Nash Equilibria and Pitfalls of Adversarial Training in Adversarial Robustness Games

Maria-Florina Balcan
Carnegie Mellon University

Rattana Pukdee
Carnegie Mellon University

Pradeep Ravikumar
Carnegie Mellon University

Hongyang Zhang
University of Waterloo

Abstract

Adversarial training is a standard technique for training adversarially robust models. In this paper, we study adversarial training as an alternating best-response strategy in a 2-player zero-sum game. We prove that even in a simple scenario of a linear classifier and a statistical model that abstracts robust vs. non-robust features, the alternating best response strategy of such game may not converge. On the other hand, a unique pure Nash equilibrium of the game exists and is provably robust. We support our theoretical results with experiments, showing the non-convergence of adversarial training and the robustness of Nash equilibrium.

1 INTRODUCTION

Deep neural networks have been widely applied to various tasks (LeCun et al., 2015; Goodfellow et al., 2016). However, these models are vulnerable to human-imperceptible perturbations, which may lead to significant performance drop (Goodfellow et al., 2015; Szegedy et al., 2014). A large body of works has improved the robustness of neural networks against such perturbations. For example, Adversarial Training (Madry et al., 2018) (AT) is a notable technique that trains a robust model by two alternative steps: 1) finding adversarial examples of training data against the current model; 2) updating the model to correctly classify the adversarial examples and returning to step 1). This procedure has a strong connection with an alternating best-response strategy in a 2-player zero-sum game. In particular, we consider a game between an adversary (row player) and a defender (column player). At each time t, a row player outputs a perturbation function that maps each data point to a perturbation, and a column player selects a model. Given a loss function, the utility of the row player is the expected loss of the model on the perturbed data and the utility of the column player is the negative expected loss. Therefore, steps 1) and 2) correspond to both players' actions that maximize their utility against the latest action of the opponents.

In this work, we show that for the adversarial robustness game, even in a simple setting, the alternating best-response strategy may not converge. We consider a general symmetric independent distribution beyond the symmetric Gaussian distribution which was typically assumed in the prior works (Tsipras et al., 2018; Ilyas et al., 2019). We call this game the Symmetric Linear Adversarial Robustness (SLAR) game. The challenge is that SLAR is not a convex-concave game, and so those known results on convex-concave zero-sum games do not apply in our setting. One of our key contributions is to analyze the dynamics of adversarial training in the SLAR game which sheds light on the behavior of adversarial training in general. On the other hand, we prove the existence of a pure Nash equilibrium and show that any Nash equilibrium provably leads to a robust classifier, i.e., a classifier that puts zero weight on the non-robust features. The Nash equilibrium is unique where any two Nash equilibria select the same classifier. Our finding motivates us to train a model that achieves a Nash equilibrium.

For linear models, there is a closed-form solution of adversarial examples for each data point (Bose et al., 2020; Tsipras et al., 2018). Different from the alternating best-response strategy, we also study the procedure of substituting the closed-form adversarial examples into the inner maximization problem and reducing the problem to a standard minimization objective. We refer to this procedure as Optimal Adversarial Training (OAT). (Tsipras et al., 2018) has shown that OAT leads to a robust classifier under symmetric Gaussian distributions. We extend their results by showing that the same conclusion also holds for the SLAR game. We support our theoretical results with experiments, demonstrating that standard adversarial training does not converge while a Nash equilibrium is robust.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

2 RELATED WORK

2.1 Adversarial Robustness

Variants of adversarial training methods have been proposed to improve adversarial robustness of neural networks (Zhang et al., 2019; Shafahi et al., 2019; Rice et al., 2020; Wong et al., 2019; Xie et al., 2019; Qin et al., 2019). Recent works utilize extra unlabeled data (Carmon et al., 2019; Zhai et al., 2019; Deng et al., 2021; Rebuffi et al., 2021) or synthetic data from a generative model (Gowal et al., 2021; Sehwag et al., 2021) to improve the robust accuracy. Another line of works consider ensembling techniques (Tramèr et al., 2018; Sen et al., 2019; Pang et al., 2019; Zhang et al., 2022). A line of theoretical works analyzed adversarial robustness by linear models, from the trade-off between robustness and accuracy (Tsipras et al., 2018; Javanmard et al., 2020; Raghunathan et al., 2020), to the generalization property (Schmidt et al., 2018). Recent works further analyze a more complex class of models such as 2-layer neural networks (Allen-Zhu and Li, 2022; Bubeck et al., 2021; Bartlett et al., 2021; Bubeck and Sellke, 2021).

Specifically, prior works considered adversarial robustness as a 2-player zero-sum game (Pal and Vidal, 2020; Meunier et al., 2021; Bose et al., 2020; Bulò et al., 2016; Perdomo and Singer, 2019; Pinot et al., 2020). For instance, Pal and Vidal (2020) proved that randomized smoothing (Cohen et al., 2019) and FGSM attack (Goodfellow et al., 2015) form Nash equilibria. Bose et al. (2020) introduced a framework to find adversarial examples that transfer to an unseen model in the same hypothesis class. Pinot et al. (2020) shows the non-existence of a Nash equilibrium in the adversarial robustness game when the classifier and the Adversary are both deterministic. However, our settings are different from prior papers. The key assumption in previous work (Pinot et al., 2020) is that an adversary is regularized and would not attack if the adversarial example does not change the model prediction. We consider the case where the adversary attacks even though the adversarial example does not change the model prediction (as in the standard adversarial training).

While most works focused on the existence of Nash equilibrium and proposed algorithms that converge to the equilibrium, to the best of our knowledge, no prior works showed that *a Nash equilibrium is robust*.

2.2 Dynamics in Games

The dynamics of a 2-player zero-sum game has been well-studied, especially when each player takes an alternating best-response strategy in the finite action space. A classical question is whether players' actions will converge to an equilibrium as the two players alternatively play a game (Nash Jr, 1950; Nash, 1951). It is known that the alternating best-response strategy converges to a Nash equilibrium for many types of games, such as potential games (Monderer and Shapley, 1996b), weakly acyclic games (Fabrikant et al., 2010), aggregative games (Dindoš and Mezzetti, 2006), super modular games (Milgrom and Roberts, 1990), and random games (Heinrich et al., 2021; Amiet et al., 2021). However, this general phenomenon may not apply to adversarial robustness games since this natural learning algorithm may not converge even in simple games (Balcan Maria-Florina, 2012), as these results rely on specific properties of those games. In addition, there are also works on different strategies such as fictitious play (Brown, 1951; Robinson, 1951; Monderer and Shapley, 1996a; Benaım and Hirsch, 1999) and its extension to an infinite action space (Oechssler and Riedel, 2001; Perkins and Leslie, 2014) or continuous time space (Hopkins, 1999; Hofbauer and Sorin, 2006). Furthermore, there is a connection between a 2-player zero-sum game with online learning where it is possible to show that an average payoff of a player with a sub-linear regret algorithm (such as follow the regularized leader or follow the perturbed leader) converges to a Nash equilibrium (Cesa-Bianchi and Lugosi, 2006; Syrgkanis et al., 2015; Suggala and Netrapalli, 2020). However, Mertikopoulos et al. (2018) studied the dynamics of such no-regret algorithms and showed that when both players play the follow-the-regularized-leader algorithms, the actions of each player do not converge to a Nash equilibrium with a cycling behavior in the game.

3 SETUP

We consider a binary classification problem where we want to learn a linear function $f : \mathbb{R}^d \to \mathbb{R}$, such that our prediction is given by $\text{sign}(f(x))$. Let $\mathcal{D}$ be the underlying distribution of (x, y) and let $x = [x_1, \ldots, x_d]$. We assume that the distribution of each feature x_i has a symmetrical mean and is independent of the others given the label.

Assumption 1. *(Symmetrical mean) The mean of each feature x_i is symmetrical over class $y = -1, 1$. That is*

$$\mathbb{E}[x_i|y] = y\mu_i,$$

where μ_i is a constant.

Assumption 2. *(Independent features given the label) Each feature x_i is independent of each other given the label.*

We study this problem on a soft-SVM objective

$$\min_w \mathcal{L}(w), \tag{1}$$

where

$$\mathcal{L}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - yw^\top x)] + \frac{\lambda}{2}||w||_2^2,$$

and λ is the regularization parameter. Assume that at the test time, we have an adversarial perturbation function $\delta : \mathcal{D} \to \mathcal{B}(\varepsilon)$, $\mathcal{B}(\varepsilon) = \{a : ||a||_\infty \leq \varepsilon\}$, that adds a perturbation $\delta(x, y)$ to a feature of each point (x, y). Our goal is to learn a function f that makes correct predictions on the perturbed data points. We denote ε as the perturbation budget.

For a given perturbation budget ε, we divide all features x_i's into robust features and non-robust features.

Definition 1 (Non-robust feature)**.** *A feature x_i is non-robust when the perturbation budget is larger than or equal to the mean of that feature*

$$|\mu_i| \leq \varepsilon.$$

Otherwise, x_i is a robust feature.

We discuss non-robust features in more details in Appendix B.

3.1 Symmetric Linear Adversarial Robustness Game

We formulate the problem of learning a robust function f as a 2-player zero-sum game between an adversary (row player) and a defender (column player). The game is played repeatedly where at each time t, the row player outputs a perturbation function $\delta^{(t)} : \mathcal{D} \to \mathcal{B}(\varepsilon)$ that maps each data point in D to a perturbation while the column player outputs a linear function $f^{(t)} = (w^{(t)})^\top x$. The utility of the row player is given by

$$U_{\text{row}}(\delta^{(t)}, w^{(t)}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(\delta^{(t)}, w^{(t)}, x, y)] + \frac{\lambda}{2}||w^{(t)}||_2^2,$$

where

$$l(\delta, w, x, y) = \max(0, 1 - yw^\top(x + \delta(x, y))).$$

The goal of the row player is to find a perturbation function that maximizes the expected loss of the perturbed data given a model from the column player. The utility of the column player is the negative expected loss:

$$U_{\text{col}}(\delta^{(t)}, w^{(t)}) = -U_{\text{row}}(\delta^{(t)}, w^{(t)}),$$

where the column player wants to output a model that minimizes the expected loss given the perturbed data.

3.2 Adversarial Training as An Alternating Best-Response Strategy

Recall that in AT (Madry et al., 2018), we first find the "adversarial examples", i.e., the perturbed data points that maximize the loss. We then optimize our model according to the given adversarial examples. In the game-theoretic setting, AT is an alternating best-response strategy:

1. The row player submits a perturbation function that maximizes the utility from the last iteration:

$$\delta^{(t)} = \underset{\delta:\mathcal{D}\to\mathcal{B}(\varepsilon)}{\text{argmax}}\ U_{\text{row}}(\delta, w^{(t-1)}).$$

2. The column player chooses a model that maximizes the utility given the perturbation $\delta^{(t)}$:

$$w^{(t)} = \underset{w\in\mathbb{R}^d}{\text{argmax}}\, U_{\text{col}}(\delta^{(t)}, w).$$

In practice, we achieve an approximation of the $w^{(t)}$ via stochastic gradient descent and an approximation of each instance $\delta^{(t)}(x, y)$ by projected gradient descent (Madry et al., 2018).

4 NON-CONVERGENCE OF ADVERSARIAL TRAINING

In this section, we start with the dynamics of AT on a SLAR game. We then provide an example of a class of data distributions on which AT does not converge. A key property of such distributions is that it has a large fraction of non-robust features.

It is known that we have a closed form solution for the worst-case adversarial perturbation w.r.t. a linear model (Bose et al., 2020; Tsipras et al., 2018).

