2020 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 21–24, 2020, ESPOO, FINLAND

ON THE ADVERSARIAL ROBUSTNESS OF LINEAR REGRESSION

Fuwei Li[*]*, Lifeng Lai*[*]*, and Shuguang Cui*[†]

[*]Department of ECE, University of California, Davis
[†] Chinese University of Hong Kong, Shenzhen, China
Email: [*]{fli, lflai}@ucdavis.edu, [†]shuguangcui@cuhk.edu.cn

ABSTRACT

In this paper, we study the adversarial robustness of linear regression problems. Specifically, we investigate the robustness of the regression coefficients against adversarial data samples. In the considered model, there exists an adversary who is able to add one carefully designed adversarial data sample into the dataset. By leveraging this poisoned data sample, the adversary tries to boost or depress the magnitude of one targeted regression coefficient under the energy constraint of the adversarial data sample. We characterize the exact expression of the optimal adversarial data sample in terms of the targeted regression coefficient, the original dataset and the energy budget. Our experiments with synthetic and real datasets show the efficiency and optimality of our proposed adversarial strategy.

Index Terms— Adversarial robustness, linear regression, poisoning attack, non-convex optimization

1. INTRODUCTION

Linear regression is one of the fundamental machine learning algorithms being used in a wide range of applications [1]. In linear regression, one makes a simple assumption that there is a linear relationship between the response and the explanatory variables. Thus, in linear regression, our goal is to find out the linear regression coefficients given the data samples, which is usually accomplished by ordinary least squares (OLS) or ridge regression approaches. The regression coefficients will be used subsequently to perform prediction or forecasting given the explanatory variables. The regression coefficients also give us a way to explain the relationship between the response variable and the explanatory variables, which is very important in biologic science [2], financial analysis [3], and environmental science [4]. Generally speaking, a large magnitude of the regression coefficient indicates a significant relationship between its corresponding explanatory variable and the response variable. Furthermore, a small regression coefficient implies a redundant or irrelevant explanatory variable. This is especially true when certain coefficients regularized regression methods, such as ridge regression and Lasso [5], are used.

Machine learning is being widely used in security and safety sensitive applications, for example, medical image analysis [6] and autonomous driving [7]. Since these applications play a vital role in our lives, the robustness of machine learning algorithms in adversarial environments has received significant research interests [8, 9, 10]. An adversary may exist in these applications and be able to observe the whole data sets. After seeing all the data samples, the adversary can corrupt our learning model and mislead the learning result by modifying our training data or adding some adversarial data samples into the original data sets. It is important to understand the robustness of machine learning algorithms in these adversary environments before they can be safely employed in critical applications.

In this paper, we consider the adversarial robustness of the linear regression problem. In the considered model, there exists a powerful adversary who can observe the whole dataset. After that, the adversary will carefully design one special adversarial data sample and add it to the original data samples in order to manipulate one specific regression coefficient. In particular, we investigate how to design the optimal adversarial data sample to minimize or maximize the absolute value of a regression coefficient under a certain energy constraint. By minimizing the absolute value of a regression coefficient, the adversary intends to make a crucial variable appear to be unimportant. Similarly, by maximizing the amplitude, the adversarial can make an irrelevant variable appear to be important. By doing these, the adversary can manipulate our interpretation of the model and can impact the performance of the downstream applications that rely on the generated model. We show that the problem of finding the optimal adversarial example can be transformed into an optimization problem with the objective being the ratio of two quadratic functions and with a quadratic constraint. To solve this non-convex problem, we convert it to a quadratically constrained quadratic programming (QCQP) problem. Although our objective is still non-convex, we identify the optimal solution from its Karush-Kuhn-Tucker (KKT) necessary condi-

The work was partially supported by the National Science Foundation with Grants CNS-1824553, CCF-1717943, and CCF-1908258.

978-1-7281-6662-9/20/$31.00

tions and are able to give the closed-form expressions for the optimal adversarial data sample in terms of the targeted regression coefficient, the original data samples and the energy budget. Compared with the related work [11] that formulated this problem as a bi-level optimization problem, our method provides the global optimality, while the bi-level method is NP hard and thus no global optimality is guaranteed.

