

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Farokhi, F

Title:
A game-theoretic approach to adversarial linear Gaussian classification

Date:
2021-09

Citation:
Farokhi, F. (2021). A game-theoretic approach to adversarial linear Gaussian classification. IFAC Journal of Systems and Control, 17, <https://doi.org/10.1016/j.ifacsc.2021.100163>.

Persistent Link:
<https://hdl.handle.net/11343/276522>

A Game-Theoretic Approach to Adversarial Linear Gaussian Classification

Farhad Farokhi

Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
E-mail: farhad.farokhi@unimelb.edu.au

Abstract

We employ a game-theoretic model to analyze the interaction between an adversary and a classifier. There are two (i.e., positive and negative) classes to which data points can belong. The adversary wants to maximize the probability of miss-detection for the positive class (i.e., false negative probability) while it does not want to significantly modify the data point so that it still maintains favourable traits of the original class. The classifier, on the other hand, wants maximize the probability of correct detection for the positive class (i.e., true positive probability) subject to a lower-bound on the probability of correct detection for the negative class (i.e., true negative probability). For conditionally Gaussian data points (conditioned on the class) and linear support vector machine classifiers, we rewrite the optimization problems of the adversary and the classifier as convex problems and use best response dynamics to learn an equilibrium of the game. This results in computing a linear support vector machine classifier that is robust against adversarial input manipulations.

Keywords: Binary classification; Linear support vector machine; Gaussian; Adversarial machine learning; Game theory.

1. Introduction

Machine learning algorithms are observed to be vulnerable to adversarial manipulations of their inputs after training and deployment, known as evasion attacks [1, 2, 3]. In fact, some machine learning models are shown to be adversely influenced by very small perturbations to the inputs [4, 2, 5]. These observations severely restrict their applications in practice.

In this paper, we propose a game-theoretic approach to model and analyze the interactions between an adversary and a decision maker (i.e., a classifier). As a starting point for research, we focus on a binary classification problem using linear support vector machines with Gaussian-distributed data in each class. This way, we can compute optimal adversarial linear support vector machine. Note that the problem of detecting and mitigating evasion attacks in support vector machines is still an ongoing debate [6, 7].

We particularly model the interaction between the adversary and the classifier using a constant-sum game. There are two classes (i.e., positive and negative classes) to which the data can belong. The adversary is interested in maximizing the probability of miss-detection for the positive class, i.e., the probability of classification of an input belonging to the negative class while it is from the positive class, also known as the false negative probability. However, the adversary does not want to significantly modify the data so that it still maintains the favourable traits of the original class. An example of such a classification problem is a simplified spam filtering in which the nature of the email determines its class (with the positive class denoting spam emails). The adversary's objective is to modify the spam emails so that they pass the spam filtering algorithm. Manipulating the email by a large amount might negate the adversarial nature of spam emails. Also, note that the adversary cannot access all the emails and thus can only manipulate the spam

emails. The classifier is interested in maximizing the probability of correct detection for the positive class, i.e., the probability of classification of an input belonging to the positive class if it is from the positive class, also known as the true positive probability. In the spam filtering example, the classifier aims to determine if an email is spam or not based on possibly modified spam email and unaltered genuine emails. Evidently, if the objective of the classifier was solely to correctly catch all data points belonging to the positive class, its optimal behaviour would have been to ignore the received data point and to mark it as belonging to the positive class. This would correctly identify all data points belonging to the positive class however it also miss-classifies all data points from the negative class. This is in fact impractical. For instance, such a policy, in the spam filtering example, would result in marking all emails as spam, which is undesirable. Therefore, the classifier enforces a lower bound on the probability of correct detection for the negative class, i.e., the probability of classification of an input belonging to the negative class if it is from the negative class, also known as the true negative probability. We rewrite the optimization problems of the adversary and the classifier as two convex optimization problems and use a best response dynamics to learn an equilibrium of the game.

Most common methods for securing machine learning algorithms against adversarial inputs are *ad hoc* in nature or based on heuristic; see, e.g., [5, 2, 8]. For instance, it has been shown that injecting adversarial examples into the training set, often referred to as adversarial training, can increase robustness to adversarial manipulations [2]. However, this approach is dependant on the method used for generating adversarial examples and the number of the required adversarial examples is often not known *a priori*. Other approach for generation of robust machine learning models is the use of regularization [9] and distributionally-robust optimization [10]. The tight rela-

relationship between distributionally-robust optimization and regularization is explored in [11]. Another method is to rely on developing computationally-friendly robustness certificates [12]. Finally, game-theoretical models have also been used in adversarial machine learning, such as adversarial support vector machines [13, 14, 15, 16, 17]. In contrast with this paper, they do not optimize the probability of true positive for the classifier and false positive for the adversary. Instead, they consider hinge losses in their modeling (motivated by the formulation of support vector machines). Furthermore, in [16, 17], gradient descent is used for characterizing the equilibrium rather than semi-definite programming as in this paper. Probabilities of true positive and false positive for adversarial detection have been considered in [18]. That paper is hence more aligned with our problem formulation. However, they only consider discrete decision spaces to get tractable results while, in this paper, we consider continuous decision spaces and use linearity for tractability.

