Bayesian Games for Adversarial Regression Problems (Online Appendix)

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

A. Proofs

Proof of Lemma 1

The partial derivative of Equation 4 for an instance i is given by

$$\frac{\partial}{\partial \bar{\mathbf{x}}_i} \hat{\theta}_d(\mathbf{w}, \bar{\mathbf{X}}, \boldsymbol{c}_d) = 2c_{d,i} \left(\bar{\mathbf{x}}_i^\mathsf{T}\mathbf{w} - z_i\right) \mathbf{w} + 2\left(\bar{\mathbf{x}}_i - \mathbf{x}_i\right).$$

Due to the convexity of $\hat{\theta}_d$ we can compute the minimum by equating it to zero and solving for $\bar{\mathbf{x}}_i$ using the Sherman-Morrison formula. This results in

$$\bar{\mathbf{x}}_i^* = \mathbf{x}_i - \left(c_{d,i}^{-1} + \|\mathbf{w}\|_2^2\right)^{-1} \left(\mathbf{x}_i^\mathsf{T}\mathbf{w} - z_i\right) \mathbf{w}.$$

The claim follows, since the optimal transformations are independent from each other. □

Proof of Lemma 2

The partial derivate of Equation 3 is given by

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{w}} \int \hat{\theta}_l(\mathbf{w}, \phi(\boldsymbol{c}_d), \boldsymbol{c}_l) \mathrm{d}q(\boldsymbol{c}_d) \\
&= \frac{\partial}{\partial \mathbf{w}} \int \mathrm{diag}\left(\boldsymbol{c}_l\right)(\phi(\boldsymbol{c}_d)\mathbf{w} - \mathbf{y})^\mathsf{T}(\phi(\boldsymbol{c}_d)\mathbf{w} - \mathbf{y}) \mathrm{d}q(\boldsymbol{c}_d) \\
&\quad + \frac{\partial}{\partial \mathbf{w}} \|\mathbf{w}\|_2^2 \\
&= 2\left(\int \phi(\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\left(\boldsymbol{c}_l\right) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right) \mathbf{w} - \\
&\quad 2\left(\int \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right)^\mathsf{T} \mathrm{diag}\left(\boldsymbol{c}_l\right) \mathbf{y} + 2\mathbf{w}.
\end{aligned}$$

Setting this gradient equal to zero and solving it for $\mathbf{w}$ yields

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

The matrix $\phi(\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\left(\boldsymbol{c}_l\right) \phi(\boldsymbol{c}_d)$ and hence the expectation $\int \phi(\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\left(\boldsymbol{c}_l\right) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)$ is positive semi-definite for any adversary's strategy $\phi(\boldsymbol{c}_d)$. Thus, all eigenvalues $\lambda_i \geq 0$ are non-negative. The corresponding eigenvectors are additionally eigenvectors of $\mathbf{I}_m + \int \phi(\boldsymbol{c}_d)^\mathsf{T} \mathrm{diag}\left(\boldsymbol{c}_l\right) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)$ with eigenvalues $1 + \lambda_i > 0$. Thus, the inverse matrix exists independently of the choice of ϕ. This proves the claim. □

Proof of Lemma 3

We define the action spaces of G' as the closure of the convex hull of optimal responses

$$\begin{aligned}
\Phi' = \mathsf{cl}(\{\phi \in \Phi \mid \phi = \tau\phi^*[\mathbf{w}_1] + (1-\tau)\phi^*[\mathbf{w}_2], \\
\tau \in [0,1], \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}\}) \\
\mathcal{W}' = \mathsf{cl}(\{\mathbf{w} \in \mathcal{W} \mid \mathbf{w} = \tau\mathbf{w}^*[\phi_1] + (1-\tau)\mathbf{w}^*[\phi_2], \\
\tau \in [0,1], \phi_1, \phi_2 \in \Phi\}),
\end{aligned}$$

where $\mathsf{cl}(M)$ denotes the closure of set M. Since a Bayesian equilibrium is a pair of optimal responses (see Definition 1), each equilibrium in the original game G is a member of the restricted space $\mathcal{W}' \times \Phi'$. Since both games have identical loss functions and costs distributions each equilibrium point in G' is an equilibrium point G and vice versa.

