Bayesian Games for Adversarial Regression Problems

Michael Großhans[1] GROSSHAN@CS.UNI-POTSDAM.DE
Christoph Sawade[1] SAWADE@CS.UNI-POTSDAM.DE
Michael Brückner[2] MICHAEL@SOUNDCLOUD.COM
Tobias Scheffer[1] SCHEFFER@CS.UNI-POTSDAM.DE
[1] University of Potsdam, Department of Computer Science, August-Bebel-Straße 89, 14482 Potsdam, Germany
[2] SoundCloud Ltd., Rosenthalerstraße 13, 10119 Berlin, Germany

Abstract

We study regression problems in which an adversary can exercise some control over the data generation process. Learner and adversary have conflicting but not necessarily perfectly antagonistic objectives. We study the case in which the learner is not fully informed about the adversary's objective; instead, any knowledge of the learner about parameters of the adversary's goal may be reflected in a Bayesian prior. We model this problem as a Bayesian game, and characterize conditions under which a unique Bayesian equilibrium point exists. We experimentally compare the Bayesian equilibrium strategy to the Nash equilibrium strategy, the minimax strategy, and regular linear regression.

1. Introduction

Standard regression algorithms are based on the *iid* assumption that data processed at training and application time are governed by identical distributions. In a variety of applications, the input distribution at application time may be influenced by an adversary whose interests are in conflict with those of the learner. For instance, in insurance risk assessment, defrauders continuously tweak specific attributes of their insurance applications to make the risk appear lower. For such applications, the *iid* assumption amounts to the naive assumption that the adversary is entirely passive.

When an adversary can exercise some control over the distribution of the data at application time, the outcome for the learner, as well as the outcome for the adversary, depends on both the predictive model that the learner chooses and the changes that the adversary imposes on the input distribution. Such interleaved optimization problems in which learner and adversary do not exchange information about their intended actions constitute *non-cooperative games*.

In the *zero-sum* case, the goals of learner and adversary are directly antagonistic. This amounts to the assumption that the adversary intends to inflict the greatest possible harm on the learner. In this case, the learner is best off by choosing a *minimax strategy* which is the minimizing argument over the parameter space of the learner, of the maximum over the action space of the adversary, of the cost function. For classification, minimax solutions were derived under several assumptions. Globerson & Roweis (2006) study the case of features that are deleted at test time; El Ghaoui et al. (2003) study features that are changed within an interval. The minimax probability machine (Lanckriet et al., 2002) minimizes the maximal probability of misclassifying new instances for a given mean and covariance matrix of each class. For regression problems, Sayed & Chen (2002) derive a minimax model that handle bounded uncertainty in the feature matrix and labels. The SVM with invariances (Teo et al., 2007) solves a convex upper bound of a minimax optimization problem for arbitrary feature transformations.

If both players have conflicting but not perfectly antagonistic goals, then the minimax strategy is overly pessimistic and does not necessarily lead to an optimal outcome. A *Nash equilibrium point* of a game is a pair of strategies that has the property that unilaterally deviating from it increases the costs for either player. If a game has a unique equilibrium and one assumes that the opponent will also act according to that Nash equilibrium, then acting according to this equilibrium point is the optimal strategy. Identifying an equilibrium point requires *complete information* about the

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

opponent's cost function. For classification games with complete information, Brückner et al. (2012) show that a unique Nash equilibrium point exists if the players' cost functions meet certain conditions.

In security applications, however, the assumption of complete information may still be too strong because the learner may not be fully informed about, for instance, the illicit profit that the adversary can make by passing a computer virus unnoticed through a detection mechanism. A further step of relaxation of the assumptions on the adversary leads to a non-cooperative *game with incomplete information* (Harsanyi, 1968). In this model, complete knowledge of the adversary's cost function is replaced by uncertainty that is expressed in terms of a Bayesian prior over parameters of the cost function.

Finding Equilibrium points in adversarial learning problems has been studied for regret minimization problems where it leads to near-optimal solutions (Freund & Schapire, 1996). If players are uncertain about some parameters of their adversaries' cost function and this uncertainty is expressed in terms of a prior distribution over these parameters, then an equilibrium of the expected costs can be identified using counterfactual regret (Zinkevich et al., 2008). However, both results rely on finite action spaces and finite ranges of the variables that players are uncertain about.

In contrast to regret minimization problems, classification and regression games usually have continuous action spaces. For classification games, the *iid* assumption—the assumption of an entirely passive adversary—has only been relaxed as far as to the point of non-cooperative non-zero-sum games with complete information; for regression, to non-cooperative zero-sum games with complete information. Here, we will extend this sequence of relaxations to non-cooperative non-zero-sum regression games with incomplete information about the adversary's cost function. We will not focus on generalization error bounds, but will instead study equilibrium points of the game defined by the regularized empirical cost functions.

The rest of this paper is organized as follows. Section 2 introduces the players and cost functions for adversarial regression problems formally. Section 3 introduces the game with incomplete information, and the concept of optimal responses and equilibrium points. In Section 4, we show sufficient conditions under which a unique Bayesian equilibrium point exists for regression games. Section 5 derives an algorithm that identifies the unique Bayesian equilibrium point. Section 6 reports on experiments, Section 7 concludes.

2. Players and Cost Functions

We study prediction games between a *learner* of a regression model and a *data generator*, which have conflicting, but not necessarily antagonistic goals.

At *training time*, the data generator produces a training matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ and a vector $\mathbf{y} \in \mathbb{R}^n$ of values of the target variable. The matrix rows and corresponding values of the target variable are governed by an unknown distribution $p(\mathbf{x}, y)$. By contrast, at *application time* the data generator produces instances and values of the target variable according to a distribution $\bar{p}(\mathbf{x}, y)$ which may differ from $p(\mathbf{x}, y)$; these instances are not yet available at training time.

The action of the learner is to select parameters $\mathbf{w} \in \mathcal{W}$ of a linear model $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{w}$. Here, $\mathcal{W}$ is called the learner's action space. We study the action space of all possible parameter vectors, $\mathcal{W} = \mathbb{R}^m$. The learner's theoretical *costs* at application time are given by the expected weighted squared loss

$$\theta_l(\mathbf{w}, \bar{p}, c_l) = \int c_l(\mathbf{x}, y) \left(f_{\mathbf{w}}(\mathbf{x}) - y\right)^2 \mathrm{d}\bar{p}(\mathbf{x}, y),$$

where $c_l(\mathbf{x}, y) \in \mathbb{R}^+$ reflects instance-specific costs.