Lemma 1. *For a fixed w, for any $(x, y) \sim D$, the perturbation $\delta(x, y) = -y\varepsilon\,\text{sign}(w)$ maximizes the inner optimization objective*

$$\max_{\delta\in\mathcal{B}(\varepsilon)} \max(0, 1 - yw^\top(x + \delta)),$$

where

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0; \\ 0, & \text{if } x = 0; \\ -1, & \text{if } x < 0. \end{cases}$$

When x is a vector, $\text{sign}(x)$ is applied to each dimension. We denote this as the worst-case perturbation.

We note that the worst-case perturbation does not depend on the feature x, which means any point in the same class has the same perturbation. Intuitively, the worst-case perturbation shifts the distribution of each class toward the decision boundary. Since there is no other incentive for the adversary to choose another perturbation, we assume that the AT always picks the worse-case perturbation $\delta(x, y) = -y\varepsilon\,\text{sign}(w)$. This implies that at time t, the row player submits the perturbation function $\delta^{(t)}$ such that

$$\delta^{(t)}(x, y) = -y\varepsilon\,\text{sign}(w^{(t-1)}).$$

However, later in this work, we do not restrict our action space to only the worst-case perturbations when analyzing

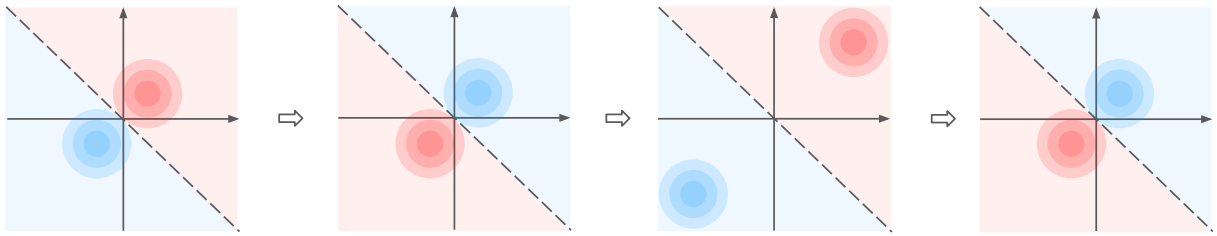

Figure 1: The space of two non-robust features, where the red and blue circles are for class $y = -1, 1$, respectively. Dashed lines represent a decision boundary of each model and background colors represent their prediction. The figure on the left is the original distribution. The adversary can always shift the mean of non-robust features across the decision boundary. A model trained with adversarial examples has to flip the decision boundary at every iteration. For example, the adversary shifts the red and blue circle across the original decision boundary leading to a decision boundary flip for a model trained on the perturbed data (second figure from the left).

a Nash equilibrium. Now, we can derive the dynamics of AT. We prove that for non-robust features x_i's, if a column player puts a positive (negative) weight of the model on x_i at time t, then the model at time $t+1$ will put a non-positive (non-negative) weight on x_i.

Theorem 1 (Dynamics of AT)**.** *Consider applying AT to learn a linear model $f(x) = w^\top x$. Let $w^{(t)} = [w_1^{(t)}, w_2^{(t)}, \ldots, w_d^{(t)}]$ be the parameter of the linear function at time t. For a non-robust feature x_i,*

1. *If $w_i^{(t)} > 0$, we have $w_i^{(t+1)} \leq 0$;*
2. *If $w_i^{(t)} < 0$, we have $w_i^{(t+1)} \geq 0$,*

for all time $t > 0$.

Proof. The key intuition is that mean of non-robust features is smaller than the perturbation budget, and the adversary can always shift the mean of these features across the decision boundary. Therefore, if we want to train a model to fit the adversarial examples, we have to flip the decision boundary at every iteration (Figure 1). Formally, consider a non-robust feature x_i with $w_i^{(t)} > 0$, the perturbation at time $t + 1$ of feature x_i is given by

$$\delta_i^{(t+1)}(x, y) = -y\varepsilon \operatorname{sign}(w_i^{(t)}) = -y\varepsilon.$$

The mean of the feature x_i of the adversarial examples at time $t + 1$ of class $y = 1$ is given by

$$\mu_i^{(t+1)} = \mathbb{E}[x_i + \delta_i^{(t+1)}(x, y) | y = 1] = \mu_i - \varepsilon \leq |\mu_i| - \varepsilon < 0.$$

The final inequality holds because x_i is a non-robust feature. We note that for a linear classifier under SVM-objective, when the mean $\mu_i^{(t+1)} < 0$ we must have $w_i^{(t+1)} \leq 0$ (see Lemma 3). □

Theorem 1 implies that the difference between the consecutive model weight is at least the magnitude of weight on non-robust features.

Corollary 1. *Consider applying AT to learn a linear model $f(x) = w^\top x$. Let $w^{(t)} = [w_1^{(t)}, w_2^{(t)}, \ldots, w_d^{(t)}]$ be the parameters of the linear function at time t. We have*

$$||w^{(t+1)} - w^{(t)}||_2^2 \geq \sum_{|\mu_i| < \varepsilon} (w_i^{(t)})^2.$$

If a large fraction of the model weight is on the non-robust features at each time t, then the model will not converge. We provide an example of data distributions where, when we train a model with AT, our model will always rely on the non-robust features at time t. We consider the following distribution:

Definition 2 (Data distribution with a large fraction of non-robust features)**.** *Let the data distribution be as follows*

1. $y \sim \text{unif}\{-1, +1\}$,
2. $x_1 = \begin{cases} +y, & \text{w.p. } p; \\ -y, & \text{w.p. } 1 - p, \end{cases}$
3. *$x_j | y$ is a distribution with mean $y\mu_j$ and variance σ_j^2, where $\varepsilon > \mu_j > 0$, for $j = 2, 3, \ldots, d + 1$.*

Given a label y, the first feature x_1 takes 2 possible values, y with probability p and $-y$ with probability $1-p$. We note that x_1 is the true label with probability p and is robust to adversarial perturbations. On the other hand, for $j \geq 2$, feature x_j is more flexible where it can follow any distribution with a condition that the feature must be weakly correlated with the true label in expectation. Each feature might be less informative compared to x_1 but combining many of them can lead to a highly accurate model. We note that x_j

is non-robust since its mean is smaller than the perturbation budget.

The data distribution is a specification of our setup in Section 3 where we have a significantly large number of non-robust features compared to the robust features. This distribution is inspired by the one studied in (Tsipras et al., 2018), where they showed that standard training on this type of distribution (when features $j = 2, \dots, d+1$ are Gaussian distributions) leads to a model that relies on non-robust features. We generalize their result to a scenario when x_j can follow any distribution (see Appendix C). We note that the lack of assumption on each distribution is a key technical challenge for our analysis.

Theorem 2 (Adversarial training uses non-robust feature (simplified version))**.** *Let the data distribution follows the distribution as in Definition 2. Consider applying AT to learn a linear model $f(x) = w^\top x$. Assume that $\varepsilon > 2\mu_j$ for $j = 2, \dots, d+1$. Let $w^{(t)} = [w_1^{(t)}, w_2^{(t)}, \dots, w_{d+1}^{(t)}]$ be the parameter of the linear function at time t. If*

$$p < 1 - \left(\frac{1}{2} \left(\frac{\sigma_{\max}}{||\mu'||_2} + \frac{\lambda}{2||\mu'||_2^2} \right) + \frac{1}{2} \sqrt{\frac{2}{\lambda}} \sigma_{\max} \right), \quad (2)$$

where

$$\sigma_i \leq \sigma_{\max}, \quad \mu' = [0, \mu_2, \dots, \mu_{d+1}],$$

then

$$\sum_{j=2}^{d+1} (w_j^{(t)})^2 \geq \frac{||w^{(t)}||_2^2 (1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1} (\mu_j + \varepsilon)^2}.$$

(For simplicity, we also assume that $\sigma_{\max} \geq 1$. For a tighter bound, see Appendix D).

Since, features $j = 2, \dots, d+1$ are not robust, Theorem 2 implies that a model at time t will put at least

$$\frac{(1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1} (\mu_j + \varepsilon)^2}$$

fraction of weight on non-robust features. We note that the term $||\mu'||_2$ at the denominator of condition (2) grows as $\mathcal{O}(d)$. Therefore, if the number of non-robust features d and the regularization parameter λ is large enough then the condition (2) holds. We discuss the full version of this theorem in Appendix D. Next, we prove that the magnitude $||w^{(t)}||_2$ is bounded below by a constant (see Lemma 6) which implies that $||w^{(t)}||_2$ does not converge to zero. Therefore, we can conclude that a model trained with AT puts a non-trivial amount of weights on non-robust features at each iteration. This implies that AT does not converge.

Theorem 3 (AT does not converge (simplified))**.** *Let the data follow the distribution as in Definition 2. Assume that the variance σ_j^2 is bounded above and $\varepsilon > 2\mu_j$ for $j = 2, \dots, d+1$. Consider applying AT to learn a linear model $f(x) = w^\top x$ on the SVM objective Let $w^{(t)}$ be the parameter of the linear function at time t. If the number of non-robust feature d and the regularization parameter λ is large enough then $w^{(t)}$ does not converge as $t \to \infty$.*

5 PROPERTIES OF THE NASH EQUILIBRIUM

Recall the definition of a Nash equilibrium:

Definition 3 (Nash Equilibrium)**.** *For a SLAR game, a pair of actions (δ^*, w^*) is called a pure strategy Nash equilibrium if the following hold*

$$\sup_{\delta} U_{\text{row}}(\delta, w^*) \leq U_{\text{row}}(\delta^*, w^*) \leq \inf_{w} U_{\text{row}}(\delta^*, w).$$

From this definition, we note that for any fixed w^*, the inequality

$$\sup_{\delta} U_{\text{row}}(\delta, w^*) \leq U_{\text{row}}(\delta^*, w^*)$$

holds if and only if δ^* is optimal. We know that the worst-case perturbation $\delta^*(x, y) = -y\varepsilon \operatorname{sign}(w^*)$ satisfies this condition. However, the optimal perturbations might not be unique because of the $\max(0, \cdot)$ operator in the hinge loss. For instance, for a point that is far away from the decision boundary such that the worst-case perturbation leads to a zero loss:

$$1 - yw^\top (x + \delta^*(x, y)) \leq 0,$$

any perturbation $\delta(x, y)$ will lead to a zero loss:

$$1 - yw^\top (x + \delta(x, y)) \leq 0.$$

Therefore, any perturbation δ leads to the same utility (Figure 2). On the other hand, if the worst-case perturbation leads to a positive loss, we can show that the optimal perturbation must be the worst-case perturbation.

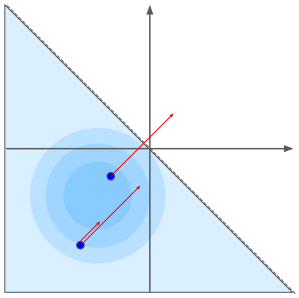

Figure 2: For points that are far enough from the decision boundary, any perturbation is optimal.

Lemma 2. *For a fixed w and any $(x, y) \sim D$, if*

$$1 - yw^\top(x - y\varepsilon\,\mathrm{sign}(w)) > 0,$$

then the optimal perturbation in is uniquely given by $\delta^*(x, y) = -y\varepsilon\,\mathrm{sign}(w)$. Otherwise, any perturbation $\mathcal{B}(\varepsilon)$ is optimal.

On the other hand, for any fixed δ^*, the inequality

$$U_{\mathrm{row}}(\delta^*, w^*) \leq \inf_w U_{\mathrm{row}}(\delta^*, w)$$

holds when w^* is an optimal solution of a standard SVM objective on the perturbed data distribution $(x + \delta^*(x, y), y) \sim D + \delta^*$. We know that for a fixed δ^*, we have a unique w^* (Lemma 7). Now, we show that a Nash equilibrium exists for the SLAR game.