It remains to be shown that the spaces Φ' and $\mathcal{W}'$ of optimal responses are bounded. We start by showing that any optimal response of the data generator is

bounded. Using the triangle inequality, the spectral norm of an optimal response to a model parameter $\mathbf{w}$ (see Lemma 1) can be upper-bounded by

$$\begin{aligned}\|\phi^*[\mathbf{w}](\boldsymbol{c}_d)\|_2 \leq \\ \|\mathbf{X}\|_2 + \left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1} \mathbf{z}\mathbf{w}^\intercal\right\|_2 + \\ \left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1} \mathbf{X}\mathbf{w}\mathbf{w}^\intercal\right\|_2 . \quad (14)\end{aligned}$$

The third summand of the right-hand side in Inequality 14 can be upper bounded as follows.

$$\begin{aligned}&\left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1} \mathbf{X}\mathbf{w}\mathbf{w}^\intercal\right\|_2 \\ &\quad\leq \left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1}\right\|_2 \|\mathbf{X}\|_2 \left\|\mathbf{w}\mathbf{w}^\intercal\right\|_2 \quad (15)\\ &\quad= \frac{1}{\|\mathbf{w}\|_2^2 + \min_i \frac{1}{c_{d,i}}} \|\mathbf{X}\|_2 \|\mathbf{w}\|_2^2 \quad (16)\\ &\quad< \|\mathbf{X}\|_2 . \quad (17)\end{aligned}$$

Equation 15 follows by the sub-multiplicativity of the spectral norm. The spectral norm of a symmetric positive definite matrix is given by its largest eigenvalue, that is, the maximal diagonal entry of the positive diagonal matrix (see Equation 16). Equation 17 holds, since $\boldsymbol{c}_d > \mathbf{0}$.

Analogously, the second summand (see Inequality 14) can be bounded using the sub-multiplicativity of the spectral norm (Inequality 18) and evaluating the largest eigenvalue of the diagonal matrix, which we split up into two factors (see Equation 19). In Inequality 20 we can bound each dominator by one of the positive summands.

$$\begin{aligned}&\left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1} \mathbf{z}\mathbf{w}^\intercal\right\|_2 \\ &\quad\leq \left\|\left(\operatorname{diag}(\boldsymbol{c}_d)^{-1} + \|\mathbf{w}\|_2^2 \mathbf{I}_n\right)^{-1}\right\|_2 \|\mathbf{z}\|_2 \|\mathbf{w}\|_2 \quad (18)\\ &\quad= \frac{\|\mathbf{z}\|_2}{\sqrt{\|\mathbf{w}\|_2^2 + \min_i \frac{1}{c_{d,i}}}} \frac{\|\mathbf{w}\|_2}{\sqrt{\|\mathbf{w}\|_2^2 + \min_i \frac{1}{c_{d,i}}}} \quad (19)\\ &\quad< \frac{\|\mathbf{z}\|_2}{\sqrt{\min_i \frac{1}{c_{d,i}}}} \frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_2} \quad (20)\\ &\quad= \|\mathbf{z}\|_2 \max_i \sqrt{c_{d,i}} \quad (21)\end{aligned}$$

Finally, using Bound 17 and 21 the function space Φ is bounded by

$$\int \|\phi[\mathbf{w}](\boldsymbol{c}_d)\|_2 \mathrm{d}q(\boldsymbol{c}_d) < \int 2\|\mathbf{X}\|_2 + \|\mathbf{z}\|_2 \max_i \sqrt{c_{d,i}} \mathrm{d}q(\boldsymbol{c}_d)$$

since $\int \max_i \sqrt{c_{d,i}} \mathrm{d}q(\boldsymbol{c}_d) < \sum_i^n \int (1 + c_{d,i}) \mathrm{d}q(\boldsymbol{c}_d) < n + \sum_i^n \int c_{d,i} \mathrm{d}q(\boldsymbol{c}_d) < \infty$ exist.