The data generator's action is to manipulate the data generation process. By changing features of individual instances, the data generator transforms the distribution $p(\mathbf{x}, y)$ at training time into a distribution $\bar{p}(\mathbf{x}, y)$ at application time. The adversary can change features, but cannot change the target value y. Intuitively, a spam sender can make a message look legitimate by adding random text, but cannot change the true nature of the message. This transformation process incurs costs which are quantified by $\Omega_d(p, \bar{p})$. This term acts as a regularizer on the transformation and implicitly constrains the possible discrepancy between the distributions at training and application time.

The data generator may incur costs when the learner classifies an instance $\mathbf{x}$ as y. We model these with a squared-loss term $(f_{\mathbf{w}}(\mathbf{x}) - z(\mathbf{x}, y))^2$ weighted by instance-specific factors $c_d(\mathbf{x}, y) \in \mathbb{R}^+$, where $z(\mathbf{x}, y)$ is the target value that renders the costs for the data generator at zero. For instance, $f_{\mathbf{w}}$ may assess the risk of financial transactions and the data generator may generate a mixture of legitimate and fraudulent transaction requests. The target value would then be $z(\mathbf{x}, y) = 0$ because both legitimate and fraudulent users want their transactions to be found risk-free and executed. For fraudulent transactions, the instance-specific costs $c_d(\mathbf{x}, y)$ are proportional to the gain that a defrauder loses when the transaction $\mathbf{x}$ is declined. For legitimate transactions, the instance-

specific costs may reflect the inconvenience experienced by customers whose transactions are denied. The theoretical *costs* of the data generator add the expected prediction costs to the transformation costs:

$$\theta_d(\mathbf{w}, \bar{p}, c_d) = \int c_d(\mathbf{x}, y) \left(f_{\mathbf{w}}(\mathbf{x}) - z(\mathbf{x}, y)\right)^2 \mathrm{d}\bar{p}(\mathbf{x}, y) + \Omega_d(p, \bar{p}).$$

The theoretical costs of both players depend on the unknown distributions p and $\bar{p}$. We therefore focus on the regularized empirical counterpart of the theoretical costs based on the training sample $(\mathbf{X}, \mathbf{y}, \mathbf{z})$, where $\mathbf{z} = (z_1, \ldots, z_n)^\mathsf{T}$ and $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$ are the empirical quantities of $z(\mathbf{x}, y)$ and y, respectively. The empirical counterpart of the data generator's regularizer $\Omega_d(p, \bar{p})$ penalizes the divergence between training matrix $\mathbf{X}$ and a perturbed sample $\bar{\mathbf{X}}$ that would be the outcome of applying the transformation that translates p into $\bar{p}$ to matrix $\mathbf{X}$. The transformed training matrix $\bar{\mathbf{X}}$ must not be mistaken for test data which are not assumed to be available at training time. The transformed training data $\bar{\mathbf{X}}$ acts as training data from the test distribution. At application time—after both players have committed to their strategies—new instances are drawn according to the distribution $\bar{p}$. In the following, we use the vectors $\mathbf{c}_v = (c_{v,1}, \ldots, c_{v,n})^\mathsf{T}$, where $v \in \{d, l\}$, to denote the players' empirical costs. Then, the empirical costs of predictive model $f_{\mathbf{w}}$ and transformation from p to $\bar{p}$ are:

$$\hat{\theta}_l(\mathbf{w}, \bar{\mathbf{X}}, \mathbf{c}_l) = \sum_{i=1}^{n} c_{l,i} \left(f_{\mathbf{w}}(\bar{\mathbf{x}}_i) - y_i\right)^2 + \Omega_l(f_{\mathbf{w}}), \quad (1)$$

$$\hat{\theta}_d(\mathbf{w}, \bar{\mathbf{X}}, \mathbf{c}_d) = \sum_{i=1}^{n} c_{d,i} \left(f_{\mathbf{w}}(\bar{\mathbf{x}}_i) - z_i\right)^2 + \Omega_d(\mathbf{X}, \bar{\mathbf{X}}). \quad (2)$$

Our analysis will focus on the standard choice of the l_2 regularizer $\Omega_l(f_{\mathbf{w}}) = \|\mathbf{w}\|_2^2$ for the learner and on the squared Frobenius norm of the difference matrix $\Omega_d(\mathbf{X}, \bar{\mathbf{X}}) = \left\|\mathbf{X} - \bar{\mathbf{X}}\right\|_F^2$ for the data generator. Note that we do not need additional regularization parameters that control the trade-off between loss functions and regularization terms for the players because this parameter is implicitly included in the scale of the cost vectors $\boldsymbol{c}_v$.

3. Bayesian Regression Game

Both players' cost functions defined in Equation 1 and 2 depend on both the parameters $\mathbf{w}$ and the transformation manifested in $\bar{\mathbf{X}}$. In general, no single value of $\mathbf{w}$ minimizes the learner's costs independently of the data generator's strategy—which is the characteristic property of a *game*.

In a game *with full information*, one assumes that both players disclose their cost functions to their opponent. We relax this assumption and model the data generator's costs as a parameter that the learner is uncertain about. We relax the full-information assumption asymmetrically: while the adversary maintains full information about the learner's cost function, the learner's full information is relaxed into uncertainty about the adversary's instance-specific costs $\mathbf{c}_d$ which is reflected in a Bayesian prior $q(\mathbf{c}_d)$. This asymmetry reflects our adoption of the learner's perspective: in modeling the learner's lack of information about the adversary, we intend to make the learner more robust against new adversaries. This setting is referred to as a *game with incomplete information* between *Bayesian players* or, for short, a *Bayesian game*.

The tuple $(\mathcal{W}, \Phi, \hat{\theta}_l, \hat{\theta}_d, \mathbf{c}_l, q)$ constitutes a Bayesian game; $\mathcal{W}$, $\hat{\theta}_l$, $\hat{\theta}_d$ and $\mathbf{c}_l$ are defined in Section 2. From the learner's perspective, the costs $\mathbf{c}_d$ are a random variable for which a value is drawn according to prior $q(\mathbf{c}_d)$ at application time. Therefore, to the learner it appears that at training time the data generator commits only to a parametric strategy $\phi : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ that maps a value of $\mathbf{c}_d$—which is only assigned at application time—to a transformation that manifests in matrix $\bar{\mathbf{X}}$. The data generator's action space Φ therefore contains functions ϕ that map from $\mathbb{R}^n$ to $\mathbb{R}^{n \times m}$.