Theorem 4 (Existence of Nash equilibrium)**.** *A Nash equilibrium exists for the SLAR game.*

Proof. We will prove this by construction. Without loss of generality, let feature x_i be a robust feature with $\mu_i > 0$ for $i = 1, \ldots, k$ and let feature x_j be a non-robust feature for $j = k+1, \ldots, d$. Consider a perturbation function δ^* such that

$$\delta^*(x, y) = [\underbrace{-y\varepsilon, \ldots, -y\varepsilon}_{k}, \underbrace{-y\mu_{k+1}, \ldots, -y\mu_d}_{d-k}].$$

Intuitively, δ^* shifts the distribution of non-robust features by the same distance of their mean so that non-robust features have zero mean, in the perturbed data. Let w^* be an optimal solution of a standard SVM objective on the perturbed data distribution $D + \delta^*$, which is known to be unique (Lemma 7). We will show that a pair (δ^*, w^*) is a Nash equilibrium. By the definition of w^*, it is sufficient to show that

$$\sup_\delta U_{\mathrm{row}}(\delta, w^*) \leq U_{\mathrm{row}}(\delta^*, w^*).$$

First, we will show that $w_i^* \geq 0$ for $i = 1, \ldots, k$ and $w_j^* = 0$ for $j = k+1, \ldots, d$. We consider the mean of each feature on the perturbed data,

1. For a robust feature x_i, we have

$$\mathbb{E}[x_i + \delta_i^*(x, y) | y = 1] = \mu_i - \varepsilon > 0.$$

2. For a non-robust feature x_j, we have

$$\mathbb{E}[x_j + \delta_j^*(x, y) | y = 1] = \mu_j - \mu_j = 0.$$

From Lemma 3 and Corollary 2 we can conclude that $w_i^* \geq 0$ for $i = 1, \ldots, k$ and $w_j^* = 0$ for $j = k+1, \ldots, d$. Next, we will show that δ^* is optimal. Recall that for a fixed model w^*, the worst-case perturbation is given by

$$\begin{aligned}\delta(x, y) &= -y\varepsilon\,\mathrm{sign}(w^*) \\ &= [-y\varepsilon\,\mathrm{sign}(w_1^*), \ldots, -y\varepsilon\,\mathrm{sign}(w_k^*), \underbrace{0, \ldots, 0}_{d-k}].\end{aligned}$$

Although $\delta(x, y) \neq \delta^*(x, y)$, we note that if $w_i^* = 0$, any perturbation of a feature x_i would still lead to the same loss

$$w_i^*(x_i + \delta_i(x, y)) = w_i^*(x_i + \delta_i^*(x, y)) = 0.$$

This implies that

$$U_{\mathrm{row}}(\delta^*, w^*) = U_{\mathrm{row}}(\delta, w^*) = \sup_\delta U_{\mathrm{row}}(\delta, w^*).$$

Therefore, δ^* is also an optimal perturbation function. We can conclude that (δ^*, w^*) is a Nash equilibrium. □

We can show further that for any Nash equilibrium, the weight on non-robust features must be zero.

Theorem 5 (Nash equilibrium is robust)**.** *Let (δ^*, w^*) be a Nash equilibrium of the SLAR game. For a non-robust feature x_i we must have $w_i^* = 0$.*

Proof. (Sketch) Let (δ^*, w^*) be a Nash equilibrium. Let x_i be a non-robust feature. We will show that $w_i^* = 0$ by contradiction. Without loss of generality, let $w_i^* > 0$. Let the risk term in the SVM objective when $w_i = w$, $w_j = w_j^*$ for $j \neq i$ and $\delta = \delta^*$ be

$$\mathcal{L}_i(w | w^*, \delta^*) := \mathbb{E}[l_i(x, y, w | w^*, \delta^*)].$$

when

$$\begin{aligned}l_i(x, y, w | w^*, \delta^*) = \max(0, 1 - y \sum_{j \neq i} w_j^*(x_j + \delta_j^*(x, y))) \\ - yw(x_i + \delta_i^*(x, y)).\end{aligned}$$

We will show that when we set $w_i^* = 0$, the risk term does not increase, that is,

$$\mathcal{L}_i(w_i^* | w^*, \delta^*) \geq \mathcal{L}_i(0 | w^*, \delta^*).$$

We use $l_i(x, y, w)$ to refer to $l_i(x, y, w | w^*, \delta^*)$ for the rest of this proof. Considering each point (x, y), we want to bound the difference

$$l_i(x, y, w_i^*) - l_i(x, y, 0).$$

The key idea is to utilize the optimality of $\delta^*(x, y)$. From Lemma 2, we know that when the worst-case perturbation leads to a positive loss, the perturbation $\delta^*(x, y)$ must be the worst-case perturbation. If

$$1 - y \sum_j w_j^* x_j + \varepsilon \sum_j |w_j^*| > 0,$$

we must have

$$\delta_j^*(x,y) = -y\varepsilon\,\mathrm{sign}(w_j),$$

for all j. For example, assume this is the case we have 2 sub-cases

Case 1.1:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| \geq 0.$$

In this case, we have

$$\begin{aligned}
&l_i(x,y,w_i^*) - l_i(x,y,0)\\
&= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x,y)))\\
&\quad - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x,y)))\\
&= \max(0, 1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|)\\
&\quad - \max(0, 1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|)\\
&= (1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|)\\
&\quad - (1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|)\\
&= -y w_i^* x_i + \varepsilon|w_i^*|.
\end{aligned}$$

We observe that as x_i is a non-robust feature, we have

$$\mathbb{E}[-y w_i^* x_i + \varepsilon|w_i^*|] \geq |w_i^*|(\varepsilon - \mu_i) > 0. \tag{3}$$

Case 1.2:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| < 0.$$

In this case, we have

$$\begin{aligned}
&l_i(x,y,w_i^*) - l_i(x,y,0)\\
&= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x,y)))\\
&\quad - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x,y)))\\
&= \max(0, 1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|)\\
&\quad - \max(0, 1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|)\\
&= (1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|) - 0\\
&\geq 0.
\end{aligned}$$

From Equation (3), the lower bound in this Case 1.1 is positive in expectation. However, we note that the condition depends on the rest of the feature x_j when $j \neq i$, so we can't just take the expectation. In addition, there is also a case when the optimal perturbation is not unique, that is when

$$1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*| \leq 0.$$

We handle these challenges in the full proof in Appendix F. Once we show that the risk term in the utility when $w_i^* \neq 0$ is no better than when $w_i^* = 0$, we note that the regularization term when $w_i^* = 0$ is higher,

$$\frac{\lambda}{2}\sum_j (w_j^*)^2 > \frac{\lambda}{2}\sum_{j\neq i}(w_j^*)^2.$$

Therefore, we can reduce the SVM objective by setting $w_i^* = 0$. This contradicts the optimality of w^*. By contradiction, we can conclude that if a feature i is not robust, then $w_i^* = 0$. □

Furthermore, we can show that any two Nash equilibria output the same model.

Theorem 6 (Uniqueness of Nash equilibrium)**.** *Let $(\delta_u, u), (\delta_v, v)$ be Nash equilibrium of the SLAR game then we have $u = v$.*

Proof. Let $(\delta_u, u), (\delta_v, v)$ be Nash equilibrium of the SLAR game. We know that

$$U_{\mathrm{row}}(\delta_v, u) \leq U_{\mathrm{row}}(\delta_u, u) \leq U_{\mathrm{row}}(\delta_u, v),$$

and

$$U_{\mathrm{row}}(\delta_u, v) \leq U_{\mathrm{row}}(\delta_v, v) \leq U_{\mathrm{row}}(\delta_v, u).$$

From these inequalities, we must have

$$U_{\mathrm{row}}(\delta_v, u) = U_{\mathrm{row}}(\delta_u, u) = U_{\mathrm{row}}(\delta_u, v) = U_{\mathrm{row}}(\delta_v, v).$$

From Lemma 7, we know that for a given perturbation function, we have a unique solution of the SVM objective on the perturbed data. Therefore,

$$U_{\mathrm{row}}(\delta_u, u) = U_{\mathrm{row}}(\delta_u, v)$$

implies that we must have $u = v$. □

Since we have a construction for a Nash equilibrium in Theorem 4, Theorem 6 implies that any Nash equilibrium will have the same model parameter as in the construction. This also directly implies that any Nash equilibrium is a robust classifier.

5.1 Optimal Adversarial Training

We note that in the SLAR game, we have a closed-form solution of worst-case perturbations in terms of model parameters,

$$\delta^*(x, y) = -y\varepsilon \operatorname{sign}(w).$$

We can substitute this to the minimax objective

$$\min_w \max_\delta U_{\text{row}}(\delta, w),$$

and directly solve for a Nash equilibrium. The objective is then reduced to a minimization objective

$$\min_w \mathbb{E}[\max(0, 1 - yw^\top(x - y\varepsilon \operatorname{sign}(w)))] + \frac{\lambda}{2}||w||_2^2.$$

We denote this as Optimal Adversarial Training (OAT). We note that (Tsipras et al., 2018) analyze OAT when the data distribution is Gaussian distributions and show that directly solving this objective lead to a robust model. We further show that OAT also leads to a robust model for any SLAR game.

Theorem 7 (Optimal adversarial training leads to a robust model)**.** *In the SLAR game, let $w^* = [w_1^*, \dots, w_d^*]$ be a solution of OAT then for a non-robust feature x_i, we have $w_i^* = 0$.*

We defer the proof to Appendix G.

6 EXPERIMENTS

We illustrate that the theoretical phenomenon also occurs in practice. We provide an experiment comparing the convergence and robustness of AT and OAT on a synthetic dataset and MNIST dataset.

6.1 Synthetic dataset

Though our theoretical finding works for much broader data distributions, the construction of our experimental setup is as follows

1. $y \sim \text{unif}\{-1, +1\}$,
2. $x_1 = \begin{cases} +y, & \text{w.p. } p; \\ -y, & \text{w.p. } 1-p, \end{cases}$
3. $x_j | y \sim \mathcal{N}(y\mu, \sigma^2)$.

We choose parameters $d = 2,000$, $p = 0.7$, $\mu = 0.01$, $\sigma = 0.01$ and set the perturbation budget $\varepsilon = 0.02$. This is an example of a distribution from Definition 2. The size of the training and testing data is 10,000 and 1,000, respectively. We use SGD with an Adam optimizer (Kingma and Ba, 2015) to train models. For more details, we refer to Appendix H

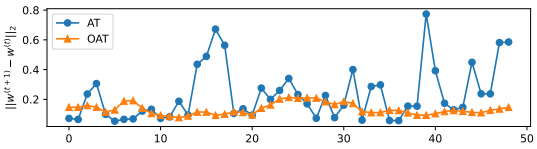

Figure 3: $||w^{(t+1)} - w^{(t)}||_2$ of a linear model trained with AT and OAT.

Non-convergence of Adversarial Training. First, we calculate the difference between weight $||w^{(t+1)} - w^{(t)}||_2$ for each timestep t. We can see that for a model trained with AT, $||w^{(t+1)} - w^{(t)}||_2$ is fluctuating while the value from a model trained with OAT is more stable (Figure 3).

Robustness. We investigate the robustness of each strategy. Since our model is linear, it is possible to calculate the distance between a point and the model's decision boundary. If the distance exceeds the perturbation budget ε then we say that the point is certifiably robust. We found that while the model trained with AT achieves a perfect standard accuracy, the model always achieves zero robust accuracy.

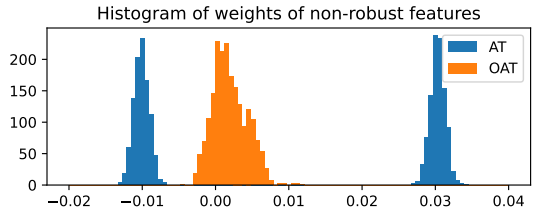

Figure 4: Weights of non-robust features at time $t = 50$.