We now show that if the data generator's action space is convex and compact the space of optimal responses $\mathcal{W}'$ of the learner is compact and convex as well. Let $\phi(\boldsymbol{c}_d)$ be a data generator's strategy given costs $\boldsymbol{c}_d$. Then, the learner's optimal response given by Lemma 2 can be bounded using the sub-multiplicativity of the l^2-norm

$$\begin{aligned}&\|\mathbf{w}^*[\phi]\|_2 \\ &\quad\leq \left\|\left(\mathbf{I}_m + \int \phi(\boldsymbol{c}_d)^\intercal \operatorname{diag}(\boldsymbol{c}_l) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right)^{-1}\right\|_2 \\ &\qquad\left\|\left(\int \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)\right)\right\|_2 \|\operatorname{diag}(\boldsymbol{c}_l)\mathbf{y}\|_2 \\ &\quad\leq \sup_{\phi' \in \Phi} \left\{\int \|\phi'(\boldsymbol{c}_d)\|_2 \mathrm{d}q(\boldsymbol{c}_d)\right\} \|\operatorname{diag}(\boldsymbol{c}_l)\mathbf{y}\|_2 . \quad (22)\end{aligned}$$

For any arbitrary symmetric, positive definite matrix the squared spectral norm equals its largest eigenvalue. Since $\int \phi(\boldsymbol{c}_d)^\intercal \operatorname{diag}(\boldsymbol{c}_l) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)$ is positive semi-definite, all eigenvalues $\lambda_i \geq 0$ are non-negative. The corresponding eigenvectors are additionally eigenvectors of $\mathbf{I}_m + \int \phi(\boldsymbol{c}_d)^\intercal \operatorname{diag}(\boldsymbol{c}_l) \phi(\boldsymbol{c}_d) \mathrm{d}q(\boldsymbol{c}_d)$ with eigenvalues $1 + \lambda_i \geq 1$. Thus, the inverse matrix exists and its norm can be upper bounded by one (see Inequality 22). Then, the claim follows since the data generator's action space is bounded. □

Proof of Theorem 3

Let $\mathbf{a} = (a_1, \ldots, a_n)^\intercal$ be an arbitrary vector. Then, following Taylor's Theorem the data generator's optimal response can be written as

$$\begin{aligned}\phi^*[\mathbf{w}](\boldsymbol{c}_d) &= \phi_{t;\mathbf{a}}[\mathbf{w}](\boldsymbol{c}_d) + R_{t;\mathbf{a}}(\boldsymbol{c}_d), \\ &= \sum_{r=0}^{t} \operatorname{diag}(\boldsymbol{c}_d - \mathbf{a})^r \mathbf{C}_r(\mathbf{a}) + R_{t;\mathbf{a}}(\boldsymbol{c}_d),\end{aligned}$$

with Cauchy remainder

$$R_{t;\mathbf{a}}(\boldsymbol{c}_d) = (t+1)\operatorname{diag}(\boldsymbol{c}_d - \boldsymbol{\xi})^t \operatorname{diag}(\boldsymbol{c}_d - \mathbf{a}) \mathbf{C}_{t+1}(\boldsymbol{\xi}),$$

for some $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\intercal$ with $\xi_i \in [c_{d,i}, a_i]$ or $\xi_i \in [a_i, c_{d,i}]$, respectively. It remains to show that $\mathbf{a}$ can be chosen, such that $R_{t;\mathbf{a}}(\boldsymbol{c}_d)$ converges to the null matrix, or equivalently, that $\lim_{t \to \infty} \|R_{t;\mathbf{a}}(\boldsymbol{c}_d)\|_2 = 0$ holds.

