate the classification problem with adversarial perturbations as a minimax hypothesis testing problem, in which the goal of the adversary is to design attack strategy to modify the data so as to maximize the error probability while the decision maker aims at designing decision rule to minimize the error probability. Our work is related to but different from the large volume of work on classic robust statistics [12]–[14]. The classic robust statistical inference mainly focuses on distributional robustness, in which the true distributions of data lie in the neighborhood of nominal distributions [15], [16]. However, these distributional robust frameworks cannot properly address the adversarial perturbations scenario. On the adversarial perturbation scenario, when sampled data is fully available for an adversary, the attacker could be much more powerful [17].

In this work, we use the maximal allowed attack amplitude to model the strength of an adversary. By restricting the strength of attack vectors, we first show that the formulated minimax problem has a saddle point solution. From this saddle point solution, we can characterize the structure of the optimal attack and defense strategies. In particular, the optimal defense strategy is to perform the Bayesian test on the corresponding probability mass functions (PMFs) after the attack. As the result, we can write the cost function as a function of the attack strategy only, and characterizing the optimal attack strategy is equivalent to solving a maximization problem over the attack strategy. However, the resulting maximization problem is a very complex non-convex optimization problem. In this paper, we solve this problem for a special case where the optimal Bayesian decision regions corresponding to the PMFs before attack consist of two consecutive regions. Using this special structure, we first relax certain constraints and construct a series of upper bounds for the error probability. We then check the achievability of each bound when the constraints are added back and find the maximum value of the error probability when all the constraints are satisfied. Even though the maximum error probability is unique, the attack strategy that achieves this maximum error probability is not unique. In our paper, we further design an efficient algorithm to characterize one of the best attack strategies. We also provide numerical examples to illustrate the analytical results obtained in this paper.

The remainder of this paper is organized as follows. In Section II, we present our problem formulation. In Section III, we analyze the saddle point of the minimax optimization problem to obtain the structure of the solution. In Section IV, we characterize the optimal attack and decision strategies. In Section V, we provide numerical examples to illustrate the analytical results. In Section VI, we offer concluding remarks. Due to space limitations, we omit details of proof.

II. Problem Formulation

Consider a discrete random variable X defined on a finite set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, where under $\mathcal{H}_k$, with $k = 0, 1$, the probability mass function(PMF) of X is $\boldsymbol{p}_k$, where $p_{k,j} = \Pr(X = x_j|\mathcal{H}_k)$.

Y. Jin and L. Lai are with the Department of Electrical and Computer Engineering, University of California, Davis, CA. The work of Y. Jin and L. Lai was supported by the National Science Foundation under Grants CCF-1717943, ECCS-1711468, CNS-1824553 and CCF-1908258.

We consider a powerful hypothesis-aware adversary who can conduct randomized attacks. In particular, the adversary can change sample $X = x_i$ to an attacked sample $X' = x_j$ with $1 \leq i, j \leq n$. Denote the attack rule as $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}$, where the adversary performs $\boldsymbol{A}$ under $\mathcal{H}_0$ and $\boldsymbol{B}$ under $\mathcal{H}_1$. The components of $\boldsymbol{A}$ are $a_{i,j} = \Pr(X' = x_j | X = x_i, \mathcal{H}_0)$. The components of $\boldsymbol{B}$ are $b_{i,j} = \Pr(X' = x_j | X = x_i, \mathcal{H}_1)$. Motivated by adversarial examples in neural network, we assume that the change introduced by the adversary has limited amplitude δ. Under the attack rule $(\boldsymbol{A}, \boldsymbol{B})$, the PMF of X' is $\boldsymbol{q}_0 = \boldsymbol{p}_0 \boldsymbol{A}$ under $\mathcal{H}_0$ and $\boldsymbol{q}_1 = \boldsymbol{p}_1 \boldsymbol{B}$ under $\mathcal{H}_1$, where $q_{k,j} = \Pr(X' = x_j | \mathcal{H}_k)$, with $k = 0, 1$.

Let $\mathcal{T} = [0, 1]^n$ be the set of all decision rules, denote $\boldsymbol{t} = [t_1, \cdots, t_n] \in \mathcal{T}$ as a randomized decision rule such that if $X = x_i$, the detector selects $\mathcal{H}_1$ with probability t_i, where $0 \leq t_i \leq 1$.