One explanation is Theorem 9, which states that AT can lead to a model that puts non-trivial weight on non-robust features. In our dataset, feature j for $j = 2, \dots, d+1$ are non-robust but predictive so that if our model relies on these features, we can have high standard accuracy but low robust accuracy. We can see that a model trained with AT puts more weight on non-robust features (see Figure 4) and puts a higher magnitude on the positive weight which help the model to achieve 100 percent standard accuracy. On the other hand, the model trained with OAT achieves 70 percent standard accuracy and robust accuracy. The number is consistent with our construction, where we assume that the robust feature is correct with probability 0.7. In addition, we can see that OAT leads to a model that puts a much lower weight on the non-robust features (see Figure 4). This is consistent with our theoretical finding that OAT leads to a robust classifier. We note that the weights are not exactly zero because we use SGD to optimize the model.

6.2 MNIST dataset

We run experiments on a binary classification task between digits 0 and 1 on MNIST dataset (LeCun et al., 1998). The

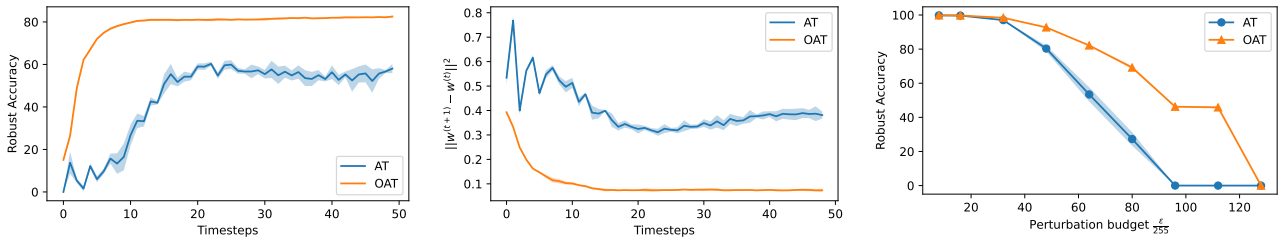

Figure 5: Robust accuracy (left) and weight difference (mid) of AT and OAT on a binary classification task between $0, 1$ on MNIST dataset when $\varepsilon = \frac{64}{255}$, and robust accuracy as we vary ε from $\frac{8}{255}$ to $\frac{128}{255}$ (right).

training and testing data have size $8,000$ and $1,500$, respectively. We train a linear classifier with AT and OAT for 50 timesteps. We use Gradient Descent with Adam optimizer and learning rate 0.01 to update our model parameter. At each timestep, for both OAT and AT, we update the model parameter with 5 gradient steps.

Non-convergence of Advesarial Training We report the difference between weight $||w^{(t+1)} - w^{(t)}||^2$ in Figure 5 (mid). The weight difference of AT fluctuates around 0.3, almost three times the weight difference of OAT.

Robustness. We report the robust accuracy at each timestep t when the perturbation budget is $\varepsilon = \frac{64}{255}$ in Figure 5 (left). We see that OAT leads to a higher robust accuracy and improved convergence than AT. For instance, at timestep 10, the robust accuracy of a model trained with OAT reaches around 80% while the value for AT is at 20%.

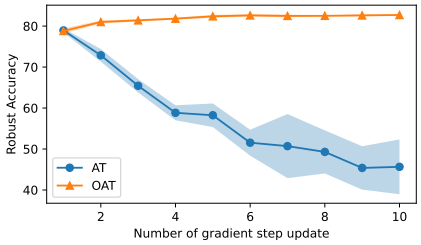

Figure 6: Robust accuracy of AT and OAT with varying number of gradient steps.

Ablations. We report robust accuracy as we vary ε from $\frac{8}{255}$ to $\frac{128}{255}$ in Figure 5 (right). We note that increasing the perturbation budget is equivalent to increasing the proportion of non-robust features. As the perturbation budget grows, robust accuracy of both model drop significantly and reach 0% when $\varepsilon = \frac{128}{255}$. We observe that OAT is more resistant to large perturbation budget than AT. For example, when $\varepsilon = \frac{96}{255}$, a model trained with OAT has a robust accuracy around 50% while the robust accuracy is 0% for AT. In addition, we report robust accuracy as we vary the number of gradient steps we take to update the model parameter at each time step (default is 5) when $\varepsilon = \frac{64}{255}$ in Figure 6. We found that as we take more gradient step, the robust accuracy of AT drop significantly while the robust accuracy of OAT increases slightly. We note that more gradient steps implies that the model parameter is closer to the optimal parameter given adversarial examples at each timestep which is the setting that we studied. Surprisingly, when taking only 1 gradient step, the robust accuracy of AT and OAT are similar but with an additional gradient step, the robust accuracy of AT drop sharply by 10% and by almost 20% with two additional gradient steps.

7 DISCUSSION

In this work, we study the dynamics of adversarial training from a SLAR game perspective. Our framework is general since the SLAR game does not make any assumption on the data distribution except that it has the symmetric means and the features given the label are independent of others. We find that iteratively training a model on adversarial examples does not suffice for a robust model, as the model manages to pick up signals from non-robust features at every epoch. One factor that leads to this phenomenon is a worst-case perturbation, which shifts the distribution across the decision boundary far enough so that the perturbed feature is predictive. On the other hand, we prove the existence, uniqueness, and robustness properties of a Nash equilibrium in the SLAR game. We note that this game has an infinite action space and is not a convex-concave game. Surprisingly, a perturbation function that leads to a Nash equilibrium is not the worst-case perturbation but is the one that perturbs the non-robust features to have zero means. Intuitively, this prevents the model from relying on non-robust features. In contrast of AT, the worst-case perturbation in OAT leads to a robust model. We remark that in our analysis, we assume that each player can find the optimal solution of their strategy at every iteration. This may not hold in practice since PGD or SGD are usually deployed to optimize for the optimal solution. However, our analysis serves as a foundation for future research on adversarial robustness game when the current assumption does not hold. That is, studying OAT in the regime when we do not have access to the closed-form adversarial examples e.g. neural networks, is an interesting future direction.

Acknowledgements

This work was supported in part by NSF grants CCF-1910321, DARPA under cooperative agreements HR00112020003, and HR00112020006, and NSERC Discovery Grant RGPIN-2022-03215, DGECR-2022-00357.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

Ben Amiet, Andrea Collevecchio, Marco Scarsini, and Ziwen Zhong. Pure Nash equilibria and best-response dynamics in random games. *Mathematics of Operations Research*, 46(4):1552–1572, 2021.

Ruta Mehta Balcan Maria-Florina, Florin Constantin. The weighted majority algorithm does not converge in nearly zero-sum games. *ICML Workshop on Markets, Mechanisms, and Multi-Agent Models.*, 2012.

Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34:9241–9252, 2021.

Michel Benaım and Morris W Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavior*, 29(1-2): 36–72, 1999.

Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and Will Hamilton. Adversarial example games. *Advances in neural information processing systems*, 33:8921–8934, 2020.

George W Brown. Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13(1):374, 1951.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pages 804–820. PMLR, 2021.

Samuel Rota Bulò, Battista Biggio, Ignazio Pillai, Marcello Pelillo, and Fabio Roli. Randomized prediction games for adversarial machine learning. *IEEE transactions on neural networks and learning systems*, 28(11): 2466–2478, 2016.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*, pages 2845–2853. PMLR, 2021.

Martin Dindoš and Claudio Mezzetti. Better-reply dynamics and global convergence to nash equilibrium in aggregative games. *Games and Economic Behavior*, 54 (2):261–292, 2006.

Alex Fabrikant, Aaron D Jaggard, and Michael Schapira. On the structure of weakly acyclic games. In *International Symposium on Algorithmic Game Theory*, pages 126–137. Springer, 2010.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Torsten Heinrich, Yoojin Jang, Luca Mungo, Marco Pangallo, Alex Scott, Bassel Tarbush, and Samuel Wiese. Best-response dynamics, playing sequences, and convergence to equilibrium in random games. Technical report, LEM Working Paper Series, 2021.

Josef Hofbauer and Sylvain Sorin. Best response dynamics for continuous zero–sum games. *Discrete & Continuous Dynamical Systems-B*, 6(1):215, 2006.

Ed Hopkins. A note on best response dynamics. *Games and Economic Behavior*, 29(1-2):138–150, 1999.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

Laurent Meunier, Meyer Scetbon, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In *International Conference on Machine Learning*, pages 7677–7687. PMLR, 2021.

Paul Milgrom and John Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, pages 1255–1277, 1990.

Dov Monderer and Lloyd S Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996a.

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1):124–143, 1996b.

John Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

Jörg Oechssler and Frank Riedel. Evolutionary dynamics on infinite strategy spaces. *Economic theory*, 17(1):141–162, 2001.

Ambar Pal and René Vidal. A game theoretic analysis of additive adversarial attacks and defenses. *Advances in Neural Information Processing Systems*, 33:1345–1355, 2020.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019.

Juan C Perdomo and Yaron Singer. Robust attacks against multiple classifiers. *arXiv preprint arXiv:1906.02816*, 2019.

Steven Perkins and David S Leslie. Stochastic fictitious play with continuous action sets. *Journal of Economic Theory*, 152:179–213, 2014.

Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. Randomization matters how to defend against strong adversarial attacks. In *International Conference on Machine Learning*, pages 7717–7727. PMLR, 2020.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32, 2019.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7909–7919, 2020.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2021.

Sanchari Sen, Balaraman Ravindran, and Anand Raghunathan. Empir: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks. In *International Conference on Learning Representations*, 2019.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 501–509, 2019.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Dinghuai Zhang, Hongyang Zhang, Aaron Courville, Yoshua Bengio, Pradeep Ravikumar, and Arun Sai Suggala. Building robust ensembles via margin boosting. *International Conference on Machine Learning*, 2022.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

A PROPERTY OF THE OPTIMAL SOLUTION OF THE SVM OBJECTIVE

We first look at the relationship between the optimal solution of this SVM objective and the underlying data distribution.

Lemma 3 (Sign of the the optimal solution)**.** *Let $w^* = [w_1^*, \dots, w_d^*]$ be an optimal solution of the SVM objective* (1)*. If each feature is independent of each other, for a feature i with*

$$\mathbb{E}[x_i|y=-1] \le 0 \le \mathbb{E}[x_i|y=1],$$

we have $w_i^* \ge 0$. Conversely, if

$$\mathbb{E}[x_i|y=1] \le 0 \le \mathbb{E}[x_i|y=-1],$$

then we have $w_i^* \le 0$.