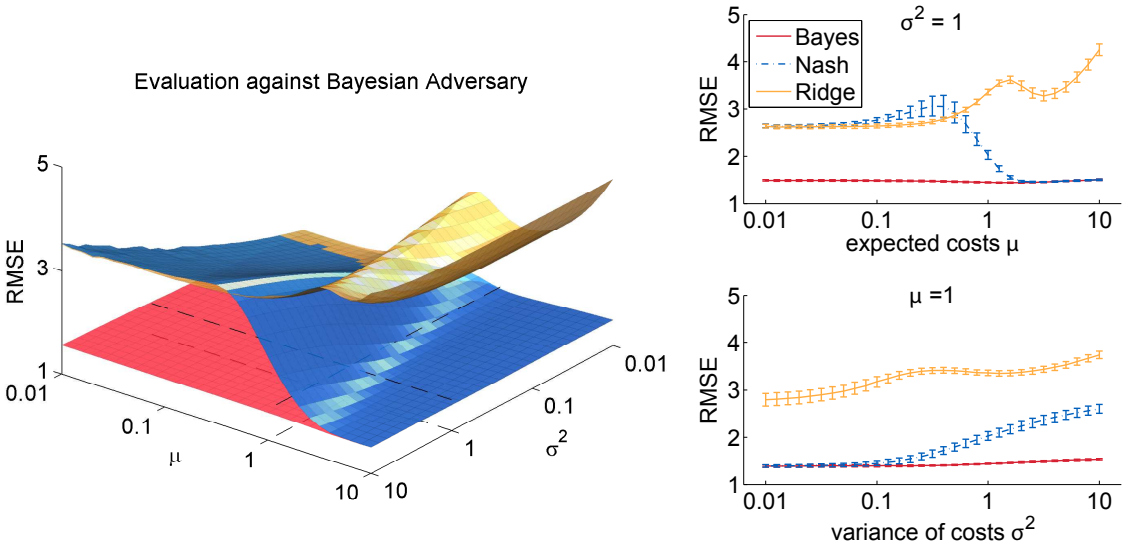

Figure 3. Evaluation against an adversary that follows a Bayesian equilibrium strategy for varying cost parameters. Error bars show standard errors.

The limit of the Cauchy remainder can be expressed as

$$\begin{aligned}
&\lim_{t\to\infty} \|R_{t;\mathbf{a}}(\boldsymbol{c}_d)\|_2 \\
&= \lim_{t\to\infty} \left\|(t+1)\mathrm{diag}\left(\boldsymbol{c}_d - \boldsymbol{\xi}\right)^t \mathrm{diag}\left(\boldsymbol{c}_d - \mathbf{a}\right)\mathbf{C}_{t+1}(\boldsymbol{\xi})\right\|_2 \\
&\propto \lim_{t\to\infty} \Big\|(t+1)\mathrm{diag}\left(\boldsymbol{c}_d - \boldsymbol{\xi}\right)^t \mathrm{diag}\left(\boldsymbol{c}_d - \mathbf{a}\right) \\
&\qquad \left(\mathbf{I}_n + \mathrm{diag}\left(\|\mathbf{w}\|_2^2\,\mathbf{a}\right)\right)^{-(t+2)} \|\mathbf{w}\|_2^{2t}\Big\|_2 \qquad (23) \\
&= \lim_{t\to\infty} \max_i \Big|(t+1)\left(c_{d,i} - \xi_i\right)^t \left(c_{d,i} - a_i\right) \\
&\qquad \left(1 + \|\mathbf{w}\|_2^2\, a_i\right)^{-t-2} \|\mathbf{w}\|_2^{2t}\Big|. \qquad (24)
\end{aligned}$$