Assuming that the prior probability of two hypotheses are equal, i.e., $\Pr(\mathcal{H}_0) = \Pr(\mathcal{H}_1)$, for decision rule $\boldsymbol{t}$, the error probability P_E can be written as

$$P_E(\boldsymbol{p}_0, \boldsymbol{p}_1, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) = \frac{1}{2}[P_F(\boldsymbol{p}_0, \boldsymbol{A}, \boldsymbol{t}) + P_M(\boldsymbol{p}_1, \boldsymbol{B}, \boldsymbol{t})]. \quad (1)$$

In the following, to simplify the notation, we will drop $\boldsymbol{p}_0, \boldsymbol{p}_1$ from the expression of P_E and will simply write it as $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$.

The goal of the attacker is to choose the attack rule $(\boldsymbol{A}, \boldsymbol{B})$ to maximize the error probability (1), while the goal of the defender is to choose the decision rule $\boldsymbol{t}$ to minimize the error probability (1). In this paper, we seek to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ and $\boldsymbol{t}^*$ by solving the minimax problem

$$\min_{\boldsymbol{t} \in \mathcal{T}} \max_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \quad (2)$$

III. Saddle Point Analysis

In this section, we characterize the structure of the optimal decision rules by analyzing the saddle point of the minimax problem (2).

Note that $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t})$ is continuous and linear, and therefore is both convex and concave in $(\boldsymbol{A}, \boldsymbol{B})$ and $\boldsymbol{t}$ respectively. Furthermore, sets $\mathcal{A} \times \mathcal{B}$ and $\mathcal{T}$ are both compact and convex. Therefore, using Von Neumann minimax theorem [18], we have

$$\begin{aligned} &\min_{\boldsymbol{t} \in \mathcal{T}} \max_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}) \\ &= \max_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{A} \times \mathcal{B}} \min_{\boldsymbol{t} \in \mathcal{T}} P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}). \end{aligned} \quad (3)$$

This implies that the solution $(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*)$ to this minimax problem satisfies the saddle point property

$$P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}) \geq P_E(\boldsymbol{A}^*, \boldsymbol{B}^*, \boldsymbol{t}^*) \geq P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t}^*). \quad (4)$$

From these two inequalities, we can characterize the structure of the optimal attack and decision strategies.

The first inequality in (4) indicates that the best decision rule must be the Bayesian test with respect to the best adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$. It is well known that, for a given arbitrary adversary attack rule $(\boldsymbol{A}, \boldsymbol{B})$, the optimal detection rule, denoted as $\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B})$, is simply a threshold rule [19]:

$$t_i^*(\boldsymbol{A}, \boldsymbol{B}) = \begin{cases} 0 & q_{0,i} > q_{1,i}, \\ \text{arbitrary} & q_{0,i} = q_{1,i}, \\ 1 & q_{0,i} < q_{1,i}. \end{cases} \quad (5)$$

For the optimal adversary $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, the optimal decision rule is $\boldsymbol{t}^* = \boldsymbol{t}^*(\boldsymbol{A}^*, \boldsymbol{B}^*)$.

Then we can use the second inequality in (4) to characterize the optimal $(\boldsymbol{A}^*, \boldsymbol{B}^*)$ by solving

$$\max_{\boldsymbol{A}, \boldsymbol{B}} \quad \frac{1}{2}[\boldsymbol{p}_0 \boldsymbol{A} (\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T + \boldsymbol{p}_1 \boldsymbol{B}(1 - (\boldsymbol{t}^*(\boldsymbol{A}, \boldsymbol{B}))^T)], (6)$$

$$a_{i,j} \geq 0, b_{i,j} \geq 0, i, j = 1, .., n, \quad (7)$$

$$\sum_{j=1}^{n} a_{i,j} = 1, \sum_{j=1}^{n} b_{i,j} = 1, i = 1, .., n, \quad (8)$$

$$1_{|i-j|>\delta} a_{i,j} = 1_{|i-j|>\delta} b_{i,j} = 0, i, j = 1, .., n. \quad (9)$$

Here, constraints (7) and (8) guarantee that each row of $\boldsymbol{A}$ and $\boldsymbol{B}$ is a conditional PMF, while constraint (9) makes sure that the changes introduced by the attacker have a limited range.