Proof. Assume that

$$\mathbb{E}[x_i|y=-1] \le 0 \le \mathbb{E}[x_i|y=1]. \tag{4}$$

We will show that for an optimal weight w^* with $w_i^* < 0$, we can reduce the SVM objective by setting $w_i^* = 0$. This would contradict with the optimality of w^* and implies that we must have $w_i^* \ge 0$ instead. Recall that the SVM objective is given by

$$\mathbb{E}_{(x,y)\sim D}[\max(0, 1 - y\sum_{j=1}^{d} w_j x_j)] + \frac{\lambda}{2}\sum_{j=1}^{d} w_j^2.$$

We denote the first term of the objective as the risk term and the second term as the regularization term. By Jensen's inequality, we know that

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim D}[\max(0, 1 - y\sum_{j=1}^{d} w_j^* x_j)] \\
&= \mathbb{E}_{(x,y)\sim D}[\max(y\sum_{j\neq i} w_j^* x_j - 1, -y w_i^* x_i) + 1 - y\sum_{j\neq i} w_j^* x_j] \\
&\ge \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(y\sum_{j\neq i} w_j^* x_j - 1, \mathbb{E}_{x_i|y}[-y w_i^* x_i]) + 1 - y\sum_{j\neq i} w_j^* x_j].
\end{aligned}$$

We can split the expectation between x_i, x_j because each feature i are independent of each other. From (4) and $w_i^* < 0$, we have

$$\mathbb{E}_{x_i|y}[-y w_i^* x_i | y=-1] = w_i^* \mathbb{E}_{x_i}[x_i|y=-1] \ge 0,$$

and

$$\mathbb{E}_{x_i|y}[-y w_i^* x_i | y=1] = -w_i^* \mathbb{E}_{x_i}[x_i|y=1] \ge 0.$$

Therefore,

$$\mathbb{E}_{x_i|y}[-y w_i^* x_i] \ge 0.$$

This implies that,

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim D}[\max(0, 1 - y\sum_{j=1}^{d} w_j^* x_j))] \\
&\ge \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(y\sum_{j\neq i} w_j^* x_j - 1, \mathbb{E}_{x_i|y}[-y w_i^* x_i])) + 1 - y\sum_{j\neq i} w_j^* x_j] \\
&\ge \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(y\sum_{j\neq i} w_j^* x_j - 1, 0)) + 1 - y\sum_{j\neq i} w_j^* x_j] \\
&= \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(0, 1 - y\sum_{j\neq i} w_j^* x_j)].
\end{aligned}$$

The risk term in the SVM objective when $w_i^* < 0$ is no better than when $w_i^* = 0$. However, the regularization term is higher.

$$\frac{\lambda}{2}\sum_{j=1}^{d}(w_j^*)^2 > \frac{\lambda}{2}\sum_{j\neq i}(w_j^*)^2.$$

Therefore, we can reduce the SVM objective by setting $w_i^* = 0$. Therefore, $w_i < 0$ can't be the optimal weight and we must have $w_i^* \geq 0$. Similarly, we can apply the same idea to the other case. □

Corollary 2. *Let $w^* = [w_1^*, \ldots, w_d^*]$ be an optimal solution of the SVM objective* (1)*. If each feature are independent of each other, for a feature i with*

$$\mathbb{E}[x_i|y = -1] = 0 = \mathbb{E}[x_i|y = 1],$$

then we have $w_i^* = 0$.

Lemma 4 (Upper bound on the magnitude)**.** *Let $w^* = [w_1^*, \ldots, w_d^*]$ be an optimal solution of the SVM objective* (1)*. We have*

$$||w^*||_2 \leq \sqrt{\frac{2}{\lambda}}.$$

Proof. Since, w^* is an optimal solution of (1), we have

$$\mathcal{L}(w^*) \leq \mathcal{L}(0) = 1.$$

Since

$$\mathcal{L}(w^*) \geq 0 + \frac{\lambda}{2}||w^*||_2^2,$$

we have

$$\frac{\lambda}{2}||w^*||_2^2 \leq 1$$
$$||w^*||_2 \leq \sqrt{\frac{2}{\lambda}}.$$

□

Lemma 5. *Let X be a random variable with mean μ and variance σ^2 then we have*

$$\max(0, \mathbb{E}[X]) \leq \mathbb{E}[\max(0, X)] \leq \max(0, \mathbb{E}[X]) + \frac{1}{2}\sqrt{\mathrm{Var}(X)}.$$

Proof. First, we know that $\max(0, x)$ is convex and by Jensen's inequality we have

$$\mathbb{E}[\max(0, X)] \geq \max(0, \mathbb{E}[X]).$$

We also know that x^2 is convex, by Jensen's inequality, we have

$$\mathbb{E}[X^2] \geq \ \mathbb{E}[|X|]^2.$$

Next, we observe that $\max(0, x) = \frac{1}{2}(x + |x|)$,

$$\begin{aligned}
\mathbb{E}[\max(0, X)] &= \mathbb{E}[\frac{1}{2}(X + |X|)] \\
&\leq \frac{1}{2}(\mathbb{E}[(X)] + \sqrt{\mathbb{E}[X^2]}) \\
&= \frac{1}{2}(\mathbb{E}[(X)] + \sqrt{\mathrm{Var}(X) + \mathbb{E}[X]^2}) \\
&\leq \frac{1}{2}(\mathbb{E}[(X)] + \sqrt{\mathrm{Var}(X)} + |\mathbb{E}[X]| \\
&= \max(0, \mathbb{E}[X]) + \frac{1}{2}\sqrt{\mathrm{Var}(X)}.
\end{aligned}$$

□

Lemma 6 (Lower bound on the magnitude). *Let $w^* = [w_1^*, \dots, w_d^*]$ be an optimal solution of the SVM objective* (1)*. Let each feature x_i has mean and variance as follows*

$$\mathbb{E}[x_i|y] = y\mu_i, \mathrm{Var}(x_i|y) = \sigma_i^2,$$

and the feature are independent of each other then

$$||w^*||_2 \geq \frac{1}{||\mu||_2}(1 - \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2})),$$

where

$$\mu = [\mu_1, \mu_2, \dots, \mu_d],\ \bar{\sigma}_\mu^2 = \frac{\sum_{i=1}^d \mu_i^2\sigma_i^2}{\sum_{i=1}^d \mu_i^2}.$$

Proof. We will prove this by contradiction. Let w^* be an optimal solution with

$$||w^*||_2 < \frac{1}{||\mu||_2}(1 - \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2})). \tag{5}$$

Rearrange to

$$1 - ||w^*||_2||\mu||_2 > \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) > 0.$$

The SVM objective is given by

$$\begin{aligned}
\mathcal{L}(w^*) &= \mathbb{E}[\max(0, 1 - y\sum_{j=1}^d w_j^* x_j)] + \frac{\lambda}{2}||w^*||_2^2 \\
&\geq \max(0, \mathbb{E}[1 - y\sum_{j=1}^d w_j^* x_j]) \\
&= \max(0, 1 - \sum_{j=1}^d w_j^*\mu_j) \\
&\geq \max(0, 1 - ||w^*||_2||\mu||_2) \\
&= 1 - ||w^*||_2||\mu||_2 \\
&> \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}).
\end{aligned}$$

Here we apply Lemma 5, the Cauchy–Schwarz inequality and (5) respectively. On the other hand, consider $w' = \frac{\mu}{||\mu||_2^2}$. We have

$$||w'||_2 = \frac{1}{||\mu||_2} > ||w^*||_2.$$

From Lemma 5, the SVM objective satisfies

$$\begin{aligned}
\mathcal{L}(w') &= \mathbb{E}[\max(0, 1 - y\sum_{j=1}^d w_j' x_j)] + \frac{\lambda}{2}||w'||_2^2 \\
&\leq \max(0, \mathbb{E}[1 - y\sum_{j=1}^d w_j' x_j]) + \frac{1}{2}\sqrt{\mathrm{Var}(1 - y\sum_{j=1}^d w_j' x_j)} + \frac{\lambda}{2}||w'||_2^2 \\
&= \max(0, \mathbb{E}[1 - \frac{||\mu||_2}{||\mu||_2}]) + \frac{1}{2}\sqrt{\sum_{j=1}^d (w_j')^2\,\mathrm{Var}(x_j)} + \frac{\lambda}{2}||w'||_2^2 \\
&= \frac{1}{2}\sqrt{\frac{\sum_{j=1}^d \mu_j^2\sigma_j^2}{||\mu||_2^4}} + \frac{\lambda}{2}\frac{1}{||\mu||_2^2} \\
&= \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) \\
&< \mathcal{L}(w^*).
\end{aligned}$$

This contradicts with the optimality of w^*. □

Lemma 7. *Let u, v be optimal solution of the SVM objective* (1) *under a data distribution $\mathcal{D}$ then we must have $u = v$.*

Proof. We will prove this by contradiction, assume that $u \neq v$. Since both are optimal solutions, we have

$$\mathcal{L}(u) = \mathcal{L}(v)$$

when

$$\mathcal{L}(u) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - yu^\top x)] + \frac{\lambda}{2}||u||_2^2$$
$$\mathcal{L}(v) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - yv^\top x)] + \frac{\lambda}{2}||v||_2^2.$$

Consider

$$\begin{aligned}
\mathcal{L}(u) &= \frac{1}{2}(\mathcal{L}(u) + \mathcal{L}(v)) \\
&= \frac{1}{2}(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - yu^\top x)] + \frac{\lambda}{2}||u||_2^2 \\
&\quad + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - yv^\top x)] + \frac{\lambda}{2}||v||_2^2) \\
&\geq \frac{1}{2}(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 2 - y(u+v)^\top x)] + \frac{\lambda}{2}(||u||_2^2 + ||v||_2^2)) \\
&> \frac{1}{2}(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 2 - y(u+v)^\top x)] + \frac{\lambda}{2}(2||\frac{u+v}{2}||_2^2) \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max(0, 1 - y\frac{(u+v)^\top}{2}x)] + \frac{\lambda}{2}||\frac{u+v}{2}||_2^2 \\
&= \mathcal{L}(\frac{u+v}{2}).
\end{aligned}$$

We utilize an inequality

$$\max(0, a) + \max(0, b) \geq \max(0, a + b)$$

and

$$||u||_2^2 + ||v||_2^2 \geq 2||\frac{u+v}{2}||_2^2.$$

The equality of the second inequality does not hold since $u \neq v$ so we have

$$||u||_2^2 + ||v||_2^2 > 2||\frac{u+v}{2}||_2^2.$$

This contradicts with the optimality of u, v as $\frac{u+v}{2}$ leads to a lower objective. Therefore, we must have $u = v$. □

B DISCUSSION ON NON-ROBUST FEATURES

We would like to point out that there is an alternative definition of robust features.

Definition 4 (Non-robust feature (alternative))**.** *We say that feature x_i is non-robust (alternative) if there exists a perturbation $\delta_i : \mathcal{D} \to \mathcal{B}(\varepsilon)$ such that after adding the perturbation, the distribution of feature i of class $y = -1$ is close to the distribution of feature i of class $y = 1$. Formally, let $D_{i,y} + \delta$ be the distribution of the perturbed feature $x_i + \delta_i(x, y)$ for a class y. The feature x_i is non robust when*

$$\mathrm{Dist}(D_{i,-1} + \delta, D_{i,1} + \delta) \leq c,$$

where Dist *is your choice of distance metric (this can be Total variation, KL divergence or etc.) and c is a constant. Otherwise, the feature x_i is robust (alternative).*

(a) An example of a feature that is only robust to mean shift but is not robust

(b) An example of a feature that is robust but is not robust to the mean shift because two distributions have the same mean

Figure 7: Features that are robust to mean shift are not necessarily robust in the alternative definition. Red and blue represents each class $y = -1, 1$ and a line represents the mean of each class.

We note that this alternative definition of non-robust feature is different from one we defined earlier which we refer to non-robust to mean shift. Figure 7 provides examples for this where each color represent the distribution of the feature given class $y = -1, 1$.

1. We can have two distributions where the distance between the means is 2.1ε that is the feature is robust to mean shift. However when perturbed, the two distributions are almost aligned with each other, thus not robust (alternative) (Figure 7a).

2. On the other hand, we can have two distributions with the same mean such that the feature is not robust to mean shift. However, the shape of these distributions are different enough so that it is not possible to perturb them to be close to each other. Therefore, this feature is robust (alternative) (Figure 7b).