To the learner, the data generator's strategy ϕ is unknown. However, if ϕ were given, then the optimal response to that strategy that minimizes the expected costs over $q(\mathbf{c}_d)$ would be

$$\mathbf{w}^*[\phi] = \arg\min_{\mathbf{w}} \int \hat{\theta}_l(\mathbf{w}, \phi(\mathbf{c}_d), \mathbf{c}_l) \mathrm{d}q(\mathbf{c}_d). \quad (3)$$

In analogy, if $\mathbf{w}$ was known to the data generator, the optimal response for costs $\boldsymbol{c}_d$ would be

$$\phi^*[\mathbf{w}](\mathbf{c}_d) = \arg\min_{\bar{\mathbf{X}}} \hat{\theta}_d(\mathbf{w}, \bar{\mathbf{X}}, \mathbf{c}_d). \quad (4)$$

A pair of parameter vector $\mathbf{w}$ and the data generator's strategy ϕ is called a *Bayesian equilibrium* (Harsanyi, 1967) if it is a fixed point with respect to the optimal response.

Definition 1 (Bayesian Equilibrium)**.** *A pair of* $\mathbf{w}$ *and* ϕ *is called a* Bayesian equilibrium *if it satisfies*

$$(\mathbf{w}, \phi(\mathbf{c}_d)) = (\mathbf{w}^*[\phi], \phi^*[\mathbf{w}](\mathbf{c}_d))$$

for all $\mathbf{c}_d$*, where* $\mathbf{w}^*[\phi]$ *and* $\phi^*[\mathbf{w}](\mathbf{c}_d)$ *are defined in Equations 3 and 4, respectively.*

If q is a single-point distribution, then a Bayesian equilibrium is a Nash equilibrium. Deviating unilaterally from the Bayesian equilibrium increases the costs

for either player—in case of the learner, the expected costs. Therefore, if one player assumes that the opponent plays a Bayesian equilibrium, it is optimal for this player to play the Bayesian equilibrium as well. However, it may be the case that more than one equilibrium exists for a game. If the players choose their actions according to distinct equilibria, then the outcome may be arbitrarily bad for either player. It is therefore crucial to study under which circumstances a regression game has a unique equilibrium.

We now characterize the optimal responses in a way that allows to infer them. In the following, the term $\operatorname{diag}(\mathbf{v})$ denotes a diagonal matrix with elements $(\operatorname{diag}(\mathbf{v}))_{ii} = v_i$ for any arbitrary vector $\mathbf{v}$ and $\mathbf{I}_k$ denotes the identity matrix of size k.

Lemma 1 (Optimal Response of the Data Generator)**.** *Let $\mathbf{X}$ be a matrix of training data, $\mathbf{z}$ the vector of target labels, and $\boldsymbol{c}_d$ a cost vector. Then, the optimal response to a model $\mathbf{w}$ as defined in Equation 4 is uniquely determined by*

$$\phi^*[\mathbf{w}](\boldsymbol{c}_d) = \mathbf{X} - \left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1} (\mathbf{X}\mathbf{w} - \mathbf{z})\,\mathbf{w}^\mathsf{T}.$$

Lemma 2 (Optimal Response of the Learner)**.** *Let $\mathbf{y}$ be the vector of labels and $\boldsymbol{c}_l$ a cost vector. Then, the optimal response to any data generator's strategy ϕ exists and is uniquely determined by*

$$\mathbf{w}^*[\phi] = \left(\mathbf{I}_m + \int \phi(\boldsymbol{c}_d)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_l)\phi(\boldsymbol{c}_d)\mathrm{d}q(\boldsymbol{c}_d)\right)^{-1} \cdot \left(\int \phi(\boldsymbol{c}_d)\mathrm{d}q(\boldsymbol{c}_d)\right)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_l)\mathbf{y}.$$

The proofs are included in the online appendix.

4. Existence and Uniqueness of Equilibrium Points

We will now identify sufficient conditions under which a game $G = (\mathcal{W}, \Phi, \hat{\theta}_l, \hat{\theta}_d, \mathbf{c}_l, q)$ has a unique Bayesian equilibrium. First we will show that for each game G with $\mathcal{W} = \mathbb{R}^m$ and $\Phi = \{\phi : \mathbb{R}^n \to \mathbb{R}^{n\times m}\}$ a game G' with compact and convex action spaces $\mathcal{W}'$ and Φ' can be constructed that has identical equilibrium points. Then, by showing that G' has at least one Bayesian equilibrium we prove that this is also the case for G.

Lemma 3 (Compactness of Action Spaces)**.** *Let $G = (\mathcal{W}, \Phi, \hat{\theta}_l, \hat{\theta}_d, \mathbf{c}_l, q)$ with $\mathcal{W} = \mathbb{R}^m$ and Φ be the set of all functions that map from $\mathbb{R}^n$ to $\mathbb{R}^{n\times m}$. Let the expected values $\int c_{d,i}\mathrm{d}q(\boldsymbol{c}_d) < \infty$ exist. Then, there is a game $G' = (\mathcal{W}', \Phi', \hat{\theta}_l, \hat{\theta}_d, \mathbf{c}_l, q)$ with nonempty, compact, and convex action spaces $\Phi' \subset \Phi$ and $\mathcal{W}' \subset \mathcal{W}$, such that each Bayesian equilibrium in G is a Bayesian equilibrium in G' and vice versa.*

The proof is included in the online appendix. Lemma 3 leads to the the following existence result.