Once we solve (6) and obtain $(\boldsymbol{A}^*, \boldsymbol{B}^*)$, we can use (5) to obtain the optimal $t^*(\boldsymbol{A}^*, \boldsymbol{B}^*)$.

IV. Derivation of the optimal adversary

In this section, we characterize the optimal solution to the complicated optimization problem in (6) under certain assumptions on $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$. Let $R_0 = \{i | p_{0,i} \geq p_{1,i}\}$ and $R_1 = \{i | p_{0,i} \leq p_{1,i}\}$, i.e., R_0 is the set of index where $p_{0,i}$ is larger while R_1 is the set of index where $p_{1,i}$ is larger. In this section, we assume that R_0 (and hence R_1) is a consecutive region in $[1, n]$. Without loss of generality, we write $R_0 = \{i | 1 \leq i \leq m\}$ and $R_1 = \{i | m + 1 \leq i \leq n\}$.

We now compare this assumption with the assumptions in [16], which assumes that the original PMFs satisfy certain monotonicity and symmetry conditions. Specifically, monotonicity means that $p_{1,i}/p_{0,i}$ is a monotonically increasing function of i and symmetry implies $p_{1,n-i+1} = p_{0,i}, 1 \leq i \leq n$. It is easy to check that the monotonicity assumption implies the assumption made in this paper. Moreover, the symmetry condition is not required here. Hence, our assumption is significantly weaker than the assumptions in [16].

We first present a lemma that simplifies $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*})$.

Lemma 1: $P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*})$ can be written as

$$P_E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{t^*}) \quad = \quad \frac{1}{2} \sum_{i=1}^{n} \min\{q_{0,i}, q_{1,i}\}. \quad (10)$$

To proceed further, we denote the mass moved into region $[1, i]$ as $I_{0,i}$ under $\mathcal{H}_0$ and $I_{1,i}$ under $\mathcal{H}_1$. Similarly, define the mass moved out from $[1, i]$ as $O_{0,i}$ under $\mathcal{H}_0$ and $O_{1,i}$ under $\mathcal{H}_1$ as shown in Figure 1.

Lemma 2: The relationship between PMFs $\boldsymbol{p}_0$, $\boldsymbol{p}_1$ before attack and $\boldsymbol{q}_0$, $\boldsymbol{q}_1$ after attack can be represented by the moved mass defined above:

$$O_{0,i-1} + I_{0,i} + p_{0,i} \quad = \quad O_{0,i} + I_{0,i-1} + q_{0,i}, \forall i, \quad (11)$$

$$O_{1,i-1} + I_{1,i} + p_{1,i} \quad = \quad O_{1,i} + I_{1,i-1} + q_{1,i}, \forall i. \quad (12)$$

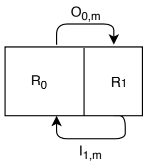

Fig. 1. Mass moved between two regions

The basic idea behind (11) and (12) is that the mass is conserved. For any component i, the sum of original mass and moved in part always equal to the sum of remaining mass and moved out part.

Define $D(\boldsymbol{A},\boldsymbol{B})$ as

$$D(\boldsymbol{A},\boldsymbol{B}) = P_E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{t}^*) - \frac{1}{2}\left(\sum_{i\in R_0} p_{1,i} + \sum_{i\in R_1} p_{0,i}\right), \tag{13}$$

It is easy to see that maximizing $P_E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{t}^*)$ is same as maximizing $D(\boldsymbol{A},\boldsymbol{B})$.

Proposition 1: $D(\boldsymbol{A},\boldsymbol{B})$ is upper bounded by the sum of mass moved between two regions:

$$\begin{aligned} 2D(\boldsymbol{A},\boldsymbol{B}) &\overset{(a)}{\leq} \sum_{i=1}^{m}[\min(q_{0,i},q_{1,i}) - p_{1,i}] + O_{0,m} \\ &\overset{(b)}{\leq} I_{1,m} + O_{0,m}, \end{aligned} \tag{14}$$

in which the equality in (a) holds when

$$\begin{aligned} q_{1,i} &\geq q_{0,i}, \forall i \in R_1, & (15)\\ I_{0,m} &= 0, & (16) \end{aligned}$$

and the equality in (b) holds when

$$\begin{aligned} q_{0,i} &\geq q_{1,i}, \forall i \in R_0, & (17)\\ O_{1,m} &= 0. & (18) \end{aligned}$$