In Equation 23 we dismiss the two terms $(-1)^{t+1}$ and $(\mathbf{X}\mathbf{w} - \mathbf{z})\,\mathbf{w}^\mathsf{T}$ from $\mathbf{C}_{t+1}(\boldsymbol{\xi})$ (see Equation 10). This leads to a diagonal matrix whose spectral norm is given by the maximal absolute diagonal entry of the matrix (see Equation 24). If $\|\mathbf{w}\| = 0$ the claim holds. So let $\|\mathbf{w}\| > 0$. Then, Equation 24 can be factorized as

$$\begin{aligned}
&\lim_{t\to\infty} \|R_{t;\mathbf{a}}(\boldsymbol{c}_d)\|_2 \propto \\
&\lim_{t\to\infty} \max_i \Bigg|(t+1)\left(\frac{c_{d,i} - \xi_i}{\|\mathbf{w}\|_2^{-2} + a_i}\right)^t \\
&\qquad \left(1 + \|\mathbf{w}\|_2^2\, a_i\right)^{-2} \left(c_{d,i} - a_i\right)\Bigg|. \qquad (25)
\end{aligned}$$

A sufficient condition for the geometric sequence and thus for Equation 25 to tend to zero is that

$$\left|\frac{c_{d,i} - \xi_i}{\|\mathbf{w}\|_2^{-2} + \xi_i}\right| \leq k < 1 \qquad (26)$$

holds for all i and a fixed k. In the following we show by case differentiation according to $c_{d,i}$ that Condition 26 is fulfilled for $a_i = \frac{1}{2}\sup\{\|\boldsymbol{c}_d\|_\infty | q(\boldsymbol{c}_d) > 0\}$, where $\|\mathbf{c}\|_\infty = \max_i(|c_i|)$ is the maximum norm. Let $a_i \leq c_{d,i}$. Since $\xi_i \in [a_i, c_{d,i}]$ it follows that $a_i \leq \xi_i \leq c_{d,i} \leq 2a_i$ and, thus, the quotient can be upper-bounded by

$$\left|\frac{\xi_i - c_{d,i}}{\|\mathbf{w}\|_2^{-2} + \xi_i}\right| \leq \frac{c_{d,i} - \xi_i}{\|\mathbf{w}\|_2^{-2} + a_i} \leq \frac{2a_i - a_i}{\|\mathbf{w}\|_2^{-2} + a_i} = k.$$

Since $\|\mathbf{w}\|_2 > 0$, it holds that $k < 1$. Now assume that $a_i > c_{d,i}$. Since the data generator's costs are bounded from below by zero, it follows that $0 \leq c_{d,i} \leq \xi_i \leq a_i$. Hence, the quotient can be upper-bounded by

$$\left|\frac{\xi_i - c_{d,i}}{\|\mathbf{w}\|_2^{-2} + \xi_i}\right| \leq \frac{\xi_i}{\|\mathbf{w}\|_2^{-2} + \xi_i} \leq \frac{a_i}{\|\mathbf{w}\|_2^{-2} + a_i} = k.$$

This proves the claim. □

B. Comprehensive Empirical Results

In Section 6 we studied the behavior of the Bayesian regression model in the context of email spam filtering. We compared the Bayesian game regression

model (denoted *Bayes*) to the Nash game regression model (denoted *Nash*), the robust ridge regression (denoted *Minimax*), and a regular ridge regression (denoted *Ridge*).

Depiction of Theorem 2

In Theorem 2 we derived sufficient conditions for unique equilibrium points. The uniqueness depends on both players' costs $\boldsymbol{c}_d$ and $\boldsymbol{c}_l$. Figure 6 (left) depicts the optimal responses of the learner (horizontal axis) and the data generator (vertical axis) for different costs $\boldsymbol{c}_l = \boldsymbol{c}_d$ in a one-dimensional regression game where $\mathbf{X} = 1, y = 1,$ and $z = 1$ without uncertainty. Their intersections constitute the equilibrium of the corresponding game. If the costs become too large, such that Condition 5 is violated for any $(w, \phi) \in [0,1] \times [0,1]$, the game G is no longer locally convex (indicated by the dashed black line).