The problem formulation of this paper is in essence close to cheap-talk and Bayesian persuasion games [19, 20, 21, 22, 23, 24, 25, 26, 27] in which a better-informed sender wants to communicate with a receiver in a strategic manner to sway its decision. However, there is a stark difference between those studies and the setup of this paper. Most importantly, in this paper, the classifier (i.e., the receiver) is restricted to follow a machine learning model (specifically, a linear support vector machine).

2. Problem Formulation

The *adversary* has access to a random variable $x \in \mathbb{R}^n$, which can belong to two classes: positive and negative. The class to which x belongs is denoted by $\theta \in \{-1, +1\}$, which is a binary random variable itself with $\mathbb{P}\{\theta = +1\} = 1 - \mathbb{P}\{\theta = -1\} = \alpha > 0$. The random variable x is assumed to be Gaussian with mean $\mu_+ \in \mathbb{R}^n$ and co-variance matrix $\Sigma_+ > 0$ if $\theta = +1$ and is assumed to be Gaussian with mean $\mu_- \in \mathbb{R}^n$ and co-variance matrix $\Sigma_- > 0$ if $\theta = -1$. The notation $A > 0$ implies that A is a symmetric positive definite matrix while $A \geq 0$ implies that A is positive semi-definite.

The adversary communicates a message $y \in \mathbb{R}^n$ to the classifier. This message may or may not be truthful. We assume that y follows

$$y = \begin{cases} Ax + w, & \theta = +1, \\ x, & \theta = -1, \end{cases} \quad (1)$$

where $A \in \mathbb{R}^{n \times n}$ is a weighting matrix and $w \in \mathbb{R}^n$ is a Gaussian random variable (i.e., additive Gaussian noise) with mean $\mu_w \in \mathbb{R}^n$ and co-variance $\Sigma_w \geq 0$. Let $\gamma := (A, \mu_w, \Sigma_w)$ be the *policy of the adversary*. The set of all policies of the adversary is denoted by Γ . Note that the adversary may not be able to access all the data points, e.g., all the emails in the spam filtering example discussed in the introduction, and thus can only manipulate the ones belonging to the positive class, e.g., spam emails.

In this paper, we restrict the adversary's policy to be linear. This is to ensure that the adversary's abilities is a match for

the classifier (as the classifier is assumed to be linear support vector machine). Furthermore, the linearity of the adversary simplifies the analysis due to the Gaussianity of the data points. We also assume that the additive noise by the adversary w is Gaussian. This can be interpreted as the adversary employing a mixed strategy that has a Gaussian distribution. Interestingly, we show that the adversary would choose not to randomize at the equilibrium. Although all these assumptions can make the problem formulation conservative, it should be noted that the analysis, at least, provides a lower bound on the influence of a more general adversary since they can find more degrees of freedom for manipulating the data points by extending their set of strategies to also cover nonlinear mappings.

The *classifier* intends to determine the class to which the random variable x actually belongs, based on the received message y . The decision of the classifier is denoted by $z \in \{-1, +1\}$ and is determined by a linear support vector machine classifier:

$$z = \text{sign}(\alpha^\top y + \beta), \quad (2)$$

where $\alpha \in \mathbb{R}^n$ is a vector of weights and $\beta \in \mathbb{R}$ is a bias. Note that scaling both α and β by a positive constant does not change the sign of $\alpha^\top y + \beta$ and thus, without loss of generality, we can assume that $\|\alpha\|_\infty \leq 1$ and $|\beta| \leq 1$. Let $\eta := (\alpha, \beta)$ be the *policy of the classifier*. The set of all policies of the classifier is denoted by Υ .

The goal of the classifier is to correctly classify as many cases belonging to the positive class as possible. Therefore, the classifier wants to maximize

$$U_c(\gamma, \eta) := \mathbb{P}\{z = +1 | \theta = +1\}. \quad (3)$$

Evidently, if the objective of the classifier was solely to maximize $\mathbb{P}\{z = +1 | \theta = +1\}$, the optimal behaviour would have been to set $\beta = 1$ and $\alpha = 0$, i.e., ignore the received messages and mark it as belonging to the positive class. This is however impractical, e.g., such a policy in the spam filtering example results in marking all emails as spam which is undesirable. Therefore, the set of the actions of the classifier is constrained by

$$C_c(\gamma, \eta) := \mathbb{P}\{z = -1 | \theta = -1\} \geq 1 - \delta, \quad (4)$$

where $\delta \in (0, 1/2)$ is a constant. By selecting δ small enough, we can ensure that data points belonging to the negative class (e.g., non-spam emails) are almost entirely correctly classified. In what follows, we use the notation

$$\Omega_c(\gamma) := \{\eta = (\alpha, \beta) \in \Upsilon : C_c(\gamma, \eta) \geq 1 - \delta, \|\alpha\|_\infty \leq 1, |\beta| < 1\}. \quad (5)$$

The goal of the adversary is to deceive the classifier into miss-classifying more data points from the positive class, e.g., accepting more spam emails. This is achieved by maximizing

$$U_a(\gamma, \eta) := \mathbb{P}\{z = -1 | \theta = +1\}. \quad (6)$$

However, the adversary does not want to make large changes to x as that might defeat its original purpose, e.g., in the spam filtering example, the email might no longer contain the desirable

traits of the spam emails like unsolicited commercial advertisements. Conceptually, this can be achieved by ensuring that the magnitude of the changes is constrained by

$$C_a(\gamma, \eta) := \mathbb{E}\{\|x - y\|_2^2 | \theta = +1\} \leq \epsilon, \quad (7)$$

where $\epsilon > 0$ is a constant. Although, in this paper, we consider a constraint on the variance of the manipulations, the analysis can be readily extended to other constraints, e.g., mean absolute deviations. In what follows, we use the notation

$$\Omega_a(\eta) := \{\gamma \in \Gamma : C_a(\gamma, \eta) \leq \epsilon\}. \quad (8)$$