Our approach to find the maximum value of $2D(\boldsymbol{A},\boldsymbol{B})$ consists of two major steps. In the first step, we maximize the upper bound $I_{1,m}+O_{0,m}$ under constraints (16) (17) and (18). This will result in a maximum value of $I_{1,m}+O_{0,m}$, denoted as F_m^*, and a new domain set $(\mathcal{A}_m^*,\mathcal{B}_m^*)$ containing all attack strategies that achieve F_m^*. In the second step, we will show that the optimal attack strategy is in $(\mathcal{A}_m^*,\mathcal{B}_m^*)$ and provide an algorithm to find such strategy.

A. *Step 1: Maximizing $I_{1,m}+O_{0,m}$*

First, we have

$$\begin{aligned} &I_{1,m} + O_{0,m} \\ &\overset{(m)}{\leq} \quad I_{1,m-1} + O_{0,m-1} + p_{0,m} - p_{1,m} \\ &\quad\ \cdots \\ &\overset{(i)}{\leq} \quad I_{1,i-1} + O_{0,i-1} + \sum_{j=i}^{m}(p_{0,j} - p_{1,j}) \\ &\quad\ \cdots \\ &\overset{(1)}{\leq} \quad \sum_{j=1}^{m}(p_{0,j} - p_{1,j}), \end{aligned} \tag{19}$$

in which inequality (i), $1 \leq i \leq m$, corresponds to the constraint $q_{1,i} \leq q_{0,i}$. To achieve the equality in (i), we need to satisfy $O_{1,i-1} = 0$, $I_{0,i-1} = 0$ and $q_{1,i} = q_{0,i}$. In general, to achieve all the equalities in (19), the following conditions need to be satisfied

$$I_{0,i} = 0, \quad O_{1,i} = 0, \quad 1 \leq i \leq m-1, \tag{20}$$

$$q_{0,i} = q_{1,i}, \quad 1 \leq i \leq m. \tag{21}$$

Condition (20) can always be satisfied if we restrict the mass moving directions, i.e., $(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}_c,\mathcal{B}_c) = \{(\boldsymbol{A},\boldsymbol{B})|a_{j,i} = 0, 1 \leq i < j \leq m, b_{j,i} = 0, 1 \leq j < i \leq m\}$. In the following sections, $(\boldsymbol{A},\boldsymbol{B})$ is limited to the set $(\mathcal{A}_c,\mathcal{B}_c) \equiv (\mathcal{A}_0^*,\mathcal{B}_0^*)$. Hence, to achieve the equalities in (19), we only need to check (21).

Starting from $\overset{(1)}{\leq}$ in (19) and going upwards, we check the achievability of each equality. In step one, there always exist some attack strategies $(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}_c,\mathcal{B}_c)$, such that $q_{0,1} = q_{1,1}$. Hence, the equality holds and the feasible set is $(\mathcal{A}_1^*,\mathcal{B}_1^*) = \{(\boldsymbol{A},\boldsymbol{B})|q_{0,1} = q_{1,1}\}$. Denote the upper bound for $I_{1,m}+O_{0,m}$ obtained in this step as $F_1^* = \sum_{j=1}^{m}(p_{0,j} - p_{1,j})$.

We continue this process. Suppose until step $\overset{(i)}{\leq}$ in (19), all the previous equalities can be reached, indicating that $(\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*) = \{(\boldsymbol{A},\boldsymbol{B})|q_{0,t} = q_{1,t}, 1 \leq t \leq i-1\}$ and $F_{i-1}^* = F_1^*$. Let $U_i = I_{1,i-1} + O_{0,i-1} + \sum_{j=i}^{m}(p_{0,j} - p_{1,j})$. If there exists $(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*)$ such that $q_{0,i} = q_{1,i}$ (the method to check this existence is stated in Lemma 3), then the upper bound derived in the step (i) is $F_i^* = F_1^*$ and the feasible set is $(\mathcal{A}_i^*,\mathcal{B}_i^*) = \{(\boldsymbol{A},\boldsymbol{B})|q_{0,t} = q_{1,t}, 1 \leq t \leq i\}$. Otherwise, we have

$$\begin{aligned} F_i^* &= \max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_c,\mathcal{B}_c)} U_{i+1} \\ &\overset{(a)}{=} \sum_{t=i+1}^{i+\delta} p_{1,t} + \sum_{t=\max\{1,i-\delta+1\}}^{i} p_{0,t} + \sum_{j=t+1}^{m}(p_{0,j} - p_{1,j}). \end{aligned}$$