Playing against a Bayesian adversary

In a first experiment, we evaluated how the methods perform against an adversary that chooses a strategy according to a Bayesian equilibrium. We choose $q(c_{d,i})$ as a gamma distribution and varied its mean μ and the variance σ^2. We set the Nash model's conjecture for all values of $c_{d,i}$ to the mean μ; this is optimal if the data generator's costs are drawn from a single-point distribution $q(c_{d,i}) = \delta(c_{d,i} = \mu)$. Figure 3 shows the root mean squared error (RMSE) for *Bayes*, *Nash*, and *Ridge* as a function of μ and σ^2 (left). Figure 3 (right) shows a sectional view along μ (top) and along σ^2 (bottom). We observe that all methods coincide if the data generator's costs vanish. The advantage of Bayes over Nash grows with the variance of q.

Playing against actual adversaries

In a second experiment, we evaluate all methods over time into the future. Here, the models play against actual spammers. Additionally, in order to artificially create a mismatch to our modeling assumptions, we also evaluate the models on test data from the past. The training sample of 200 instances are drawn from month k. To evaluate into the future, the regularization parameters of all learners (for *Bayes*, we use a single cost parameter for all instances) are tuned on 1,000 instances from month $k+1$; for evaluation into the past, tuning data are drawn from month $k-1$. Test data are then drawn from months $k+2$ to $k+1$ and from months $k-2$ to $k-6$, respectively. 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 4 (left) shows the RMSE over time for fixed mean $\mu = 0.01$ and variance $\sigma^2 = 0.01$. Figure 4 (center) depicts the RMSE for a fixed point in the past over a range of different values for μ. Figure 4 (right) shows the training times of the Bayesian equilibrium model and reference models for varying number of attributes.

Additionally we study the shift in spam mails in reality and compare it with the computed equilibrium point. For depiction we choose the two most discriminant principal components with respect to spam and non-spam. We train separate *Nash* models on March 2007 (see section 6), April 2007, May 2007 and June 2007. Again we use 200 instances and the data generator's costs are set to $\mu = 0.01$. The learner's costs are tuned on the subsequent month. Given the *Nash* model we are able to extract the optimal transformed data from training month according to Lemma 1. Figure 5 depicts a training samples (blue), test samples from the six subsequent months (green/yellow) and the computed equilibrium data (red) for three different training months (April - June 2007); a fourth is shown in Figure 1. If changes of spam mails are continuous they can be well predicted (see e.g. upper right corner). If instead the changes are more volatile (see e.g. lower right corner) the *Nash* model is not able to predict the appearance of new spam mails.

Approximation of adversary's responses

Finally, we study the impact of the degree t of the Taylor approximation on the accuracy and execution time of *Bayes*. Figure 6 (center) shows the RMSE evaluated over time for $t = 1, 2, 3$ (see Equation 11). Figure 6 (right, top) shows the execution time depending on the number of training emails for a fixed number of attributes ($m = 10$). Figure 6 (right, bottom) shows the execution time depending on the number of attributes for a fixed number of training mails ($n = 200$).