Definition 1 (Adversarial Classification Game). An adversarial classification game is defined as a strategic game between two players: an adversary and a classifier. The utilities of the adversary and the classifier are $U_c(\gamma, \eta)$ and $U_a(\gamma, \eta)$, respectively. The action spaces of the adversary and the classifier are $\Omega_c(\gamma, \eta)$ and $\Omega_a(\eta)$, respectively.

In the language of [28], an adversarial classification game is in fact a competitive economy (a generalization of the Nash equilibrium) as the action spaces of the players potentially depends on the actions of the other players. In this paper, we use game instead of competitive economy in line with more recent game theory literature.

Definition 2 (Equilibrium). A pair of policies (γ^*, η^*) constitutes an equilibrium if

$$\gamma^* \in \arg \max_{\gamma \in \Omega_a(\eta^*)} U_a(\gamma, \eta^*), \quad (9a)$$

$$\eta^* \in \arg \max_{\eta \in \Omega_c(\gamma^*)} U_c(\gamma^*, \eta). \quad (9b)$$

3. Main Results

We can prove an important result regarding the adversarial classification game illustrating the direct conflict of interest between the adversary and the classifier, as expected.

Proposition 1 (Constant-Sum Game). The adversarial classification game is a constant-sum game.

Proof: Note that $U_c(\gamma, \eta) + U_a(\gamma, \eta) = \mathbb{P}\{z = +1 | \theta = +1\} + \mathbb{P}\{z = -1 | \theta = +1\} = 1$ for any γ and η . ■

In the remainder of this section, we provide a method for computing equilibria of an adversarial classification game. We first show that the best responses of the players can be computed using convex optimizations. In what follows, $\text{erf} : \mathbb{R} \rightarrow [0, 1]$ denotes the error function, defined as $\text{erf}(x) := (\int_{-x}^x \exp(-t^2) dt) / \sqrt{\pi}$.

Theorem 1. Let (γ^*, η^*) be such that (1) and (2) holds with probability one with $A = \bar{A}/t_1$, $\mu_w = \bar{\mu}_w/t_1$, $\Sigma_w = R_w R_w^\top / t_1$, $\alpha =$

$\bar{\alpha} / \max\{\|\bar{\alpha}\|_\infty, \|\bar{\beta}\|\}$, and $\beta = \bar{\beta} / \max\{\|\bar{\alpha}\|_\infty, \|\bar{\beta}\|\}$, where $(\bar{A}, \bar{\mu}_w, R_w, Z', t_1)$ is given by

$$\arg \max_{\bar{A}, \bar{\mu}_w, R_w, Z', t} -\alpha^\top (\bar{A}\mu_+ + \bar{\mu}_w) - \beta \quad (10a)$$

$$\text{s.t.} \quad \begin{bmatrix} \frac{1}{\sqrt{2}}I & 0 & R_w^\top \alpha \\ 0 & \Sigma_+^{-1} & \bar{A}^\top \alpha \\ \alpha^\top R_w & \alpha^\top \bar{A} & t \end{bmatrix} \geq 0, \quad (10b)$$

$$\text{trace}(Z') \leq t\epsilon, \quad (10c)$$

$$\begin{bmatrix} tI & 0 & 0 & R_w^\top \\ 0 & t & 0 & \bar{\mu}_w^\top \\ 0 & 0 & t(\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & tI - \bar{A}^\top \\ R_w & \bar{\mu}_w & tI - \bar{A} & Z' \end{bmatrix} \geq 0, \quad (10d)$$

$$t \geq 0. \quad (10e)$$

and $(\bar{\alpha}, \bar{\beta}, t_2)$ is given by

$$\arg \max_{\bar{\alpha}, \bar{\beta}, t} \bar{\alpha}^\top (A\mu_+ + \mu_w) + \bar{\beta}, \quad (11a)$$

$$\text{s.t.} \quad 2\bar{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \bar{\alpha} \leq 1, \quad (11b)$$

$$-\bar{\alpha}^\top \mu_- - \bar{\beta} \geq \text{erf}^{-1}(1 - 2\delta)t, \quad (11c)$$

$$\|\sqrt{2}\Sigma_-^{1/2}\bar{\alpha}\|_2 \leq t, \quad (11d)$$

$$t \geq 0. \quad (11e)$$

Then (γ^*, η^*) is an equilibrium of the adversarial classification game.