Theorem 1 (Existence of an Equilibrium)**.** *Let the expected value $\int c_{d,i}\mathrm{d}q(\boldsymbol{c}_d) < \infty$ exist. Then, the Bayesian regression game $G = (\mathcal{W}, \Phi, \hat{\theta}_l, \hat{\theta}_d, \boldsymbol{c}_l, q)$ has at least one Bayesian equilibrium.*

Proof. Following Lemma 3 it is sufficient to show that game G' has at least one Bayesian equilibrium. Since the action spaces $\mathcal{W}'$ and Φ' are bounded, $\mathbf{w}^*[\phi]$ is continuous in ϕ (see proof of Lemma 3), and $\phi^*[\mathbf{w}]$ is continuous in $\mathbf{w}$, Brouwers theorem implies that there exists at least one fixed point of $(\phi, \mathbf{w}) \mapsto (\phi^*[\mathbf{w}], \mathbf{w}^*[\phi])$ in $\Phi' \times \mathcal{W}'$. □

We will now derive sufficient conditions for the uniqueness of a Bayesian equilibrium. The equilibrium is unique if $\mathbf{c}_l$ and $\mathbf{c}_d$ are sufficiently small in relation to the regularizers; the exact condition is detailed in Equation 5 and can be validated given $\boldsymbol{c}_l$ and q.

Theorem 2 (Uniqueness of Equilibria)**.** *Let $G' = (\mathcal{W}', \Phi', \hat{\theta}_l, \hat{\theta}_d, \boldsymbol{c}_l, q)$ be a regression game, where $\mathcal{W}'$ and Φ' are nonempty, convex and compact sets. Furthermore, let the expected values $\int c_{d,i}\mathrm{d}q(\boldsymbol{c}_d) < \infty$ exist. Then, G' has a unique Bayesian equilibrium if for all distinct points $(\mathbf{w}, \phi), (\bar{\mathbf{w}}, \bar{\phi}) \in \mathcal{W}' \times \Phi'$:*

$$\begin{aligned}
&\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2 + \int \left\|\phi(\boldsymbol{c}_d) - \bar{\phi}(\boldsymbol{c}_d)\right\|_F^2 \mathrm{d}q(\boldsymbol{c}_d) > \\
&\int \left(\bar{\phi}(\boldsymbol{c}_d)\bar{\mathbf{w}} - \mathbf{y}\right)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_l)\bar{\phi}(\boldsymbol{c}_d)(\mathbf{w} - \bar{\mathbf{w}})\mathrm{d}q(\boldsymbol{c}_d) - \\
&\int \left(\phi(\boldsymbol{c}_d)\mathbf{w} - \mathbf{y}\right)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_l)\phi(\boldsymbol{c}_d)(\mathbf{w} - \bar{\mathbf{w}})\mathrm{d}q(\boldsymbol{c}_d) + \\
&\int \left(\bar{\phi}(\boldsymbol{c}_d)\bar{\mathbf{w}} - \mathbf{z}\right)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_d)(\phi(\boldsymbol{c}_d) - \bar{\phi}(\boldsymbol{c}_d))\bar{\mathbf{w}}\mathrm{d}q(\boldsymbol{c}_d) - \\
&\int \left(\phi(\boldsymbol{c}_d)\mathbf{w} - \mathbf{z}\right)^\mathsf{T} \operatorname{diag}(\boldsymbol{c}_d)(\phi(\boldsymbol{c}_d) - \bar{\phi}(\boldsymbol{c}_d))\mathbf{w}\mathrm{d}q(\boldsymbol{c}_d). \quad (5)
\end{aligned}$$

Theorem 2 generalizes Theorem 8 of Brückner et al. (2012) on the uniqueness of equilibria for games with complete information. If the players' costs $\boldsymbol{c}_d$ and $\boldsymbol{c}_l$ are too large, Equation 5 is violated, the game G is no longer locally convex and multiple equilibria can exist. An experiment on the link between uniqueness and cost parameters can be found in the online appendix.

Proof. The existence of a Bayesian equilibrium follows from Theorem 1. We now reformulate the two-player

game G' into an $(n+1)$-player game G'', where each instance $\mathbf{x}_i$ is transformed by an individual strategy $\phi_i : \mathbb{R} \to \mathbb{R}^m$ and any Bayesian equilibrium in G' corresponds to a distinct Nash equilibrium in G''. Let $G'' = (\mathcal{W}'', \Phi_1'', \ldots, \Phi_n'', \hat{\theta}_l'', \hat{\theta}_{d,1}'', \ldots, \hat{\theta}_{d,n}'')$ be an $(n+1)$-player game without uncertainty, where $\mathcal{W}'' = \mathcal{W}'$ and Φ_i'' is the strategy space Φ' restricted to the data generator of instance i; the functions

$$\hat{\theta}_{d,i}''(\mathbf{w}, \phi_i) = \int \|(\phi_i(c_{d,i}) - \mathbf{x}_i)\|_2^2 \mathrm{d}q(\boldsymbol{c}_d) + \int c_{d,i}(\phi_i(c_{d,i})^\mathsf{T}\mathbf{w} - z_i)^2 \mathrm{d}q(\boldsymbol{c}_d) \text{ and} \tag{6}$$

$$\hat{\theta}_l''(\mathbf{w}, \phi_1, \ldots, \phi_n) = \|\mathbf{w}\|_2^2 + \sum_{i=1}^n \int c_{l,i}(\phi_i(c_{d,i})^\mathsf{T}\mathbf{w} - y_i)^2 \mathrm{d}q(\boldsymbol{c}_d)$$

are the corresponding loss functions for $i = 1, \ldots, n$. Following Theorem 1 of Harsanyi (1968), we now show that if $(\mathbf{w}, (\phi_1, \ldots, \phi_n))$ is a Bayesian equilibrium in G' then $(\mathbf{w}, \phi_1, \ldots, \phi_n)$ is a Nash equilibrium in G''. Suppose that $(\mathbf{w}, \phi_1, \ldots, \phi_n)$ is *not* a Nash equilibrium point in G''. Then, there exists a better strategy for at least one player. If $\mathbf{w}$ is not an optimal choice, the point $(\mathbf{w}, (\phi_1, \ldots, \phi_n))$ is not a Bayesian equilibrium point in G' since the learner's loss functions are equal for fixed data transformation strategies. However, if any ϕ_i is not optimal, then the expected loss of the i-th data generator (see Equation 6) can be reduced; there exists some point $\boldsymbol{c}_d$, where the i-th data generator benefits from changing her strategy unilaterally. Hence, the i-th summand in the loss function $\hat{\theta}_d$ (see Equation 2) decreases while the rest remain unchanged when the other strategies are kept fixed. Consequently, $(\mathbf{w}, (\phi_1, \ldots, \phi_n))$ is not a Bayesian equilibrium point in G'. Hence, it is sufficient to show that there exists at most one Nash equilibrium point in G''.