C STANDARD TRAINING RELIES ON NON-ROBUST FEATURES

Theorem 8 (Standard training uses non-robust feature)**.** *Let the data distribution follows the distribution as in Definition 2. Let $w^* = [w_1^*, w_2^*, \dots, w_{d+1}^*]$ be the optimal solution under a standard SVM objective,*

$$w^* = \underset{w}{\operatorname{argmin}} \mathbb{E}[\max(0, 1 - yw^\top x)] + \frac{\lambda}{2}||w||_2^2.$$

If

$$p < 1 - \frac{1}{2}\left(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}\right) - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_\mu, \tag{6}$$

where

$$\mu = [\mu_1, \mu_2, \dots, \mu_{d+1}],\ \bar{\sigma}_\mu = \sqrt{\frac{\sum_{j=1}^{d+1} \mu_j^2 \sigma_j^2}{||\mu||_2^2}},$$

then w^* will rely on the non-robust feature j,

$$w_1^* \leq \sum_{j \geq 2} w_j^* \mu_j.$$

This also implies that

$$\sum_{j=2}^{d+1} (w_j^*)^2 \geq \frac{||w^*||_2^2}{1 + \sum_{j=2}^{d+1} \mu_j^2}.$$

Proof. We will prove by contradiction. Assume that w^* is an optimal solution of the SVM objective and (6) holds, and

$$w_1^* > \sum_{j=2}^{d+1} w_j^* \mu_j.$$

The SVM objective is given by

$$\begin{aligned}\mathcal{L}(w^*) &= \mathbb{E}[\max(0, 1 - \sum_{i=1}^{d+1} y w_i^* x_i)] + \frac{\lambda}{2}||w^*||_2^2 \\ &= p\mathbb{E}[\max(0, 1 - w_1^* - \sum_{j=2}^{d+1} y w_j^* x_j)] + (1-p)\mathbb{E}[\max(0, 1 + w_1^* - \sum_{j=2}^{d+1} y w_j^* x_j)] + \frac{\lambda}{2}||w^*||_2^2 \\ &\geq (1-p)\mathbb{E}[\max(0, 1 + w_1^* - \sum_{j=2}^{d+1} y w_j^* x_j)] + \frac{\lambda}{2}||w^*||_2^2.\end{aligned}$$

From Lemma 5, $\mathbb{E}[\max(0, X)] \geq \max(0, \mathbb{E}[X])$. Therefore,

$$\begin{aligned}\mathcal{L}(w^*) &\geq (1-p)\max(0, \mathbb{E}[1 + w_1^* - \sum_{j=2}^{d+1} y w_j^* x_j])) + \frac{\lambda}{2}||w^*||_2^2 \\ &= (1-p)\max(0, 1 + w_1^* - \sum_{j=2}^{d+1} w_j^* \mu_j)) + \frac{\lambda}{2}||w^*||_2^2 \\ &\geq (1-p) + \frac{\lambda}{2}||w^*||_2^2.\end{aligned}$$

The last inequality holds because

$$w_1^* > \sum_{j=2}^{d+1} w_j^* \mu_j.$$

Now, consider $w' = \frac{||w^*||_2}{||\mu||_2}\mu$ when $\mu = [\mu_1, \mu_2, \ldots, \mu_{d+1}]$. We have

$$||w'||_2 = \frac{||w^*||_2}{||\mu||_2}||\mu||_2 = ||w^*||_2,$$

and from Lemma 5,

$$\mathbb{E}[\max(0, X)] \leq \max(0, \mathbb{E}[X]) + \frac{1}{2}\sqrt{\mathrm{Var}(X)}.$$

we have

$$\begin{aligned}\mathcal{L}(w') &= \mathbb{E}[\max(0, 1 - \sum_{i=1}^{d+1} y w_i' x_i)] + \frac{\lambda}{2}||w'||_2^2 \\ &\leq \max(0, \mathbb{E}[1 - y\sum_{j=1}^{d+1} w_j' x_j]) + \frac{1}{2}\sqrt{\mathrm{Var}(1 - y\sum_{j=1}^{d+1} w_j' x_j)} + \frac{\lambda}{2}||w^*||_2^2 \\ &= \max(0, 1 - ||w^*||_2||\mu||_2) + \frac{1}{2}\sqrt{\sum_{j=1}^{d+1}(w_j')^2\,\mathrm{Var}(x_j)} + \frac{\lambda}{2}||w^*||_2^2 \\ &= \max(0, 1 - ||w^*||_2||\mu||_2) + \frac{1}{2}||w^*||_2\sqrt{\frac{\sum_{j=1}^{d+1}\mu_j^2\sigma_j^2}{||\mu||_2^2}} + \frac{\lambda}{2}||w^*||_2^2 \\ &= \max(0, 1 - ||w^*||_2||\mu||_2) + \frac{1}{2}||w^*||_2\bar{\sigma_\mu} + \frac{\lambda}{2}||w^*||_2^2.\end{aligned}$$

From Lemma 6

$$||w^*||_2 \geq \frac{1}{||\mu||_2}(1 - \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2})),$$

and Lemma 4

$$||w^*||_2 \leq \sqrt{\frac{2}{\lambda}},$$

we have the upper bound of $\mathcal{L}(w')$ as follows

$$\mathcal{L}(w') \leq \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma_\mu} + \frac{\lambda}{2}||w^*||_2^2.$$

From (6) we know that

$$p < 1 - \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_\mu.$$

So

$$1 - p > \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_\mu,$$

and that

$$\mathcal{L}(w^*) < \mathcal{L}(w').$$

This contradicts with the fact that w^* is an optimal solution. Therefore, if w^* is an optimal solution and (6) holds, we must have

$$w_1^* \leq \sum_{j=2}^{d+1} w_j^* \mu_j.$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&(w_1^*)^2 \leq (\sum_{j=2}^{d+1} w_j^* \mu_j)^2 \leq (\sum_{j=2}^{d+1}(w_j^*)^2)(\sum_{j=2}^{d+1}\mu_j^2) \\
\iff &\sum_{i=1}^{d+1}(w_i^*)^2 \leq (\sum_{j=2}^{d+1}(w_j^*)^2)(1 + \sum_{j=2}^{d+1}\mu_j^2) \\
\iff &\frac{||w^*||_2^2}{1 + \sum_{j=2}^{d+1}\mu_j^2} \leq \sum_{j=2}^{d+1}(w_j^*)^2.
\end{aligned}$$

□

The condition in Theorem 8 holds when the number of non-robust features d and the regularization parameter λ are large enough,

$$p < 1 - \frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_\mu.$$

When $||\mu||_2$ is large the RHS will be larger so the condition holds for more value of p. In an extreme case when $||\mu||_2 \to \infty$, we have

$$\frac{1}{2}(\frac{\bar{\sigma}_\mu}{||\mu||_2} + \frac{\lambda}{2||\mu||_2^2}) \to 0.$$

The condition becomes

$$p < 1 - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_\mu.$$

Therefore, a necessary condition is that the regularization parameter has to be large enough

$$\lambda > \frac{\bar{\sigma}_\mu^2}{2(1-p)^2}$$

where λ scales with the weighted average of the variance $\bar{\sigma}_\mu^2$. In general, the term $||\mu||_2 = \sqrt{\sum_{j=1}^{d}\mu_j^2}$ depends on the magnitude and the number of non-robust features. If the each μ_i is large then we only need a smaller d for $||\mu||_2$ to be large.

We note that the terms μ_1, σ_1^2 are fuctions of p. However, we can bound them with constants

$$\mu_1 = \mathbb{E}[x_1 y] = 2p - 1 \leq 1$$

and

$$\sigma_1^2 = 1 - (2p-1)^2 \leq 1.$$

In addition, the contribution of μ_1, σ_1 would be in $\mathcal{O}(\frac{1}{d})$ as we have a large number of non-robust feature d.

D ADVERSARIAL TRAINING RELIES ON NON-ROBUST FEATURE

Theorem 9 (Adversarial training uses non-robust feature)**.** *Let the data distribution follows the distribution as in Definition 2. Let δ be a perturbation given by adversarial training with a perturbation budget ε. We assume that the perturbation is in the form of the worst case perturbation where*

$$\delta(x, y) \in \{-y\varepsilon, 0, y\varepsilon\}^{d+1}.$$

Let $w^* = [w_1^*, w_2^*, \ldots, w_{d+1}^*]$ be the optimal solution under a standard SVM objective on the perturbed data $x + \delta$,

$$w^* = \underset{w}{\operatorname{argmin}}\, \mathbb{E}[\max(0, 1 - yw^\top(x + \delta(x, y)))] + \frac{\lambda}{2}||w||_2^2.$$

If

$$p < 1 - \sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}), \tag{7}$$

when

$$s \in \{-1, 0, 1\}^{d+1}, \bar{\sigma}_{\mu,s} = \sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j + \varepsilon s_j)^2\sigma_j^2}{||\mu + \varepsilon s||_2^2}},$$

then w^* satisfies

$$w_1^*(1 - \varepsilon) \leq \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|.$$

This implies

$$\sum_{j=2}^{d+1}(w_j^*)^2 \geq \frac{||w^*||_2^2(1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1}(\mu_j + \varepsilon)^2}.$$

Proof. Let

$$\delta(x, y) = [\delta_1(x, y), \delta_2(x, y), \ldots, \delta_{d+1}(x, y)],$$

when

$$\delta_i(x, y) = y\varepsilon s_i,$$

where s_i can take 3 possible values

$$s_i = \begin{cases} 1; \\ 0; \\ -1. \end{cases}$$

The perturbation from adversarial training does not depends on x. We can see this as shifting the whole distribution for each feature. For the first feature

$$x_1 + \delta_1(x, y) = \begin{cases} y(1 + \varepsilon s_1), & \text{w.p. } p; \\ -y(1 - \varepsilon s_1), & \text{w.p. } 1 - p. \end{cases}$$

For each feature j for $j = 2, \ldots, d+1$, this perturbation will only change the mean of the perturbed data but will preserve the variance.

$$\mathbb{E}[x_j + \delta_j(x, y)] = y\mu_j + y\varepsilon s_j, \mathrm{Var}(x_j + \delta_j(y)) = \sigma_j^2.$$

We refer $\delta(y,s)$ to the perturbation where $\delta(x,y) = y\varepsilon s$. Denote the SVM objective on the data with perturbation δ as

$$\begin{aligned}\mathcal{L}(w,\delta) &= \mathbb{E}[\max(0, 1 - yw^\top(x + \delta(x,y)))] + \frac{\lambda}{2}||w||_2^2 \\ &= \mathbb{E}[\max(0, 1 - \sum_{i=1}^{d+1} yw_i(x_i + \delta_i(x,y)))] + \frac{\lambda}{2}||w||_2^2.\end{aligned}$$

For a fixed s, let w^* be an optimal solution of the SVM objective on the perturbed data $x + \delta(y,s)$ and assume that

$$w_1^*(1-\varepsilon) > \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|.$$

First,

$$\begin{aligned}\mathcal{L}(w^*,\delta) &\geq (1-p)\max(0, \mathbb{E}[1 + w_1(1-\varepsilon s_1) - \sum_{j=2}^{d+1} yw_j^*(x_j + y\varepsilon s_j)])) + \frac{\lambda}{2}||w^*||_2^2 \\ &\geq (1-p)\max(0, 1 + w_1(1-\varepsilon s_1) - \sum_{j=2}^{d+1} w_j^*(\mu_j + \varepsilon s_j)) + \frac{\lambda}{2}||w^*||_2^2 \\ &\geq (1-p)\max(0, 1 + w_1(1-\varepsilon) - \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|) + \frac{\lambda}{2}||w^*||_2^2 \\ &\geq 1 - p + \frac{\lambda}{2}||w^*||_2^2.\end{aligned}$$

The last inequality holds because

$$w_1^*(1-\varepsilon) > \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|.$$