We denote the set that contains all the strategies satisfying (a) as $(\mathcal{A}_i^*,\mathcal{B}_i^*)$. Then, with restriction to $(\mathcal{A}_i^*,\mathcal{B}_i^*)$, continue the previous steps to check the remaining equalities in (19). When it comes to step $\overset{(m)}{\leq}$, the maximum achievable value of $I_{1,m}+O_{0,m}$, denoted as F_m^*, is found. The solutions to the above maximization problem make up the set $(\mathcal{A}_m^*,\mathcal{B}_m^*)$, which might contain multiple elements.

Lemma 3: The possibility of $q_{0,i} = q_{1,i}, 1 \leq i \leq m$, can be easily determined by the original PMF as follows

(a) $\min_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*)}(q_{1,i} - q_{0,i}) \leq 0$ is always true;

(b) If $1 \leq i \leq \delta$, $\max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*)}(q_{1,i} - q_{0,i}) \geq 0$ is true.

(c) For $i \geq \delta+1$, when $F_{i-1}^* = F_1^*$, $\max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*)}(q_{1,i} - q_{0,i}) \geq 0$ if and only if $\sum_{t=1}^{i+\delta} p_{1,t} - \sum_{t=1}^{i-\delta} p_{0,t} \geq 0$;
When $F_{i-1}^* = F_k^*, k \geq 1$, $\max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_{i-1}^*,\mathcal{B}_{i-1}^*)}(q_{1,i} - q_{0,i}) \geq 0$ if and only if $\sum_{t=k+\delta+1}^{i+\delta} p_{1,t} - \sum_{t=k-\delta+1}^{i-\delta} p_{0,t} \geq 0$.

B. Step 2: Showing the optimal attack strategy is in $(\mathcal{A}_m^*, \mathcal{B}_m^*)$

In Step 1, we have derived the maximum value F_m^* of $F_m = I_{1,m} + O_{0,m}$ under constraints (16), (17) and (18). In this step, we will show that the optimal attack strategy, when considering all constraints (15), (16), (17) and (18), is contained in $(\mathcal{A}_m^*, \mathcal{B}_m^*)$. We will also provide an efficient procedure to find such a strategy in $(\mathcal{A}_m^*, \mathcal{B}_m^*)$.

Lemma 4: There is an optimal solution to $\max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}_c,\mathcal{B}_c)} D(\boldsymbol{A},\boldsymbol{B})$ in the set $(\mathcal{A}_m^*, \mathcal{B}_m^*)$.

Before providing outline of the proof, we introduce some definitions first. Denote $\mathcal{F}_m = \{F_m | (\boldsymbol{A},\boldsymbol{B}) \text{ satisfies (16), (17) and (18) } \} = \{F_m | q_{0,i} \geq q_{1,i}, i \in R_0, (\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}_c, \mathcal{B}_c)\}$. For an arbitrary $F'_m \in \mathcal{F}_m$, define a set $(\mathcal{A}'_m, \mathcal{B}'_m) = \{(\boldsymbol{A},\boldsymbol{B}) : I_{1,m} + O_{0,m} = F'_m\}$.

Here, we describe the main ideas of our proof. It is easy to see that Lemma 4 is true if for any valid set $(\mathcal{A}'_m, \mathcal{B}'_m)$, the following inequality holds

$$\max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}'_m,\mathcal{B}'_m)} D(\boldsymbol{A},\boldsymbol{B}) \leq \max_{(\boldsymbol{A},\boldsymbol{B})\in(\mathcal{A}^*_m,\mathcal{B}^*_m)} D(\boldsymbol{A},\boldsymbol{B}). \tag{22}$$

To prove (22), for any given F'_m, we will develop series of upper bounds for $D(\boldsymbol{A},\boldsymbol{B})$ using similar steps in Section IV-A. Denote the upper bound obtained in step j as H'_j when $(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}'_m, \mathcal{B}'_m)$ and H^*_j when $(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}^*_m, \mathcal{B}^*_m)$. We will show $H^*_j \geq H'_j$, for $m+1 \leq j \leq n$. Finally, H'_n is found to be the achievable upper bound or the maximum value for $D(\boldsymbol{A},\boldsymbol{B})$ and has the property $H'_n \leq H^*_n$.