The data generators' action spaces Φ_i'' constitute Hilbert spaces of square differentiable functions $\phi_i : \mathbb{R} \to \mathbb{R}^m$ on a Lebesgue measurable set with measure $q(\boldsymbol{c}_d)$. The directional Gâteaux derivative of $\hat{\theta}_{d,i}''$ in the direction $d \in \Phi_i''$ is given by

$$\left\langle \partial_{\phi_i} \hat{\theta}_{d,i}'', d \right\rangle_{\Phi_i''} = 2\int \langle \phi_i(c_{d,i}))\rangle \mathrm{d}q(\boldsymbol{c}_d) - 2\int \langle \mathbf{x}_i, d(\boldsymbol{c}_d)\rangle \mathrm{d}q(\boldsymbol{c}_d) + 2\int c_{d,i} \left\langle \mathbf{w}\left(\mathbf{w}^\mathsf{T}\phi_i(c_{d,i}) - z_i\right), d(\boldsymbol{c}_d)\right\rangle \mathrm{d}q(\boldsymbol{c}_d)$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product in $\mathbb{R}^m$. Since $\mathcal{W}'' \subset \mathbb{R}^m$, the learner's directional derivate for any $\mathbf{d} \in \mathcal{W}''$ is given by

$$\left\langle \partial_{\mathbf{w}} \hat{\theta}_l'', \mathbf{d} \right\rangle_{\mathcal{W}''} = \langle \mathbf{w}, \mathbf{d}\rangle + 2\sum_{i=1}^n c_{l,i} \int \langle \phi_i(\boldsymbol{c}_d)\phi_i(\boldsymbol{c}_d)^\mathsf{T}\mathbf{w} - y_i\phi_i(\boldsymbol{c}_d), \mathbf{d}\rangle \mathrm{d}q(\boldsymbol{c}_d).$$

Let $(\mathbf{w}, \phi_1, \ldots, \phi_n), (\bar{\mathbf{w}}, \bar{\phi}_1, \ldots, \bar{\phi}_n) \in \mathcal{W}'' \times \Phi_1'' \times \cdots \times \Phi_n''$ be two distinct points. Then, by Theorem 2.5 of Carlson (2001), a unique equilibrium in G'' exists if

$$\begin{aligned} 0 < &\left\langle \partial_{\mathbf{w}} \hat{\theta}_l''(\mathbf{w}, \phi_1, \ldots, \phi_n), \mathbf{w} - \bar{\mathbf{w}} \right\rangle_{\mathcal{W}''} - \\ &\left\langle \partial_{\mathbf{w}} \hat{\theta}_l''(\bar{\mathbf{w}}, \bar{\phi}_1, \ldots, \bar{\phi}_n), \mathbf{w} - \bar{\mathbf{w}} \right\rangle_{\mathcal{W}''} + \\ &\sum_{i=1}^n \left\langle \partial_{\phi_i} \hat{\theta}_{d,i}''(\mathbf{w}, \phi_i) - \partial_{\phi_i} \hat{\theta}_{d,i}''(\bar{\mathbf{w}}, \bar{\phi}_i), \phi_i - \bar{\phi}_i \right\rangle_{\Phi_i''} \\ = &\langle \mathbf{w} - \bar{\mathbf{w}}, \mathbf{w} - \bar{\mathbf{w}}\rangle + \\ &\sum_{i=1}^n \int k_i(\mathbf{w}, \bar{\mathbf{w}}, \phi_i(\boldsymbol{c}_d), \bar{\phi}_i(\boldsymbol{c}_d)) \mathrm{d}q(\boldsymbol{c}_d), \end{aligned} \tag{7}$$

where the instance-specific terms k_i are given by

$$\begin{aligned} k_i(\mathbf{w}, \bar{\mathbf{w}}, \mathbf{x}, \bar{\mathbf{x}}) = &\langle \mathbf{x} - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}}\rangle + \\ &c_{l,i} \left\langle \left(\mathbf{x}^\mathsf{T}\mathbf{w} - y_i\right)\mathbf{x} - \left(\bar{\mathbf{x}}^\mathsf{T}\bar{\mathbf{w}} - y_i\right)\bar{\mathbf{x}}, \mathbf{w} - \bar{\mathbf{w}} \right\rangle + \\ &c_{d,i} \left\langle \left(\mathbf{x}^\mathsf{T}\mathbf{w} - z_i\right)\mathbf{w} - \left(\bar{\mathbf{x}}^\mathsf{T}\bar{\mathbf{w}} - z_i\right)\bar{\mathbf{w}}, \mathbf{x} - \bar{\mathbf{x}} \right\rangle. \end{aligned} \tag{8}$$

Inserting Equation 8 in 7 and rewriting it in matrix form yields Inequality 5. It is a sufficient condition for the uniqueness of a Nash equilibrium in the proposed Hilbert spaces. In the input space, the data generators can play multiple strategies to reach this unique equilibrium; they differ for costs $q(\boldsymbol{c}_d) = 0$ of probability measure zero. However, exactly one of these strategies corresponds to a Bayesian equilibrium in G' since the data generator has to play optimal for all $\boldsymbol{c}_d$ regardless of the probability distribution q. Hence, there exists at most one Bayesian equilibrium in G'. □

5. Finding the Unique Bayesian Equilibrium

We have derived sufficient conditions for the existence of a unique Bayesian equilibrium in the linear regression game. Since the optimal responses are uniquely defined, the equilibrium in a two-player regression game is already uniquely determined by one single action. Hence, the number of parameters to be estimated in order to find a Bayesian equilibrium can be reduced from $m(n+1)$ to m. Therefore, we now define a surrogate function $w : \mathbf{w} \mapsto \mathbf{w}^*[\phi^*[\mathbf{w}]]$; every fixed point $w(\mathbf{w}) = \mathbf{w}$ with dimension m corresponds to a Bayesian equilibrium $(\mathbf{w}, \phi[\mathbf{w}])$ and vice versa.