On the other hand, consider

$$w' = \frac{||w^*||_2}{||\mu + \varepsilon s||_2}[\mu_1 + \varepsilon s_1, \ldots, \mu_{d+1} + \varepsilon s_{d+1}].$$

we have

$$||w'||_2 = ||w^*||_2.$$

Consider

$$\begin{aligned}\mathcal{L}(w',\delta) &= \mathbb{E}[\max(0, 1 - \sum_{j=1}^{d+1} yw_j'(x_j + y\varepsilon s_j))] + \frac{\lambda}{2}||w'||_2^2 \\ &\leq \max(0, \mathbb{E}[1 - y\sum_{j=1}^{d+1} w_j'(x_j + y\varepsilon s_j)]) + \frac{1}{2}\sqrt{\mathrm{Var}(1 - y\sum_{j=1}^{d+1} w_j'(x_j + y\varepsilon s_j))} + \frac{\lambda}{2}||w^*||_2^2 \\ &= \max(0, 1 - ||w^*||_2||\mu + \varepsilon s||_2) + \frac{1}{2}||w^*||_2\sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j + \varepsilon s_j)^2\sigma_j^2}{||\mu + \varepsilon s||_2^2}} + \frac{\lambda}{2}||w^*||_2^2 \\ &= \max(0, 1 - ||w^*||_2||\mu + \varepsilon s||_2) + \frac{1}{2}||w^*||_2\bar{\sigma}_{\mu,s} + \frac{\lambda}{2}||w^*||_2^2,\end{aligned}$$

when

$$\bar{\sigma}_{\mu,s} = \sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j + \varepsilon s_j)^2\sigma_j^2}{||\mu + \varepsilon s||_2^2}}.$$

From Lemma 6

$$||w^*||_2 \geq \frac{1}{||\mu + \varepsilon s||_2}(1 - \frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2})),$$

and Lemma 4

$$||w^*||_2 \leq \sqrt{\frac{2}{\lambda}},$$

we have the upper bound of $\mathcal{L}(w')$ as follows

$$\mathcal{L}(w', \delta) \leq \frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s} + \frac{\lambda}{2}||w^*||_2^2.$$

Recall that we have

$$\mathcal{L}(w^*, \delta) \geq 1 - p + \frac{\lambda}{2}||w^*||_2^2.$$

Therefore, if

$$p < 1 - \frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2}) - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s},$$

we would have

$$\mathcal{L}(w', \delta) < \mathcal{L}(w^*, \delta),$$

which lead to a contradiction with the fact that w^* is an optimal solution. This implies that for a fixed perturbation $\delta(s)$, if

$$p < 1 - \frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2}) - \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s},$$

then the optimal solution of the SVM objective on the perturbed data $x + \delta(s)$ satisfies

$$w_1^*(1 - \varepsilon) \leq \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|.$$

Now, if we have

$$p < 1 - \sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu + \varepsilon s||_2} + \frac{\lambda}{2||\mu + \varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}),$$

we can conclude that for any perturbation $s \in \{-1, 0, 1\}^{d+1}$, the optimal solution of the SVM objective on the perturbed data $x + \delta(s)$ satisfies

$$w_1^*(1 - \varepsilon) \leq \sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|.$$

Moreover, we can apply Cauchy-Schwarz inequality to have

$$\begin{aligned}(w_1^*)^2 &\leq \frac{(\sum_{j=2}^{d+1} w_j^*|\mu_j + \varepsilon|)^2}{(1 - \varepsilon)^2} \\ &\leq \frac{(\sum_{j=2}^{d+1}(w_j^*)^2)(\sum_{j=2}^{d+1}(\mu_j + \varepsilon)^2)}{(1 - \varepsilon)^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}||w^*||_2^2 &\leq \sum_{j=2}^{d+1}(w_j^*)^2(\frac{\sum_{j=2}^{d+1}(\mu_j + \varepsilon)^2 + (1 - \varepsilon)^2}{(1 - \varepsilon)^2}) \\ \iff \frac{||w^*||_2^2(1 - \varepsilon)^2}{\sum_{j=2}^{d+1}(\mu_j + \varepsilon)^2 + (1 - \varepsilon)^2} &\leq \sum_{j=2}^{d+1}(w_j^*)^2.\end{aligned}$$

□

The condition in Theorem 9 make sure that for any perturbation, the model would still rely on non-robust feature,

$$p < 1 - \sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu+\varepsilon s||_2} + \frac{\lambda}{2||\mu+\varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}).$$

If we assume that the variance $\sigma_{\mu,s} \approx \sigma$ is about the same for all s then the condition becomes

$$p < 1 - \sup_s((\frac{1}{2}(\frac{\sigma}{||\mu+\varepsilon s||_2} + \frac{\lambda}{2||\mu+\varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\sigma).$$

We know that s^* that achieve the supremum would also minimize $||\mu+\varepsilon s||_2$. The optimal s^* follows

1. If $2\mu_i > \varepsilon$ then $s_i^* = -1$;
2. If $\varepsilon > 2\mu_i$ then $s_i^* = 0$.

If the perturbation budget is large enough where for all i, we have $\varepsilon > 2\mu_i$ then this condition is equivalent to the condition in Theorem 8.

D.1 Simplified condition

We will reduce the condition,

$$p < 1 - \sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu+\varepsilon s||_2} + \frac{\lambda}{2||\mu+\varepsilon s||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}),$$

when

$$s \in \{-1,0,1\}^{d+1}, \bar{\sigma}_{\mu,s} = \sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j+\varepsilon s_j)^2\sigma_j^2}{||\mu+\varepsilon s||_2^2}},$$

to a condition in the simplified version of Theorem 9 in the main text,

$$p < 1 - (\frac{1}{2}(\frac{\sigma_{\max}}{||\mu'||_2} + \frac{\lambda}{2||\mu'||_2^2}) + \frac{1}{2}\sqrt{\frac{2}{\lambda}}\sigma_{\max}),$$

when

$$\sigma_i \le \sigma_{\max}, \quad \mu' = [0, \mu_2, \ldots, \mu_{d+1}].$$

We make assumptions that $\varepsilon > 2\mu_i$ for $i = 2, \ldots, d+1$ so that for any $s \in \{-1,0,1\}^{d+1}$,

$$||\mu+\varepsilon s||_2 > \sum_{j=2}^{d+1}\mu_j^2 = ||\mu'||_2^2$$

We note that the terms μ_1, σ_1^2 are fuctions of p. However, we can bound them with constants

$$\mu_1 = \mathbb{E}[x_1 y] = 2p - 1 \le 1,$$

and

$$\sigma_1^2 = 1 - (2p-1)^2 \le 1.$$

We have

$$\begin{aligned}
\bar{\sigma}_{\mu,s} &= \sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j+\varepsilon s_j)^2\sigma_j^2}{||\mu+\varepsilon s||_2^2}} \\
&= \sqrt{\frac{(2p-1+\varepsilon s_1)^2\sigma_1^2+\sum_{j=2}^{d+1}(\mu_j+\varepsilon s_j)^2\sigma_j^2}{(2p-1+\varepsilon s_1)^2+\sum_{j=2}^{d+1}(\mu_j+\varepsilon s_j)^2}} \\
&\leq \sqrt{\frac{(2p-1+\varepsilon s_1)^2+\sum_{j=2}^{d+1}(\mu_j+\varepsilon s_j)^2\sigma_{\max}^2}{(2p-1+\varepsilon s_1)^2+\sum_{j=2}^{d+1}(\mu_j+\varepsilon s_j)^2}} \\
&= \sqrt{\sigma_{\max}^2+\frac{(1-\sigma_{\max}^2)(2p-1+\varepsilon s_1)^2}{(2p-1+\varepsilon s_1)^2+\sum_{j=2}^{d+1}(\mu_j+\varepsilon s_j)^2}} \\
&\leq \sigma_{\max}.
\end{aligned}$$

In the last line, we assume that $\sigma_{\max} > 1$. However, when $\sigma_{\max} \leq 1$, we can split into 2 terms

$$\begin{aligned}
\bar{\sigma}_{\mu,s} &\leq \sqrt{\sigma_{\max}^2+\frac{(1-\sigma_{\max}^2)(1+\varepsilon)^2}{\sum_{j=2}^{d+1}\mu_j^2}} \\
&\leq \sigma_{\max}+\frac{(1+\varepsilon)\sqrt{1-\sigma_{\max}^2}}{||\mu'||_2}.
\end{aligned}$$

For simplicity, we stick with the former case, when $\sigma_{\max} > 1$. We have

$$\sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu+\varepsilon s||_2}+\frac{\lambda}{2||\mu+\varepsilon s||_2^2})+\frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}) \leq (\frac{1}{2}(\frac{\sigma_{\max}}{||\mu'||_2}+\frac{\lambda}{2||\mu'||_2^2})+\frac{1}{2}\sqrt{\frac{2}{\lambda}}\sigma_{\max}.$$

Therefore, if p satisfies

$$p < 1-(\frac{1}{2}(\frac{\sigma_{\max}}{||\mu'||_2}+\frac{\lambda}{2||\mu'||_2^2})+\frac{1}{2}\sqrt{\frac{2}{\lambda}}\sigma_{\max}),$$

we would have the condition in Theorem 9.

E ADVERSARIAL TRAINING DOES NOT CONVERGE

Theorem 10. *(AT does not converge) Consider applying AT to learn a linear model $f(x) = w^\top x$ on the SVM objective when the data follows the distribution as in Definition 2. Let $w^{(t)} = [w_1^{(t)}, w_2^{(t)}, \ldots, w_{d+1}^{(t)}]$ be the parameter of the linear function at time t. If*

$$p < 1-\sup_s((\frac{1}{2}(\frac{\bar{\sigma}_{\mu,s}}{||\mu+\varepsilon s||_2}+\frac{\lambda}{2||\mu+\varepsilon s||_2^2})+\frac{1}{2}\sqrt{\frac{2}{\lambda}}\bar{\sigma}_{\mu,s}), \tag{8}$$

when

$$s \in \{-1,0,1\}^{d+1}, \bar{\sigma}_{\mu,s} = \sqrt{\frac{\sum_{j=1}^{d+1}(\mu_j+\varepsilon s_j)^2\sigma_j^2}{||\mu+\varepsilon s||_2^2}},$$

then $w^{(t)}$ does not converge as $t \to \infty$.

Proof. The difference between w of two consecutive iterations is given by

$$||w^{(t+1)}-w^{(t)}||_2^2 = \sum_{i=1}^{d+1}(w_i^{(t+1)}-w_i^{(t)})^2.$$

From Theorem 1, for a non-robust feature $j \geq 2$, the sign of $w_j^{(t)}$, $w_j^{(t+1)}$ cannot be both positive or negative. If

$$w_j^{(t)} > 0,$$

then

$$w_j^{(t+1)} \leq 0,$$

and if

$$w_j^{(t)} < 0,$$

then

$$w_j^{(t+1)} \geq 0.$$

This implies that

$$(w_j^{(t+1)} - w_j^{(t)})^2 = (|w_j^{(t+1)}| + |w_j^{(t)}|)^2 \geq (w_j^{(t)})^2.$$

We have

$$||w^{(t+1)} - w^{(t)}||_2^2 \geq \sum_{j=2}^{d+1} (w_j^{(t)})^2.$$

Because (8) holds, from Theorem 9, we have

$$\sum_{j=2}^{d+1} (w_j^{(t)})^2 \geq \frac{||w^{(t)}||_2^2 (1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1} (\mu_j + \varepsilon)^2}.$$

Therefore,

$$||w^{(t+1)} - w^{(t)}||_2^2 \geq \frac{||w^{(t)}||_2^2 (1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1} (\mu_j + \varepsilon)^2}. \tag{9}$$

Assume that $w^{(t)}$ converge to w^* as $t \to \infty$ then we must have

$$||w^{(t+1)} - w^{(t)}||_2^2 \to 0,$$

and

$$||w^{(t)}||_2^2 \to ||w^*||_2^2.$$

From inequality (9), take $t \to \infty$, we have

$$0 \geq \frac{||w^*||_2^2 (1-\varepsilon)^2}{(1-\varepsilon)^2 + \sum_{j=2}^{d+1} (\mu_j + \varepsilon)^2}.$$

Therefore,

$$||w^*||_2 = 0.$$

If $w^{(t)}$ converge then it can only converge to 0. However, from Lemma 6

$$||w^{(t)}||_2 \geq \frac{1}{||\mu + \varepsilon s^{(t)}||_2} (1 - \frac{1}{2} (\frac{\bar{\sigma}_{\mu, s^{(t)}}}{||\mu + \varepsilon s^{(t)}||_2} + \frac{\lambda}{2||\mu + \varepsilon s^{(t)}||_2^2})),$$

when

$$\bar{\sigma}_{\mu, s^{(t)}} = \sqrt{\frac{\sum_{j=1}^{d+1} (\mu_j + \varepsilon s_j^{(t)})^2 \sigma_j^2}{||\mu + \varepsilon s^{(t)}||_2^2}},$$

and $s^{(t)} = \frac{1}{y\varepsilon} \delta^{(t)}$ is the sign of the perturbation at time t. $||w^{(t)}||_2$ is bounded below therefore it cannot converge to zero. This leads to a contradiction. We can conclude that $w^{(t)}$ does not converge as $t \to \infty$.