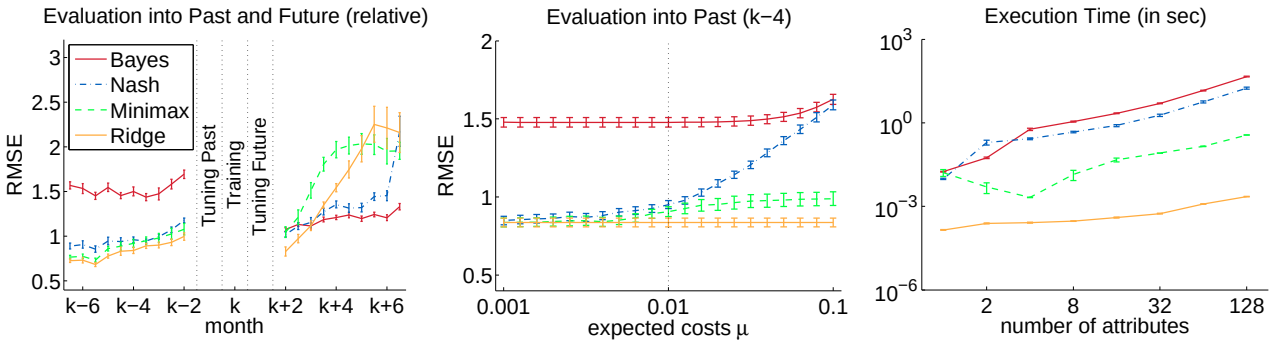

Figure 4. Evaluation of regression models with fixed expected costs into the past and future (left) and varying expected costs into the past (center). Execution Time (right). Error bars show standard errors.

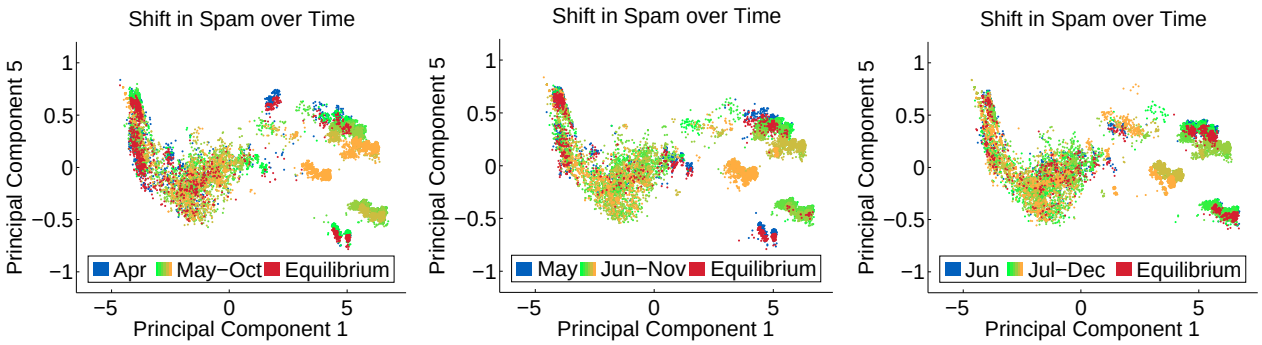

Figure 5. Shift in spam mails over time for three different training months. Training data is shown in blue; the imaginary Equilibrium data are shown in red.

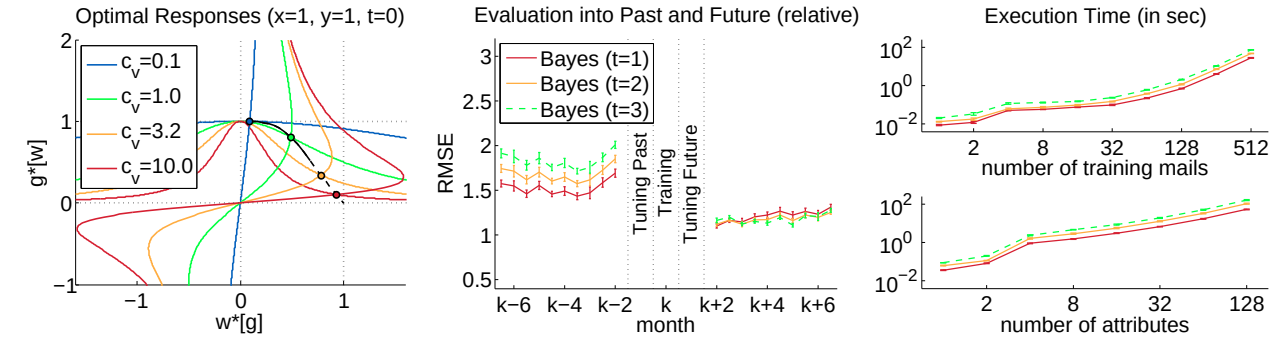

Figure 6. Optimal responses for one-dimensional regression games (left). Evaluation of Bayesian model with different Taylor degrees t over time (center). Execution Time (right).