Definition 2 (Surrogate Function)**.** *For any weight vector* $\mathbf{w}$ *and a given cost vector* $\boldsymbol{c}_l$*, the function*

$$w(\mathbf{w}; \boldsymbol{c}_l) = \arg\min_{\mathbf{w}'} \int \hat{\theta}_l(\mathbf{w}', \phi^*[\mathbf{w}](\boldsymbol{c}_d), \boldsymbol{c}_l) \mathrm{d}q(\boldsymbol{c}_d)$$

returns the optimal response for a data transformation $\phi^*[\mathbf{w}]$ *which was itself an optimal response to the vector* $\mathbf{w}$*.*

Following Lemma 2, function w can be expressed as

$$w(\mathbf{w}; \boldsymbol{c}_l) = \left(\mathbf{I}_m + \int \phi^*[\mathbf{w}](\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\,(\boldsymbol{c}_l) \phi^*[\mathbf{w}](\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right)^{-1} \cdot \left(\int \phi^*[\mathbf{w}](\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right)^\mathsf{T} \mathrm{diag}\,(\boldsymbol{c}_l)\mathbf{y}. \quad (9)$$

The fixed point of w can be found by fixed-point algorithms, which evaluate the optimal responses and possibly their gradients iteratively. Unfortunately, Equation 9 depends on the matrices $\int \phi^*[\mathbf{w}](\boldsymbol{c}_d)^\mathsf{T} \mathrm{d}q(\boldsymbol{c}_d)$ and $\int \phi^*[\mathbf{w}](\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\,(\boldsymbol{c}_l) \phi^*[\mathbf{w}](\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)$ which have no closed-form solutions for arbitrary prior distributions q. To tackle this problem, we approximate the data generator's optimal response $\phi^*[\mathbf{w}](\boldsymbol{c}_d)$ by the t-th-order Taylor expansion $\phi_{t;\mathbf{a}}[\mathbf{w}](\boldsymbol{c}_d)$ at point $\mathbf{a}$:

$$\phi_{t;\mathbf{a}}[\mathbf{w}](\boldsymbol{c}_d) = \sum_{r=0}^{t} \mathrm{diag}\,(\boldsymbol{c}_d - \mathbf{a})^r \mathbf{C}_r(\mathbf{a}), \text{ where} \quad (10)$$

$$\mathbf{C}_0(\mathbf{a}) = \mathbf{X} - \mathrm{diag}\,(\mathbf{a}) \left(\mathbf{I}_n + \mathrm{diag}\left(\|\mathbf{w}\|_2^2\, \mathbf{a}\right)\right)^{-1} (\mathbf{X}\mathbf{w} - \mathbf{z})\, \mathbf{w}^\mathsf{T},$$

$$\mathbf{C}_i(\mathbf{a}) = (-1)^i \left(\mathbf{I}_n + \mathrm{diag}\left(\|\mathbf{w}\|_2^2\, \mathbf{a}\right)\right)^{-(i+1)} \|\mathbf{w}\|_2^{2(i-1)} (\mathbf{X}\mathbf{w} - \mathbf{z})\, \mathbf{w}^\mathsf{T}$$

for all $0 < i \leq t$. Theorem 3 states that the Approximation 10 becomes more accurate with increasing t if the costs $\boldsymbol{c}_d$ are bounded.

Theorem 3 (Convergence criterion)**.** *Let* $\|\boldsymbol{c}_d\|$ *be bounded from above. Then, there exists a point* $\mathbf{a}$*, such that for all points* $\boldsymbol{c}_d$ *with non-zero density* $q(\boldsymbol{c}_d) > 0$*:*

$$\lim_{t \to \infty} \|\phi^*[\mathbf{w}](\boldsymbol{c}_d) - \phi_{t;\mathbf{a}}[\mathbf{w}](\boldsymbol{c}_d)\|_2 = 0.$$

The proof is included in the online appendix. If the costs $\boldsymbol{c}_d$ are *unbounded*, there is typically some degree t that approximates $\phi^*[\mathbf{w}](\boldsymbol{c}_d)$ better than larger values.

The Taylor expansion requires that the first $2t$ moments of q exist; let $\boldsymbol{\mu}_i(\mathbf{a}) = \int \mathrm{diag}\,(\boldsymbol{c}_d - \mathbf{a})^i \mathrm{d}q(\boldsymbol{c}_d)$ be the i-th central moment around $\mathbf{a}$. Equation 11 defines a surrogate function according to Equation 9, where we made use of the Taylor approximation $\phi_{t;\mathbf{a}}[\mathbf{w}](\boldsymbol{c}_d)$ of the response $\phi^*[\mathbf{w}](\boldsymbol{c}_d)$ of the data generator:

$$w_{t;\mathbf{a}}(\mathbf{w}; \boldsymbol{c}_l) = \left(\mathbf{I}_m + \sum_{r,s=0}^{t} \mathbf{C}_r(\mathbf{a})^\mathsf{T} \mathrm{diag}\,(\boldsymbol{c}_l) \boldsymbol{\mu}_{r+s}(\mathbf{a}) \mathbf{C}_s(\mathbf{a})\right)^{-1} \left(\sum_{r=0}^{t} \mathbf{C}_r(\mathbf{a}) \boldsymbol{\mu}_r(\mathbf{a})\right) \mathrm{diag}\,(\boldsymbol{c}_l)\mathbf{y}. \quad (11)$$

The existence of the inverse matrices in Equations 9 and 11 follows from Lemma 2.

The choice of the pivotal point $\mathbf{a}$ influences the radius of convergence (see proof of Theorem 3); for small values of t, it should be located in a high-density region of q. We will now derive an algorithm that infers a Bayesian equilibrium based on the Taylor expansion at $\mathbf{a} = \mathbb{E}[\boldsymbol{c}_d]$. A fixed point of $w_{t;\mathbb{E}[\boldsymbol{c}_d]}(\cdot; \boldsymbol{c}_l)$ can be found by simplex algorithms (see, *e.g.,* Van der Laan & Talman, 1982), which, unfortunately, have exponential worst-case execution time (Hirsch et al., 1989). On the other hand, standard gradient descend approaches guarantee only local convergence and are not robust against poor starting points. We use a graduated optimization algorithm to find the fixed point.

Algorithm 1 uses a sequence of increasing costs, terminating at the original costs of the learner. For function FixedPoint, we use a Newton-like method that approximates the Jacobi matrix by difference quotients. The k-th fixed point is a Taylor approximation to the Bayesian equilibrium (Line 5).