□

F NASH EQUILIBRIUM IS ROBUST

Proof. Let (δ^*, w^*) be a Nash equilibrium. Let x_i be a non-robust feature. We will show that $w_i^* = 0$ by contradiction. Without loss of generality, let $w_i^* > 0$. Let the risk term in the SVM objective when $w_i = w$, $w_j = w_j^*$ for $j \neq i$ and $\delta = \delta^*$ be

$$\mathcal{L}_i(w|w^*, \delta^*) := \mathbb{E}[l_i(x, y, w|w^*, \delta^*)].$$

when

$$l_i(x, y, w|w^*, \delta^*) = \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) - yw(x_i + \delta_i^*(x, y)).$$

We will show that when we set $w_i^* = 0$, the risk term does not increase, that is,

$$\mathcal{L}_i(w_i^*|w^*, \delta^*) \geq \mathcal{L}_i(0|w^*, \delta^*).$$

We use $l_i(x, y, w)$ to refer to $l_i(x, y, w|w^*, \delta^*)$ for the rest of this proof. Considering each point (x, y), we have 2 cases:

Case 1:

$$1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*| > 0.$$

From Lemma 2, we have

$$\delta_j^*(x, y) = -y\varepsilon\,\mathrm{sign}(w_j),$$

for all j with $w_j^* \neq 0$.

Case 1.1:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*| \geq 0.$$

In this case, we have

$$\begin{aligned}
l_i(x, y, w_i^*) - l_i(x, y, 0) &= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x, y))) - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&= \max(0, 1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|) - \max(0, 1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|) \\
&= (1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|) - (1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|) \\
&= -yw_i^* x_i + \varepsilon|w_i^*|.
\end{aligned}$$

Case 1.2:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*| < 0.$$

In this case, we have

$$\begin{aligned}
&l_i(x, y, w_i^*) - l_i(x, y, 0) \\
&\quad= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x, y))) - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&\quad= \max(0, 1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|) - \max(0, 1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i} |w_j^*|) \\
&\quad= (1 - y\sum_j w_j^* x_j + \varepsilon\sum_j |w_j^*|) - 0 \\
&\quad\geq 0.
\end{aligned}$$

Case 2:

$$1 - y\sum_j w_j^* x_j + \varepsilon \sum_j |w_j^*| \leq 0.$$

From Lemma 2, $\delta_j^*(x, y)$ can take any value in $[-\varepsilon, \varepsilon]$ and

$$\max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x, y))) = 0.$$

Case 2.1:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon \sum_{j\neq i} |w_j^*| \geq 0.$$

This implies that

$$-yw_i^* x_i + \varepsilon|w_i^*| \leq 0.$$

We have 2 further cases:

Case 2.1.1:

$$1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y)) \geq 0.$$

In this case, we have

$$\begin{aligned}
l_i(x, y, w_i^*) - l_i(x, y, 0) &= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x, y))) - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&= 0 - (1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&= -(1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&\geq -(1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon \sum_{j\neq i} |w_j^*|) \\
&\geq -yw_i^* x_i + \varepsilon|w_i^*|.
\end{aligned}$$

The final inequality holds since

$$1 - y\sum_j w_j^* x_j + \varepsilon \sum_j |w_j^*| \leq 0,$$

which implies

$$-yw_i^* x_i + \varepsilon|w_i^*| \leq -(1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon \sum_{j\neq i} |w_j^*|).$$

Case 2.1.2:

$$1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y)) < 0.$$

In this case, we have

$$\begin{aligned}
l_i(x, w_i^*) - l_i(x, 0) &= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x, y))) - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x, y))) \\
&= 0 - 0 \\
&\geq -yw_i^* x_i + \varepsilon|w_i^*|.
\end{aligned}$$

The last inequality holds because we know that

$$-yw_i^* x_i + \varepsilon|w_i^*| \leq 0.$$

Case 2.2:

$$1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon \sum_{j\neq i}|w_j^*| < 0.$$

In this case, we have

$$1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x,y)) < 1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| < 0,$$

and

$$\begin{aligned} l_i(x, w_i^*) - l_i(x,0) &= \max(0, 1 - y\sum_j w_j^*(x_j + \delta_j^*(x,y))) - \max(0, 1 - y\sum_{j\neq i} w_j^*(x_j + \delta_j^*(x,y))) \\ &= 0 - 0 = 0. \end{aligned}$$

From every case, we can conclude that

$$\begin{aligned} \mathcal{L}_i(w_i^*|w_*, \delta^*) - \mathcal{L}_i(0|w*, \delta^*) &:= \mathbb{E}[l_i(x,y,w_i^*) - l_i(x,y,0)] \\ &\geq \mathbb{E}[(-yw_i^* x_i + \varepsilon|w_i^*|)\mathbb{1}\{1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| \geq 0\}] \\ &= \mathbb{E}[-yw_i^* x_i + \varepsilon|w_i^*|]\mathbb{P}\left(1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| \geq 0\right) \\ &\geq |w_i^*|(\varepsilon - |\mu_i|)\mathbb{P}\left(1 - y\sum_{j\neq i} w_j^* x_j + \varepsilon\sum_{j\neq i}|w_j^*| \geq 0\right) \\ &\geq 0, \end{aligned}$$

where the third line holds since the features are independent of each other. The risk term in the utility when $w_i \neq 0$ is no better than when $w_i = 0$. However, the regularization term is higher,

$$\frac{\lambda}{2}\sum_j (w_j^*)^2 > \frac{\lambda}{2}\sum_{j\neq i}(w_j^*)^2.$$

Therefore, we can reduce the SVM objective by setting $w_i^* = 0$. This contradicts with the optimality of w^*. By contradiction, we can conclude that if a feature i is not robust, then $w_i^* = 0$. □

G OPTIMAL ADVERSARIAL TRAINING LEADS TO A ROBUST MODEL

Proof. We are learning a function $f(x) = w^\top x$ where $w = [w_1, \ldots, w_d] \in \mathbb{R}^d$. For a fixed w, we know that the perturbation that maximizes the inner loss is $\delta^*(x,y) = -y\varepsilon\,\mathrm{sign}(w)$. Substitute this in the objective, we are left to solve

$$\min_w \mathbb{E}_{(x,y)\sim D}[\max(0, 1 - yw^\top x + \varepsilon||w||_1)] + \frac{\lambda}{2}||w||_2^2. \tag{10}$$

Assume that w^* is an optimal solution of (10). For a non-robust feature x_i, we will show that $w_i^* = 0$ by contradiction. Assume that $|w_i^*| > 0$. Consider the expected contribution of w_i^* to the first term of the objective (risk) is given by

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim D}[\max(0, 1 - yw^\top x + \varepsilon||w||_1)] \\
&= \mathbb{E}_{(x,y)\sim D}[\max(0, 1 + \sum_j (\varepsilon|w_j| - yw_j x_j))] \\
&= \mathbb{E}_{(x,y)\sim D}[\max(0, 1 + \sum_j p_j)] \\
&= \mathbb{E}_{(x,y)\sim D}[\max(-1 - \sum_{j\neq i} p_j, p_i) + 1 + \sum_{j\neq i} p_j],
\end{aligned}$$

when we denote $p_j = \varepsilon|w_j| - yw_j x_j$. Since $\max(0, \cdot)$ is a convex function, by Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}_{(x,y)\sim D}[\max(-1 - \sum_{j\neq i} p_j, p_i) + 1 + \sum_{j\neq i} p_j] \\
&= \mathbb{E}_y \mathbb{E}_{x_j|y} \mathbb{E}_{x_i|y}[\max(-1 - \sum_{j\neq i} p_j, p_i) + 1 + \sum_{j\neq i} p_j] \\
&\geq \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(-1 - \sum_{j\neq i} p_j, \mathbb{E}_{x_i|y}[p_i]) + 1 + \sum_{j\neq i} p_j].
\end{aligned}$$

Because features are independent, we can split the expectation between x_i, x_j. We note that as feature x_i is non-robust, we have $|\mu_i| \leq \varepsilon$ so that

$$\begin{aligned}
\mathbb{E}_{x_i|y}[p_i] &= \mathbb{E}_{x_i|y}[\varepsilon|w_i| - yw_i x_i] \\
&= \varepsilon|w_i| - \mathbb{E}_{x_i|y}[yw_i x_i] \\
&\geq |w_i|(\varepsilon - |\mu_i|) \\
&\geq 0.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\mathbb{E}_y \mathbb{E}_{x_j|y}[\max(-1 - \sum_{j\neq i} p_j, \mathbb{E}_{x_i|y}[p_i]) + 1 + \sum_{j\neq i} p_j] \\
&\geq \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(-1 - \sum_{j\neq i} p_j, 0) + 1 + \sum_{j\neq i} p_j] \\
&= \mathbb{E}_y \mathbb{E}_{x_j|y}[\max(0, 1 + \sum_{j\neq i} p_j)].
\end{aligned}$$

The right-hand side term is just the loss term when we set $w_i = 0$. Therefore, setting $w_i = 0$ for non-robust features does not increase the loss. At the same time, setting $w_i = 0$ reduces the second term of the objective $\frac{\lambda}{2}||w||_2^2$. Thus, we can reduce the objective (10) by setting $w_i = 0$ for non-robust feature i . This contradicts the optimality of w^*. By contradiction, we have $w_i^* = 0$ for all feature x_i that is non-robust. □

H EXPERIMENT

The loss function is a hinge loss with an ℓ_2 regularization of $\lambda = 0.01$. We train a linear model with AT and OAT for 50 time steps. For OAT, we directly optimize the weight with respect to the loss

$$\mathcal{L}_{OAT}(w) \sum_{i=1}^{200} [\max(0, 1 - y_i w^\top x_i + \varepsilon||w||_1)] + \frac{\lambda}{2}||w||_2^2.$$

We use an Adam optimizer (Kingma and Ba, 2015) with a learning rate 0.01 and batch size 200.

For AT, at each time step t, we first generate the worst-case perturbations

$$\delta^{(t)}(x, y) = -y\varepsilon \operatorname{sign}(w^{(t-1)})$$

then we generate adversarial examples accordingly. Next, we update our model with an Adam optimizer with a learning rate 0.01 and a batch size of 200. The loss of each batch is given by

$$\mathcal{L}(w)_{AT} = \sum_{i=1}^{200} \max(0, 1 - y_i w^\top (x_i + \delta^{(t)}(x_i, y_i))) + \frac{\lambda}{2}||w||_2^2.$$