In particularly, by restricting $(\boldsymbol{A},\boldsymbol{B})$ in $(\mathcal{A}'_m, \mathcal{B}'_m)$, we have

$$\begin{aligned} F'_m &\overset{(m+1)}{\geq} I_{1,m} + \min\{q_{0,m+1}, q_{1,m+1}\} - p_{0,m+1} \\ &\quad +O_{0,m+1} \\ &\quad ... \\ &\overset{(j)}{\geq} I_{1,m} + \sum_{t=m+1}^{j} (\min\{q_{0,t}, q_{1,t}\} - p_{0,t}) \\ &\quad +O_{0,j} \\ &\quad ... \\ &\overset{(n)}{\geq} I_{1,m} + \sum_{t=m+1}^{n} (\min\{q_{0,t}, q_{1,t}\} - p_{0,t}) \\ &= 2D(\boldsymbol{A},\boldsymbol{B}), \end{aligned} \tag{23}$$

in which the equality in (j), $m+1 \leq j \leq n$, holds when one of the constraints in (15), $q_{1,j} - q_{0,j} \geq 0$, is satisfied.

In the following, we will maximize $I_{1,m} + \sum_{t=m+1}^{j} (\min\{q_{0,t}, q_{1,t}\} - p_{0,t}) + O_{0,j}$ for each j, which will leads to a series of upper bound H'_j mentioned above.

Suppose $H'_j = H'_i$, indicating that $\forall(\boldsymbol{A}', \boldsymbol{B}') \in (\mathcal{A}'_j, \mathcal{B}'_j)$, $q_{0,t} \leq q_{1,t}, i+1 \leq t \leq j$ are true. Then we have the following theorem.

Theorem 1:

$$\begin{aligned} H'_{j+1} &= \min\Big\{H'_j, \\ &\qquad \sum_{t=j-\delta+2}^{j} p_{0,t} - \sum_{t=m+1}^{j} p_{0,t} + \sum_{t=m+1}^{j+\delta+1} p_{1,t} \Big\}. \end{aligned}$$

To make it true, $q_{0,t} = q_{1,t}, i+1 \leq t \leq j$ needs to be satisfied if possible.

Since $H'_m = F'_m$ and $F'_m \leq F^*_m$, for $\forall F'_m \in \mathcal{F}_m$, $H'_{j+1} \leq H^*_{j+1}$ can be obtained recursively. Finally, we have $H'_n \leq H^*_n$ and Lemma 4 is proved.

C. Algorithm to find the optimal adversary

In practice, to find the exact maxima of $D(\boldsymbol{A},\boldsymbol{B})$ and the attack strategy, we design an efficient algorithm to generate one particular optimal strategy. We use two loops to represent the two steps in Section IV-A and IV-B. For each component i, we calculate $\max q_{1,i}, \min q_{1,i}, \max q_{1,i}, \min q_{0,i}$ and assign values to $q^*_{0,i}$ and $q^*_{1,i}$ according to the following principles:

1) If $\exists(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}^*_{i-1}, \mathcal{B}^*_{i-1})$, s.t. $q_{0,i} = q_{1,i}$, then
$$q^*_{0,i} = q^*_{1,i} = \min\{\max_{\boldsymbol{A}\in\mathcal{A}^*_{i-1}} q_{0,i}, \max_{\boldsymbol{B}\in\mathcal{B}^*_{i-1}} q_{1,i}\}.$$
2) If $\forall(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}^*_{i-1}, \mathcal{B}^*_{i-1})$, $q_{0,i} > q_{1,i}$, then
$$q^*_{1,i} = \max_{\boldsymbol{B}\in\mathcal{B}^*_{i-1}} q_{1,j} \text{ and } q^*_{0,i} = \min_{\boldsymbol{A}\in\mathcal{A}^*_{i-1}} q_{0,j}.$$
3) If $\forall(\boldsymbol{A},\boldsymbol{B}) \in (\mathcal{A}^*_{i-1}, \mathcal{B}^*_{i-1})$, $q_{0,i} < q_{1,i}$, then
$$q^*_{1,i} = \min_{\boldsymbol{B}\in\mathcal{B}^*_{i-1}} q_{1,j} \text{ and } q^*_{0,i} = \max_{\boldsymbol{A}\in\mathcal{A}^*_{i-1}} q_{0,j}.$$