Algorithm 1 Bayesian Equilibrium by Graduated Optimization

Input: Sequence of costs $\mathbf{c}_{l,1}, \ldots, \mathbf{c}_{l,k}$, where $\|\mathbf{c}_{l,i}\| < \|\mathbf{c}_{l,i+1}\|$, $\mathbf{c}_{l,1} = \mathbf{0}^n$; $\mathbf{c}_{l,k} = \mathbf{c}_l$; Taylor degree t
Output: Bayesian equilibrium $(\mathbf{w}^*, \phi^*)$

1: $\mathbf{w}_0 \leftarrow \mathbf{0}^m$
2: **for** $i = 1, \ldots, k$ **do**
3: $\quad \mathbf{w}_i \leftarrow \textsc{FixedPoint}(w_{t;\mathbb{E}[\boldsymbol{c}_d]}(\cdot; \mathbf{c}_{l,i}), \mathbf{w}_{i-1})$
4: **end for**
5: return $(\mathbf{w}_k, \phi^*[\mathbf{w}_k])$

6. Experimental Evaluation

In this section we study the behavior of the Bayesian regression model in the context of email spam filtering. Our motivating application is to predict the fraction of users who perceive an email as unwanted. We collected about 190,000 emails from an email service provider between September 2007 and December 2008.

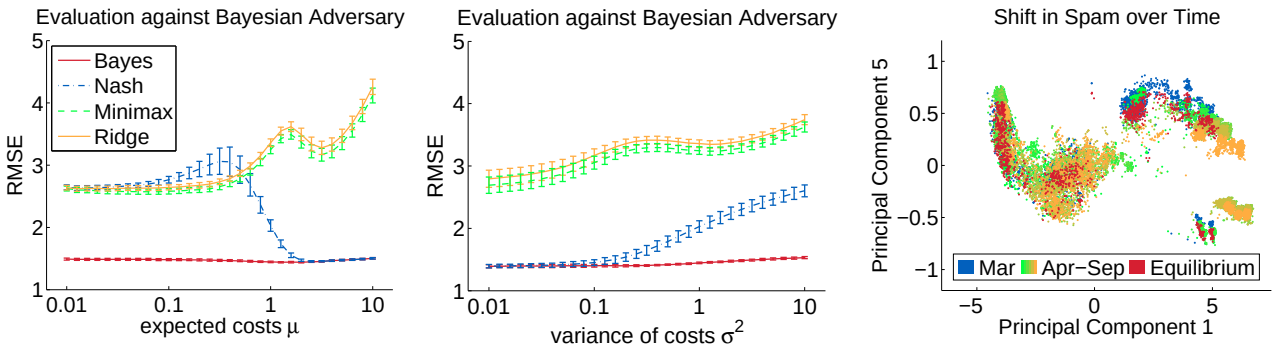

Figure 1. Evaluation against an adversary that follows a Bayesian equilibrium strategy for varying cost parameters (left/center). Shift in spam mails over time (right). Error bars show standard errors.

We approximate the actual fraction of users who perceive the message as spam by the log-likelihood for the class spam of a classifier; we train classifiers using ten-fold cross validation on all data and label the held-out data in each iteration with the log-likelihood inferred by the classifier. Emails are represented by the first ten principal components ($m = 10$) of the binary bag-of-words features (around 226,000 words). The target score $z = 0$ reflects that all senders desire their emails to be perceived as non-spam by all recipients. We measure the root mean squared error (RMSE).

For the Bayesian regression model, we use the first-order Taylor approximation, $t = 1$ (except when we study different values of t). For $t = 1$, the Bayesian regression model only depends on mean value μ and variance σ^2 of the prior q; the distributional assumption about $q(c_{d,i})$ has no additional influence on the equilibrium. For higher values of t, we use a gamma distribution. We compare the Bayesian regression model (denoted *Bayes*) to three reference methods: the Nash regression model that emerges as a special case for one-point distributions q (denoted *Nash*), robust ridge regression (denoted *Minimax*; Sayed & Chen 2002), and a regular ridge regression (denoted *Ridge*). We set the Nash model's conjecture for all values of $c_{d,i}$ to the mean value μ. The perturbation parameter for *Minimax* is chosen as minimal value such that the space of possible transformations of the input matrix still includes the solutions of *Bayes* and *Nash*.

In the first experiment, we study how the methods perform against an adversary that chooses a strategy according to a Bayesian equilibrium for varying parameters of $q(\boldsymbol{c}_d)$. In each repetition, we compute two Bayesian equilibrium points on separate, disjoint sets drawn at random from September 2007. We extract the learner's model from the first, and transformed data points from the second equilibrium point after drawing actual costs from $q(c_{d,i})$ and playing according to Lemma 1. The regularization parameters of all methods are set to match the learner's costs of $c_{l,i} = 0.1$. Figure 1 shows the RMSE for varying expected values (left, with fixed variances $\sigma^2 = 1$) and variances (center, with fixed $\mu = 1$) of the data generator's costs, averaged over ten training samples of size 200. We observe that *Bayes* outperforms the *iid* baselines consistently. The advantage of *Bayes* over *Nash* grows with the variance of q. This is plausible because the Bayesian game accounts for uncertainty on the data generator's costs whereas the Nash model assumes that all values are μ. More details are documented in the online appendix.

In a second experiment, we evaluate all methods into the future. Here, the models play against actual spammers. The training sample of 200 instances is drawn from month k. The regularization parameters of all learners (for *Nash* and *Bayes*, we use a single cost parameter for all instances) are tuned on 1,000 instances from month $k + 1$. Test data are drawn from months $k+2$ to $k+6$. Additionally, in order to artificially create a mismatch to the assumed adversity of the data generator, we also evaluate the models on test data drawn in reverse chronological order; for evaluation into the past, tuning data are drawn from month $k-1$ and test data are drawn from months $k-2$ to $k-6$. This process is repeated and RMSE measurements are averaged over ten resampling iterations of the training set and, in an outer loop, over four training months k (March to June 2008). The data generator's costs parameters are set to $\mu = 0.01$ and $\sigma^2 = 0.01$ for *Bayes* and to $\mu = 0.01$ for *Nash*.