Proof: Let us consider the classifier. The utility of the classifier can be simplified as

$$\begin{aligned} U_c(\gamma, \eta) &= \mathbb{P}\{z = +1 | \theta = +1\} \\ &= \mathbb{P}\{\alpha^\top y + \beta > 0 | \theta = +1\} \\ &= \frac{1}{2} - \frac{1}{2} \text{erf}\left(\frac{-\alpha^\top (A\mu_+ + \mu_w) - \beta}{\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}}\right), \end{aligned}$$

where the last equality follows from that $\alpha^\top y + \beta$, conditioned on the observation that $\theta = +1$, is a Gaussian random variable with mean $\mathbb{E}\{\alpha^\top y + \beta | \theta = +1\} = \alpha^\top (A\mu_+ + \mu_w) + \beta$ and variance $\text{Var}(\alpha^\top y + \beta | \theta = +1) = \alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha$. Because $\text{erf}(\cdot)$ is an increasing function, maximizing $U_c(\gamma, \eta)$ is equivalent to minimizing $(-\alpha^\top (A\mu_+ + \mu_w) - \beta) / (\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha})$. The constraint of the classifier can also be rewritten as

$$\begin{aligned} C_c(\gamma, \eta) &= \mathbb{P}\{z = -1 | \theta = -1\} \\ &= \mathbb{P}\{\alpha^\top y + \beta < 0 | \theta = -1\} \\ &= \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{-\alpha^\top \mu_- - \beta}{\sqrt{2\alpha^\top \Sigma_- \alpha}}\right) \end{aligned}$$

where the last equality again follows from that $\alpha^\top y + \beta$, conditioned on the observation that $\theta = -1$, is a Gaussian random variable with mean $\mathbb{E}\{\alpha^\top y + \beta | \theta = -1\} = \alpha^\top \mu_- + \beta$ and variance $\text{Var}(\alpha^\top y + \beta | \theta = -1) = \alpha^\top \Sigma_- \alpha$. The constraint that $C_c(\gamma, \eta) \geq 1 - \delta$ is equivalent to

$$\text{erf}\left(\frac{-\alpha^\top \mu_- - \beta}{\sqrt{2\alpha^\top \Sigma_- \alpha}}\right) \geq 1 - 2\delta.$$

Again, because $\text{erf}(\cdot)$ is an increasing function, we can rewrite the classifier's constraint as

$$\frac{-\alpha^\top \mu_- - \beta}{\sqrt{2\alpha^\top \Sigma_- \alpha}} \geq \delta' := \text{erf}^{-1}(1 - 2\delta).$$

Note that $\text{erf}^{-1}(1 - 2\delta) > 0$ because $\delta \in (0, 1/2)$. These derivations allow us to transform the optimization problem in (9b) into

$$\max_{\alpha, \beta} \frac{\alpha^\top (A\mu_+ + \mu_w) + \beta}{\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}}, \quad (12a)$$

$$\text{s.t.} \quad \frac{-\alpha^\top \mu_- - \beta}{\sqrt{2\alpha^\top \Sigma_- \alpha}} \geq \delta', \quad (12b)$$

$$\|\alpha\|_\infty \leq 1, \quad (12c)$$

$$|\beta| \leq 1. \quad (12d)$$

We use the approach of [29] for the constraint to eliminate the fractional constraint. We define the change of variables $g = 1/\sqrt{2\alpha^\top \Sigma_- \alpha}$, $\tilde{\alpha} = (1/\sqrt{2\alpha^\top \Sigma_- \alpha})\alpha$, and $\tilde{\beta} = (1/\sqrt{2\alpha^\top \Sigma_- \alpha})\beta$. Hence, we can rewrite (12) as

$$\max_{\alpha, \beta} \frac{\tilde{\alpha}^\top (A\mu_+ + \mu_w) + \tilde{\beta}}{\sqrt{2\tilde{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \tilde{\alpha}}}, \quad (13a)$$

$$\text{s.t.} \quad -\tilde{\alpha}^\top \mu_- - \tilde{\beta} \geq \delta' \quad (13b)$$

$$2\tilde{\alpha}^\top \Sigma_- \tilde{\alpha} \leq 1, \quad (13c)$$

$$\|\tilde{\alpha}\|_\infty \leq g, \quad (13d)$$

$$|\tilde{\beta}| \leq g. \quad (13e)$$

We can drop the last two inequalities noting that we can set $g = \max\{\|\alpha^*\|_\infty, |\beta^*|\}$, where

$$(\alpha^*, \beta^*) \in \arg \max_{\alpha, \beta} \frac{\tilde{\alpha}^\top (A\mu_+ + \mu_w) + \tilde{\beta}}{\sqrt{2\tilde{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \tilde{\alpha}}}, \quad (14a)$$

$$\text{s.t.} \quad -\tilde{\alpha}^\top \mu_- - \tilde{\beta} \geq \delta', \quad (14b)$$

$$2\tilde{\alpha}^\top \Sigma_- \tilde{\alpha} \leq 1. \quad (14c)$$

Again, we use the approach of [29], but this time for the utility, to rewrite this fractional optimization problem. To do so, define $t = 1/\sqrt{2\tilde{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \tilde{\alpha}}$, $\bar{\alpha} = (1/\sqrt{2\tilde{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \tilde{\alpha}})\tilde{\alpha}$, $\bar{\beta} = (1/\sqrt{2\tilde{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \tilde{\alpha}})\tilde{\beta}$. Following [29], the optimization problem in (14) is equivalent with

$$\max_{\bar{\alpha}, \bar{\beta}, t} \bar{\alpha}^\top (A\mu_+ + \mu_w) + \bar{\beta}, \quad (15a)$$

$$\text{s.t.} \quad 2\bar{\alpha}^\top (A\Sigma_+ A^\top + \Sigma_w) \bar{\alpha} \leq 1, \quad (15b)$$