2. PROBLEM FORMULATION

Suppose we have n data samples, $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, where y_i is the response variable, $\mathbf{x}_i \in \mathbb{R}^m$ is the feature vector, and each component of $\mathbf{x}_i$ represents an explanatory variable. In this paper, we consider an adversarial setup in which the adversary first observes the the whole dataset $(\mathbf{y}, \mathbf{X})$, in which $\mathbf{y} := [y_1, y_2, \cdots, y_n]^\top$ and $\mathbf{X} := [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\top$, and then carefully designs an adversarial data sample, $(y_0, \mathbf{x}_0)$, and inserts it into the existing data samples. After inserting this adversarial data sample, we have the poisoned dataset $(\hat{\mathbf{y}}, \hat{\mathbf{X}})$, where $\hat{\mathbf{y}} := [y_0, y_1, y_2, \cdots, y_n]^\top$, $\hat{\mathbf{X}} := [\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\top$.

From the dataset, we intend to learn a linear regression model. From the poisoned dataset, the learned model is obtained by solving

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^m}{\text{argmin}} \quad \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2. \tag{1}$$

Let $\hat{\boldsymbol{\beta}}$ be the optimal solution of problem (1).

The goal of the attacker is to design the adversarial data sample $(y_0, \mathbf{x}_0)$ to decrease (or increase) the importance of a certain explanatory variable. If the goal is to decrease the importance of explanatory variable i, the objective function can be written as

$$\min_{\mathbf{x}_0, y_0} \ |\hat{\beta}_i|, \tag{2}$$

where $\hat{\beta}_i$ is the ith component of $\hat{\boldsymbol{\beta}}$. Similarly, if the goal of the attacker is to increase the importance of an explanatory variable, the objective function can be written as

$$\max_{\mathbf{x}_0, y_0} \ |\hat{\beta}_i|. \tag{3}$$

To make the problem nontrivial, we need to impose certain constraint on $(y_0, \mathbf{x}_0)$. In this paper, we impose energy constraint on the adversarial data sample and we use the ℓ_2 norm to measure the energy of the adversarial data sample. As the result, the optimization problem with respect to objective (2) can be written as

$$\begin{aligned} \min_{\|[\mathbf{x}_0^\top, y_0]\|_2 \le \eta} \quad & |\hat{\beta}_i| \\ \text{s.t.} \quad & \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2, \end{aligned} \tag{4}$$

where η is the energy budget. Similarly, we have the objective

$$\begin{aligned} \max_{\|[\mathbf{x}_0^\top, y_0]\|_2 \le \eta} \quad & |\hat{\beta}_i| \\ \text{s.t.} \quad & \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2, \end{aligned} \tag{5}$$

with respect to problem (3). Problem (4) and problem (5) are complicated bi-level optimization problems. In this paper, we will fully characterize their optimal solutions. These optimal solutions will enable us to understand the impact of the adversarial data sample on the linear regression problem.

3. OPTIMAL SOLUTIONS

To solve problem (4) or problem (5), we first solve the following two optimization problems

$$\min_{\|[\mathbf{x}_0^\top, y_0]\|_2 \le \eta} \quad \hat{\beta}_i \tag{6}$$

$$\text{s.t.} \quad \hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2. \tag{7}$$

and

$$\max_{\|[\mathbf{x}_0^\top, y_0]\|_2 \le \eta} \quad \hat{\beta}_i \tag{8}$$

$$\text{s.t.} \quad \hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\boldsymbol{\beta}\|_2^2. \tag{9}$$

It is easy to check that the solutions to problems (4) and (5) can be obtained from the solutions to (6) and (8). In particular, let $(\hat{\beta}_i^*)_{\min}$ and $(\hat{\beta}_i^*)_{\max}$ be optimal values of problems (6) and (8) respectively. Then, if $\hat{\beta}_i \ge 0$, we can check that $\max\{0, (\hat{\beta}_i^*)_{\min}\}$ and $\max\{|(\hat{\beta}_i^*)_{\min}|, |(\hat{\beta}_i^*)_{\max}|\}$ are the solutions to problems (4) and (5) respectively. Similar arguments can be made if $\hat{\beta}_i < 0$. Compared with (4) and (5), the objective functions in (6) and (8) also provide additional benefits. For example, we can use these formulations to study how to change a positive regression coefficient to a negative one.