After finishing the first loop in R_0, the optimal F^*_m is obtained. Then we step into the second loop corresponding to R_1 and calculate H^*_i for each i until H^*_n is generated, which leads to the maximum P_E. Therefore, by conducting this algorithm, we find one of the best PMF after attack and the maximum error probability.

V. Numerical Results

In this section, we first give an example about the optimal attack strategy found by our algorithm. We then give an example to illustrate the relationship between attack amplitude and minimax prediction error.

In the first example, we set the original PMF under $\mathcal{H}_0$ and $\mathcal{H}_1$ as $\boldsymbol{p}_0 = \frac{1}{103}[9, 10, 20, 29, 13, 7, 3, 5, 3, 4]$ and $\boldsymbol{p}_1 = \frac{1}{103}[8, 7, 3, 8, 6, 16, 20, 21, 7, 7]$ respectively. Clearly, $R_0 = \{1, 2, 3, 4, 5\}$ and $R_1 = \{6, 7, 8, 9, 10\}$.

For this setup, we first consider $\delta = 1$. By calculating $\max\limits_{\boldsymbol{B}\in\mathcal{B}_c} \sum_{j=1}^{k} q_{1,k}$ and $\min\limits_{\boldsymbol{A}\in\mathcal{A}_c} \sum_{j=1}^{k} q_{0,j}$, we find the smallest $k = 4$ that makes $\max\limits_{\boldsymbol{B}\in\mathcal{B}} \sum_{j=1}^{k} q_{1,k} < \min\limits_{\boldsymbol{A}\in\mathcal{A}} \sum_{j=1}^{k} q_{0,j}$, indicating that $q_{0,4} > q_{1,4}$ must be true if we maintain $q_{0,i} = q_{1,i}, 1 \leq i \leq 3$. Hence, for this attack amplitude, it is impossible for the attacker to make two hypotheses indistinguishable. Eventually, the mass functions after attack are $\boldsymbol{q}^*_0 = \frac{1}{103}[9, 9, 8, 13, 29, 20, 3, 5, 3, 4]$ and $\boldsymbol{q}^*_1 = \frac{1}{103}[9, 9, 8, 6, 16, 20, 3, 18, 7, 7]$ respectively. When applying Bayesian test, the prediction error is 0.2816 before attack and 0.4029 after attack. When we set $\delta = 2$, it is possible to make $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$ the same. The PMF after attack is $\boldsymbol{q}^*_0 = \boldsymbol{q}^*_1 = [9, 10, 13, 16, 20, 20, 3, 5, 3, 4]$. The minimax prediction error is 0.5.

We also find that the vulnerabilities of similar distributions to adversary could be quite different. For example, by