$$-\bar{\alpha}^\top \mu_- - \bar{\beta} \geq \delta' t, \quad (15c)$$

$$\left\| \sqrt{2} \Sigma_-^{1/2} \bar{\alpha} \right\|_2 \leq t, \quad (15d)$$

$$t \geq 0. \quad (15e)$$

Now, we consider the adversary. The utility of the adversary is given by

$$\begin{aligned} U_a(\gamma, \eta) &= \mathbb{P}\{z = -1 | \theta = +1\} \\ &= 1 - \mathbb{P}\{z = +1 | \theta = +1\} \\ &= 1 - U_c(\gamma, \eta) \\ &= \frac{1}{2} + \frac{1}{2} \text{erf} \left(\frac{-\alpha^\top (A\mu_+ + \mu_w) - \beta}{\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}} \right) \end{aligned}$$

Because $\text{erf}(\cdot)$ is an increasing function, maximizing $U_a(\gamma, \eta)$ is the same as maximizing $(-\alpha^\top (A\mu_+ + \mu_w) - \beta)/\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}$. Furthermore, the constraint of the adversary can be rewritten as

$$C_a(\gamma, \eta) = \mathbb{E}\{\|x - y\|_2^2 | \theta = +1\}$$

$$= \mathbb{E}\{\|(I - A)x - w\|_2^2 | \theta = +1\}$$

$$= \text{trace}((I - A)(\Sigma_+ + \mu_+ \mu_+^\top)(I - A)^\top) + \text{trace}(\mu_w \mu_w^\top + \Sigma_w)$$

Following these derivations, we can transform the optimization problem in (9a) to

$$\max_{A, \Sigma_w, \mu_w} \frac{-\alpha^\top (A\mu_+ + \mu_w) - \beta}{\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}} \quad (16a)$$

$$\text{s.t.} \quad \text{trace}((I - A)(\Sigma_+ + \mu_+ \mu_+^\top)(I - A)^\top) + \text{trace}(\mu_w \mu_w^\top + \Sigma_w) \leq \epsilon, \quad (16b)$$

$$\Sigma_w \geq 0. \quad (16c)$$

Note that the inequality constraint (16b) can be replaced with the following three constraints:

$$(I - A)(\Sigma_+ + \mu_+ \mu_+^\top)(I - A)^\top + \mu_w \mu_w^\top + \Sigma_w \leq Z, \quad (17a)$$

$$Z \geq 0, \quad (17b)$$

$$\text{trace}(Z) \leq \epsilon. \quad (17c)$$

Further, the constraint (17a) can be linearized using its Schur complement because $\Sigma_+ > 0$ to get

$$\begin{bmatrix} (\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & I - A^\top \\ I - A & Z - \Sigma_w - \mu_w \mu_w^\top \end{bmatrix} \geq 0. \quad (18)$$

Again, we can use the Schur complement, to transform the constraint (18) into

$$\begin{bmatrix} 1 & 0 & \mu_w^\top \\ 0 & (\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & I - A^\top \\ \mu_w & I - A & Z - \Sigma_w \end{bmatrix} \geq 0. \quad (19)$$

Therefore, the optimization problem in (16) can be transformed into

$$\max_{A, \Sigma_w, \mu_w, Z} \frac{-\alpha^\top (A\mu_+ + \mu_w) - \beta}{\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}} \quad (20a)$$

$$\text{s.t.} \quad \begin{bmatrix} 1 & 0 & \mu_w^\top \\ 0 & (\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & I - A^\top \\ \mu_w & I - A & Z - \Sigma_w \end{bmatrix} \geq 0, \quad (20b)$$

$$Z \geq 0, \text{trace}(Z) \leq \epsilon, \Sigma_w \geq 0. \quad (20c)$$

Define change of variables $t = 1/\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha}$, $\bar{A} = (1/\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha})A$, $\bar{\mu}_w = (1/\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha})\mu_w$, $\bar{\Sigma}_w = (1/\sqrt{2\alpha^\top (A\Sigma_+ A^\top + \Sigma_w) \alpha})\Sigma_w$. Following [29], the optimization problem in (20) is equivalent with

$$\max_{\bar{A}, \bar{\Sigma}_w, \bar{\mu}_w, Z, t} -\alpha^\top (\bar{A}\mu_+ + \bar{\mu}_w) - \beta \quad (21a)$$

$$\text{s.t.} \quad \alpha^\top (\bar{A}\Sigma_+ \bar{A}^\top + t\bar{\Sigma}_w) \alpha \sqrt{2} \leq t, \quad (21b)$$

$$\begin{bmatrix} t & 0 & \bar{\mu}_w^\top \\ 0 & t(\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & tI - \bar{A}^\top \\ \bar{\mu}_w & tI - \bar{A} & tZ - \bar{\Sigma}_w \end{bmatrix} \geq 0, \quad (21c)$$

$$\text{trace}(Z) \leq \epsilon, \quad (21d)$$

$$\bar{\Sigma}_w \geq 0, Z \geq 0, t \geq 0. \quad (21e)$$

Using the Schur complement of (21b) and defining $Z' = tZ$, we can transform (21) into

$$\max_{\bar{A}, \bar{\Sigma}_w, \bar{\mu}_w, Z', t} -\alpha^\top (\bar{A}\mu_+ + \bar{\mu}_w) - \beta \quad (22a)$$