Figure 2 (left) shows that *Bayes* and *Nash* are more robust over time; they outperform the minimax and ridge regression models for emails received at least two months after training. Additionally, *Bayes* significantly outperforms *Nash*. When test data are drawn in reverse chronological order, the relative performance of

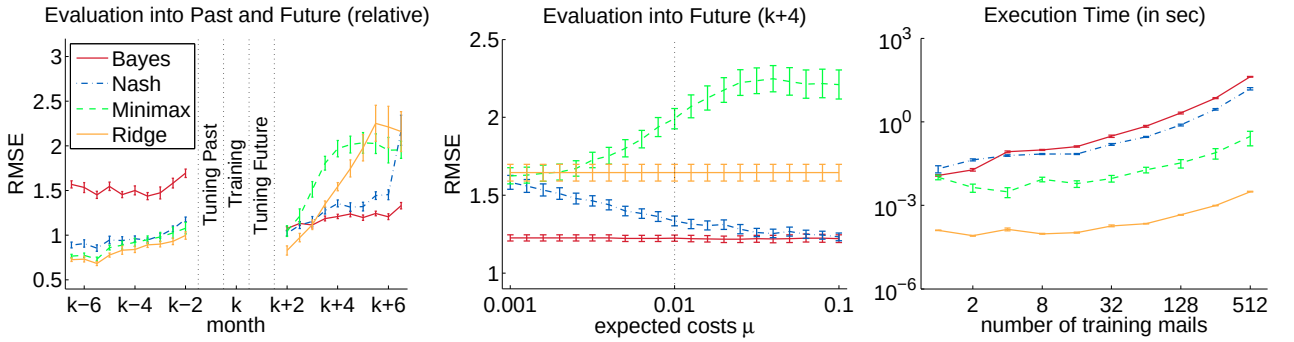

Figure 2. Evaluation of regression models with fixed expected costs into the past and future (left) and varying expected costs into the future (center). Execution Time (right). Error bars show standard errors.

Bayes, *Nash* and *Ridge* is reversed. This corresponds to the mismatch between the assumed adversity of the test data and the actual tendency of historic email spam to be less innovative and difficult. Barely regularized ridge regression excels in this setting.

Figure 2 (center) shows experiments in which we measure the RMSE for a fixed point in the future over a range of values for μ. The curve shows that *Bayes* is robust with respect to the parameter μ of prior q due to the dominating variance. The online appendix includes more details.

Figure 1 (right) visualizes the actual chronological shift of spam emails and compares it to the equilibrium point. The axes are the two most discriminative principal components of the input space. We train a *Nash* model on 200 instances from March 2008; the green to yellow dots visualize the actual chronological shift of class spam. The red dots visualize the training data, transformed according to the equilibrium point given by Lemma 1. Generally, the equilibrium anticipates the principal trend of the data shift. In the lower right-hand corner, an entirely new cluster emerges in the test data that is not present in the equilibrium. Similar experiments for April to June 2008 are included in the online appendix.

All game-theoretical models are computationally more intense. Figure 2 (right) shows the execution time over the number of training emails. The execution time as a function of the number of attributes can be found in the online appendix. *Bayes* computes multiple fixed points using the gradient of the optimal responses. The optimal responses in a Nash game ignore the variance of the data generator's costs leading to a simplified optimization problem. *Minimax* has to solve a costly inner optimization problem to determine the worst perturbation.

Finally, we study the impact of the degree t of the Taylor approximation on the accuracy and execution time of *Bayes*. Figure 6 in the online appendix shows that its influence on the RMSE is minimal; $t = 3$ gives marginally lower RMSE values than 2 and 1. The execution time for $t = 3$ is by a constant factor of approximately 3 higher than for $t = 1$.

7. Conclusion

Previous work on adversarial classification has relaxed the *iid* assumption—the assumption of an entirely passive adversary—as far as to the point of non-cooperative *non-zero-sum* games with *complete information* (Brückner et al., 2012); for adversarial regression, to non-cooperative *zero-sum* games with *complete information* (Sayed & Chen, 2002); and for finite action spaces, to non-cooperative *zero-sum* games with *incomplete information* (Zinkevich et al., 2008). This paper extends this to non-cooperative *non-zero-sum* regression games with *incomplete information* about the cost function of the adversary. We have shown that regression games have at least one Bayesian equilibrium, and that the equilibrium is unique when the cost functions are sufficiently strongly regularized. We derived an algorithm that identifies the unique Bayesian equilibrium. From our experiments, we conclude that the Bayesian model achieves a smaller RMSE than the Nash model, the minimax model and ridge regression when playing against a Bayesian adversary for email data. When evaluating against actual future emails, the Bayesian models predict the log-likelihood of the class spam for future emails more accurately than the Nash, minimax, and ridge regression models.

Acknowledgment

This work was supported by the German Science Foundation DFG under grant SCHE 540/12-1.

References

Brückner, M., Kanzow, C., and Scheffer, T. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13:2589–2626, 2012.

Carlson, D. A. The existence and uniqueness of equilibria in convex games with strategies in Hilbert spaces. *Advances in Dynamic Games and Applications*, 6:79–97, 2001.

El Ghaoui, L., Lanckriet, G., and Natsoulis, G. Robust classification with interval data. Technical Report UCB/CSD-03-1279, Computer Science Division (EECS), University of California, 2003.

Freund, Y. and Schapire, R. Game theory, on-line prediction and boosting. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, 1996.

Globerson, A. and Roweis, S. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Harsanyi, J. C. Games with incomplete information played by Bayesian players, i-iii. part i. the basic model. *Management Science*, 14(3):159–182, 1967.

Harsanyi, J. C. Games with incomplete information played by Bayesian players, i-iii. part ii. Bayesian equilibrium points. *Management Science*, 14(5): 320–334, 1968.

Hirsch, M. D., Papadimitriou, C. H., and Vavasis, S. A. Exponential lower bounds for finding Brouwer fix points. *Journal of Complexity*, 5(4):379–416, 1989.

Lanckriet, G., El Ghaoui, L., Bhattacharyya, C., and Jordan, M. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3: 552–582, 2002.

Sayed, A. H. and Chen, H. A uniqueness result concerning a robust regularized least-squares solution. *Systems and Control Letters*, 46(5):361–369, 2002.

Teo, C. H., Globerson, A., Roweis, S., and Smola, A. Convex learning with invariances. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, 2007.

Van der Laan, G. and Talman, A. J. J. On the computation of fixed points in the product space of unit simplices and an application to noncooperative n person games. *Mathematics of Operations Research*, 7(1):1–13, 1982.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems*, 2008.