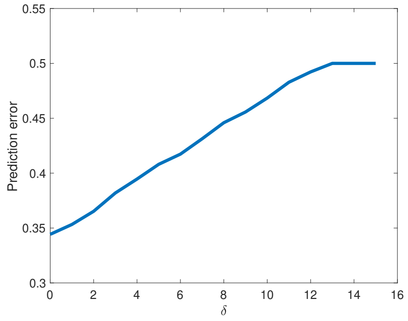

Fig. 2. Prediction error v.s. $\delta, n = 100$

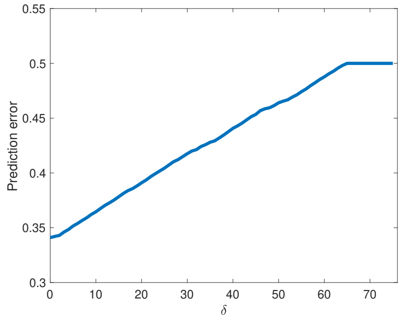

Fig. 3. Prediction error v.s. $\delta, n = 500$

switching the mass of component 1 and 4 under $\mathcal{H}_0$, i.e., $\boldsymbol{p}_0' = \frac{1}{103}[29, 10, 20, 9, 13, 7, 3, 5, 3, 4]$ and $\boldsymbol{p}_1' = \boldsymbol{p}_1$, the prediction error $P_E(\delta)$ has the value $P_E(1) = 0.3689 < 0.4029$, $P_E(2) = 0.4854 < 0.5$ and $P_E(3) = 0.5$, meaning that this small change weakens the power of attackers. In other words, $(\boldsymbol{p}_0', \boldsymbol{p}_1')$ has stronger robustness compared with $(\boldsymbol{p}_0, \boldsymbol{p}_1)$ which might be due to the boundary effect. Hence, we will allow the attacker to perform circularly and explore this boundary effect in the future.

In the second example, we explore how δ affects the prediction error for given $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$. In our experiment, for $\boldsymbol{p}_0$, we first generate a vector with n components each of which is independently generated using uniform distribution between 0 and 1, and then normalize the vector so that it is a valid PMF. For $\boldsymbol{p}_1$, we follow the same process. After that, we adjust the indexes of entries of these two vectors so that $\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ satisfy the assumption of in the paper. We then apply the proposed algorithm to find the best attack strategy and its prediction error under Bayesian test. Fig. 2 illustrate the result for the case with $n = 100, m = 50$. For this case, the result shows that the minimum δ to make $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$ the same is $\delta = 13$. When $n = 500, m = 263$, to make $P_E = 0.5$, the minimum attack amplitude is $\delta = 65$, as shown in Fig. 3. Hence, the attacker becomes more powerful as the allowed attack amplitude increases and experiments show that most PMFs can be made the same when $\delta = 0.15n$ or even smaller, indicating that quite small perturbations could cause negative effect on the prediction accuracy.

VI. Conclusion

In this paper, we have solved a minimax hypothesis testing problem corresponding to realistic classification tasks with adversarial perturbations. The optimal attack strategy has been found to maximize the error probability and the best decision maker has been shown to perform Bayesian test on the attacked PMF to minimize error probability. Numerical results are provided to support our work and the robustness of different hypotheses has been discussed.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv pp.1312.6199*, Dec. 2013.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conference on Learning Representations*, (San Diego, CA), May. 2015.

[3] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symposium on Security and Privacy*, (San Jose, CA), pp. 39–57, May. 2017.

[4] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," *arXiv pp.1710.11342*, Oct. 2017.

[5] W. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep learning models in natural language processing: A survey," *arXiv, pp.1901.06796*, Jan. 2019.

[6] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori, and A. Criminisi, "Measuring neural net robustness with constraints," in *Proc. Advances in neural information processing systems*, (Quebec, Canada), pp. 2613–2621, Dec. 2016.

[7] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symposium on Security and Privacy*, (San Jose, CA), pp. 582–597, May. 2016.

[8] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, Sep. 2018.

[9] A. Fawzi, O. Fawzi, and P. Frossard, "Fundamental limits on adversarial robustness," in *Proc. Int. Conference on Machine Learning, Workshop on Deep Learning*, (Lille, France), Jul. 2015.

[10] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," in *Proc. Advances in Neural Information Processing Systems*, (Quebec, Canada), pp. 1178–1187, Dec. 2018.

[11] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford, UK: Oxford university press, Feb. 2013.

[12] Y. Yang, "Robust estimation for dependent observations," *Manuscripta geodaetica*, vol. 19, no. 1, pp. 10–17, Oct. 1994.

[13] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, vol. 196. San Francisco, CA: John Wiley & Sons, 2011.

[14] P. J. Huber, *Robust statistics*. New York, NY: Springer, 2011.

[15] G. Gül and A. M. Zoubir, "Minimax robust hypothesis testing," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5572–5587, Apr. 2017.

[16] B. C. Levy, "Robust hypothesis testing with a relative entropy tolerance," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 413–421, Dec. 2008.

[17] L. Lai and E. Bayraktar, "On the adversarial robustness of robust estimators," *IEEE Transactions on Information Theory*, Jun. 2018. Submitted.

[18] J.-P. Aubin and I. Ekeland, *Applied nonlinear analysis*. North Chelmsford, MA: Courier Corporation, 2006.

[19] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, vol. 40. San Francisco, CA: John Wiley & Sons, 2011.