$$\text{s.t.} \quad \begin{bmatrix} \Sigma_+^{-1} & \bar{A}^\top \alpha \\ \alpha^\top \bar{A} & t - t\alpha^\top \bar{\Sigma}_w \alpha \sqrt{2} \end{bmatrix} \geq 0, \quad (22b)$$

$$\text{trace}(Z') \leq t\epsilon, \quad (22c)$$

$$\begin{bmatrix} t & 0 & \bar{\mu}_w^\top \\ 0 & t(\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & tI - \bar{A}^\top \\ \bar{\mu}_w & tI - \bar{A} & Z' - \bar{\Sigma}_w \end{bmatrix} \geq 0, \quad (22d)$$

$$\bar{\Sigma}_w \geq 0, \quad (22e)$$

$$t \geq 0. \quad (22f)$$

By defining R_w such that $R_w R_w^\top = t\bar{\Sigma}_w$ and using the Schur complements of (22b) and (22d), we can transform (22) into

$$\begin{aligned} & \max_{\bar{A}, R_w, \bar{\mu}_w, Z', t} -\alpha^\top (\bar{A}\mu_+ + \bar{\mu}_w) - \beta \\ & \text{s.t.} \quad \begin{bmatrix} \frac{1}{\sqrt{2}}I & 0 & R_w^\top \alpha \\ 0 & \Sigma_+^{-1} & \bar{A}^\top \alpha \\ \alpha^\top R_w & \alpha^\top \bar{A} & t \end{bmatrix} \geq 0, \\ & \quad \text{trace}(Z') \leq t\epsilon, \\ & \quad \begin{bmatrix} tI & 0 & 0 & R_w^\top \\ 0 & t & 0 & \bar{\mu}_w^\top \\ 0 & 0 & t(\Sigma_+ + \mu_+ \mu_+^\top)^{-1} & tI - \bar{A}^\top \\ R_w & \bar{\mu}_w & tI - \bar{A} & Z' \end{bmatrix} \geq 0, \\ & \quad t \geq 0. \end{aligned}$$

This concludes the proof. $\blacksquare$

Remark 1 (Computing the Equilibrium). We can use the best response dynamics (i.e., each player in the game updates its strategy by computing its optimal response based on the strategy of the other player) to extract an equilibrium of the game. The convergence of the best response dynamics follows from [30] because of the constant-sum nature of the game; see Proposition 1.

Corollary 1. There exists a best response policy for the adversary with

- $A \neq 0$ if $\alpha \neq 0$, $\mu_+ \neq 0$, and $\lambda_{\max}(\Sigma_+ + \mu_+ \mu_+^\top) \leq \epsilon/n$;
- $\Sigma_w = 0$.

Proof: We use $(\bar{A}, \bar{\mu}_w, R_w, Z', t)$ to denote the optimal solution of (10). We can prove $A \neq 0$ by *reductio ad absurdum*. Let us assume that this not true. Therefore, the optimal solution to (10) is such that $\bar{A} = 0$. For $\epsilon \ll 1$, let us perturb $t \rightarrow t + \epsilon$ and $Z' \rightarrow Z' + \rho I$ for $\lambda_{\max}(\Sigma_+ + \mu_+ \mu_+^\top)\epsilon \leq \rho \leq \epsilon\epsilon/n$. With this perturbation all the constraints in (10b)–(10e) are satisfied as strict inequality while the cost function remains unchanged (so we still have an optimal solution). Now, let us select $\epsilon' \ll \epsilon \ll 1$ and perturb $\bar{A} \rightarrow -\epsilon'(\alpha^\top \mu)I$. By continuity of the eigenvalues

of matrix and that (10b)–(10e) are attained as strict inequality, the resulting decision still remains feasible while the cost function has strictly increased. This results in a contradiction.

Now, we prove that $R_w = 0$ so it must be that Σ_w is also zero. We use $(\bar{A}, \bar{\mu}_w, R_w, Z', t)$ to denote the optimal solution of (10). If $R_w \neq 0$, we can set it equal to zero. Doing so the cost function would not change as it is not a function of R_w . Also, setting $R_w = 0$ keeps the constraints still satisfied. Therefore, $(\bar{A}, \bar{\mu}_w, 0, Z', t)$ must also be a solution of (10). $\blacksquare$

Corollary 1 provides structural insights into the best response policy. Particularly, we have $\Sigma_w = 0$, and therefore, the adversary does not need to randomize its decisions. This lack of randomization has been interestingly observed in other signaling games as well [20, 31]. Also, if $\alpha \neq 0$, $\mu_+ \neq 0$, and $\lambda_{\max}(\Sigma_+ + \mu_+ \mu_+^\top)n \leq \epsilon$, it gives that $\bar{A} \neq 0$, which implies that the adversary would not cancel x entirely even if ϵ is large enough.

4. Numerical Example

In this section, we illustrate the applicability of the developed game-theoretic framework on two numerical problems: an illustrative example using Gaussian data and a practical example using real data on heart disease classification.

4.1. Illustrative Example

Consider an example in which the data for the positive and negative classes are Gaussian random variables with the following means and co-variance matrices:

$$\mu_+ = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \mu_- = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_+ = \begin{bmatrix} 1 & 0 \\ 0 & 1/5 \end{bmatrix}, \Sigma_- = \begin{bmatrix} 1/5 & 0 \\ 0 & 1 \end{bmatrix}.$$

For the following experiments 500 data points from each class are randomly generated to illustrate the effects of the adversary and the classifier.

We first consider there is no adversary. In this case, the classifier is merely interested in identifying the optimal linear support vector machine by maximizing the true positive probability subject to a constraint that the true negative probability is greater than or equal to $1 - \delta = 0.99$. Figure 1 (a) shows the optimal non-adversarial support vector machine with the dashed line. The data points from the positive class are shown by the red pluses and the data points from the negative class are shown by the blue squares. In this case, the true negative probability is 0.99 and the true positive probability (which is equal to one minus the false negative probability) is 0.9993. This high performance is of course an artifact of the setup of the example in which the data points for both classes are linearly separable with high probability and mixing is only due to outliers.

Now, let the classifier select this optimal non-adversarial support vector machine as its policy. Assume that the adversary can now manipulate the data in the positive class but the classifier is not prepared to respond to these manipulations. The adversary is interested in manipulating the data from the positive class in order to maximize the false negative probability subject to a bound on the expected variance of the changes with

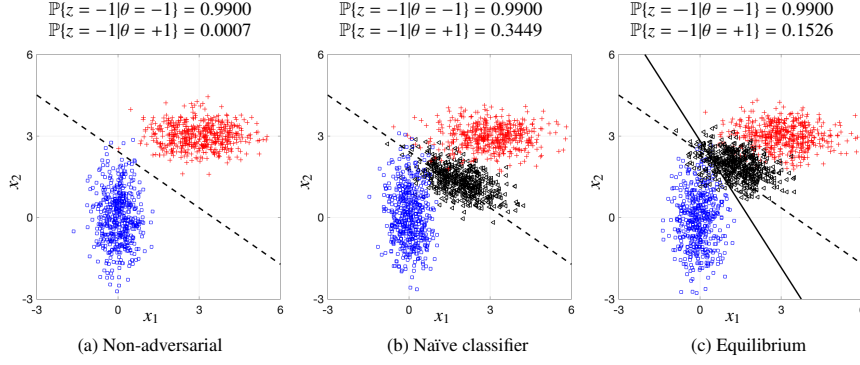


Figure 1: Illustration of the effect of the adversary on linear support vector machine classifiers using synthetic Gaussian data. There is no adversary in the non-adversarial case. The naïve case refers to when the classifier is not prepared for the adversary. The optimal non-adversarial and adversarial support vector machines are illustrated with dashed and solid lines, respectively. The original data points from the positive class are shown by the red pluses and the data points from the negative class are shown by the blue squares. The manipulated data points from the positive class are shown by the black triangles.

Table 1: The effect of adversary on linear support vector machine classifiers using the Cardiovascular Disease dataset.

	True negative probability $\mathbb{P}\{z = -1 \theta = -1\}$	False negative probability $\mathbb{P}\{z = -1 \theta = +1\}$
Non-adversarial	0.8986	0.2000
Naïve classifier	0.8986	0.9758
Equilibrium	0.9130	0.6303

$\epsilon = 2$. Figure 1 (b) the effect of the adversary on the performance of this naïve classifier. The naïveté refers to that the classifier is not prepared to respond to the adversary. The manipulated data points from the positive class are depicted by the black triangles. Similarly, the original data points from the positive class are shown by red pluses and the data points from the negative class are shown by the blue squares. In this case, the false negative probability increases 493% in comparison to the non-adversarial case. This is of course due to unprepared, naïve nature of the classifier.

Now, consider the case where the classifier uses the optimal linear support vector machine extracted from the equilibrium of the game extracted by the best response dynamics. At the equilibrium, the optimal strategy for the adversary is

$$A = \begin{bmatrix} 0.77 & -0.05 \\ -0.15 & 0.97 \end{bmatrix}, \quad \mu_w = \begin{bmatrix} -0.85 \\ -0.54 \end{bmatrix}, \quad \Sigma_w = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

This is in line with Corollary 1. Figure 1 (c) shows the optimal adversarial support vector machine with the solid line (compare with the non-adversarial support vector machine depicted still by the dashed line). Again, the manipulated data points from the positive class are shown by the black triangles, the original data points from the positive class are shown by red pluses, and the data points from the negative class are shown by the blue squares. By employing the optimal adversarial support vector machine, the classifier can reduce the false negative probability 56% in comparison to the naïve case.

4.2. Heart Disease

In this section, we use the Cardiovascular Disease dataset¹. The dataset contains age, height, weight, gender, Systolic and Diastolic blood pressures, Cholesterol level, Glucose level as well as smoking, drinking, and activity levels of 70,000 individuals. There are two classes of individuals: those with no heart disease (negative class) and those with a heart disease (positive class).

Consider the case where an adversary is interested in misclassifying an individual from the positive class. This could be motivated by that the adversary wants to forge medical documents with minimal changes to pass a life insurance test. For the sake of numerical stability, we scale the data with the inverse of the Cholesky factor of the co-variance matrix of the data. This is just to ensure that all the entries of the data are of the same size. This is indeed a linear transformation whose effect can be compensated for in the linear support vector machine and the linear manipulations of the adversary. Therefore, there is no loss of generality in using this transformation.

To be able to use the developed framework, we fit Gaussian density functions to data points for each class and follow the approach of this paper for designing the classifier and computing the optimal manipulation by the adversary. In what follows, we set $\delta = 10^{-1}$ and $\epsilon = 1$. Table 1 shows the effect of the adversary on the linear support vector machine classifier. Although our Gaussian assumption might not be entirely valid, the optimal non-adversarial support vector machine almost meets the constraints on the true negative probability (with true negative probability of 0.8986 instead of $1 - \delta = 0.9000$). In the naïve case, the adversary can use the ignorance of the classifier to improve the false negative probability by 488%; nearly all individuals from the positive class can be made to pass the test. Following the equilibrium extracted by the best response dynamics, the performance of the adversary is degraded by 35%. This is of course a significant improvement for the classifier. To be able to further reduce the false negative probability, we

¹<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

need to increase δ . This portrays the trade-off that the classifier faces.

5. Conclusions and Future Work

We used a constant-sum game to model the interaction between an adversary and a classifier. For Gaussian data and linear support vector machine classifiers, we transformed the optimization problems of the adversary and the classifier to convex optimization problems. We then utilized best response dynamics to learn an equilibrium of the game in order to extract linear support vector machine classifiers that are robust to adversarial input manipulations.

References

- [1] N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 99–108, ACM, 2004.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015.
- [3] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.
- [4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrđić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.
- [5] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597, IEEE, 2016.
- [6] C. Frederickson, M. Moore, G. Dawson, and R. Polikar, "Attack strength vs. detectability dilemma in adversarial machine learning," in *2018 International Joint Conference on Neural Networks*, pp. 1–8, 2018.
- [7] Y. Han and B. Rubinstein, "Adequacy of the gradient-descent method for classifier evasion attacks," in *Workshops at the 32nd AAAI Conference on Artificial Intelligence*, 2018.
- [8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- [9] P. Russu, A. Demontis, B. Biggio, G. Fumera, and F. Roli, "Secure kernel machines against evasion attacks," in *Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security*, pp. 59–69, 2016.
- [10] A. Sinha, H. Namkoong, and J. Duchi, "Certifiable distributional robustness with principled adversarial training," in *International Conference on Learning Representations (ICLR)*, 2017.
- [11] F. Farokhi, "Regularization helps with mitigating poisoning attacks: Distributionally-robust machine learning using the wasserstein distance," *arXiv preprint arXiv:2001.10655*, 2020.
- [12] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2018.
- [13] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi, "Adversarial support vector machine learning," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1059–1067, 2012.
- [14] S. Weerasinghe, S. M. Erfani, T. Alpcan, and C. Leckie, "Support vector machines resilient against training data integrity attacks," *Pattern Recognition*, vol. 96, p. 106985, 2019.
- [15] A. Globerson and S. Roweis, "Nightmare at test time: robust learning by feature deletion," in *Proceedings of the 23rd international conference on Machine learning*, pp. 353–360, 2006.
- [16] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2617–2654, 2012.
- [17] M. Brückner and T. Scheffer, "Nash equilibria of static prediction games," in *Advances in Neural Information Processing Systems*, pp. 171–179, 2009.
- [18] L. Dritsoula, P. Loiseau, and J. Musacchio, "A game-theoretic analysis of adversarial classification," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 3094–3109, 2017.
- [19] V. P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, pp. 1431–1451, 1982.
- [20] F. Farokhi, A. M. H. Teixeira, and C. Langbort, "Estimation with strategic sensors," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 724–739, 2016.
- [21] J. Farrell and M. Rabin, "Cheap talk," *Journal of Economic Perspectives*, vol. 10, no. 3, pp. 103–118, 1996.
- [22] S. Sarıtaş, S. Yüksel, and S. Gezici, "Quadratic multi-dimensional signaling games and affine equilibria," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 605–619, 2016.
- [23] S. Sarıtaş, S. Yüksel, and S. Gezici, "Dynamic signaling games with quadratic criteria under nash and stackelberg equilibria," *Automatica*, vol. 115, p. 108883, 2020.
- [24] S. Sarıtaş, P. Furrer, S. Gezici, T. Linder, and S. Yüksel, "On the number of bins in equilibria for signaling games," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 972–976, IEEE, 2019.
- [25] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.
- [26] S. Dughmi and H. Xu, "Algorithmic bayesian persuasion," in *Proceedings of the 48th annual ACM symposium on Theory of Computing*, pp. 412–425, ACM, 2016.
- [27] V. S. S. Nadendla, C. Langbort, and T. Başar, "Effects of subjective biases on strategic information transmission," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6040–6049, 2018.
- [28] K. J. Arrow and G. Debreu, "Existence of an equilibrium for a competitive economy," *Econometrica*, pp. 265–290, 1954.
- [29] S. Schaible, "Parameter-free convex equivalent and dual programs of fractional programming problems," *Zeitschrift für Operations Research*, vol. 18, no. 5, pp. 187–196, 1974.
- [30] E. N. Barron, R. Goebel, and R. R. Jensen, "Best response dynamics for continuous games," *Proceedings of the American Mathematical Society*, vol. 138, no. 3, pp. 1069–1083, 2010.
- [31] F. Farokhi, H. Sandberg, I. Shames, and M. Cantoni, "Quadratic Gaussian privacy games," in *2015 54th IEEE conference on decision and control (CDC)*, pp. 4505–4510, IEEE, 2015.