In the following, we will focus on solving the minimization problem (6). The solution to the maximization problem (8) can be obtained using similar approach. To solve this bi-level optimization problem, we can first solve the optimization problem in the constraint (7). Problem (7) is just an ordinary least squares problem, which has a simple closed-form solution: $\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^\top \hat{\mathbf{y}}$. Plug in $\hat{\mathbf{X}} = [\mathbf{x}_0, \mathbf{X}^\top]^\top$ and $\hat{\mathbf{y}} = [y_0, \mathbf{y}^\top]^\top$, and we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X} + \mathbf{x}_0 \mathbf{x}_0^\top)^{-1} [\mathbf{x}_0, \mathbf{X}^\top][y_0, \mathbf{y}^\top]^\top.$$

According to the Sherman-Morrison formula [12], we have

$$\begin{aligned} &(\mathbf{X}^\top \mathbf{X} + \mathbf{x}_0 \mathbf{x}_0^\top)^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} - \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0 \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 + \mathbf{x}_0^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_0}. \end{aligned}$$

The inverse of $\mathbf{X}^\top\mathbf{X} + \mathbf{x}_0\mathbf{x}_0^\top$ always exists because $1 + \mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0 \neq 0$. Plug this inverse in the expression of $\hat{\boldsymbol{\beta}}$, we get

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}_0 + y_0\mathbf{A}\mathbf{x}_0 - \frac{\mathbf{A}\mathbf{x}_0\mathbf{x}_0\mathbf{A}}{1+\mathbf{x}_0^\top\mathbf{A}\mathbf{x}_0}(\mathbf{X}^\top\mathbf{y} + y_0\mathbf{x}_0) \\ &= \boldsymbol{\beta}_0 + \frac{\mathbf{A}\mathbf{x}_0(y_0 - \mathbf{x}_0^\top\boldsymbol{\beta}_0)}{1+\mathbf{x}_0^\top\mathbf{A}\mathbf{x}_0},\end{aligned}$$

where

$$\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1}, \tag{10}$$

$$\boldsymbol{\beta}_0 = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{11}$$

As $\boldsymbol{\beta}_0$ is independent of $(y_0, \mathbf{x}_0)$, problem (6) is equivalent to

$$\min_{\mathbf{x}_0, y_0} \quad \frac{\mathbf{a}^\top\mathbf{x}_0(y_0 - \mathbf{x}_0^\top\boldsymbol{\beta}_0)}{1+\mathbf{x}_0^\top\mathbf{A}\mathbf{x}_0} \tag{12}$$

$$\text{s.t.} \quad \|[\mathbf{x}_0^\top,\, y_0]\|_2 \leq \eta, \tag{13}$$

where $\mathbf{a}$ is the ith column of $\mathbf{A}$. The optimization problem (12) is the ratio of two quadratic functions with a quadratic constraint. To further simplify this optimization problem, we can write our objective and constraints in a more compact form by performing variable change: $\mathbf{u} = [\mathbf{x}_0^\top,\, y_0]^\top$. Using this compact representation, the optimization problem (12) can be written as

$$\min_{\mathbf{u}} \quad \frac{\frac{1}{2}\mathbf{u}^\top\mathbf{H}\mathbf{u}}{1+\mathbf{u}^\top\left[\begin{smallmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & 0\end{smallmatrix}\right]\mathbf{u}} \tag{14}$$

$$\text{s.t.} \quad \mathbf{u}^\top\mathbf{u} \leq \eta^2,$$

in which

$$\mathbf{H} = \begin{bmatrix} -\mathbf{a}\boldsymbol{\beta}_0^\top - \boldsymbol{\beta}_0\mathbf{a}^\top & \mathbf{a} \\ \mathbf{a}^\top & 0 \end{bmatrix}. \tag{15}$$

(14) is a non-convex optimization problem. To solve this problem, we employ the technique introduced in [13]. We first do variable change $\mathbf{u} = \frac{\mathbf{z}}{s}$ by introducing variable $\mathbf{z}$ and scalar s. Inserting this into problem (14), adding constraint 1 to the denominator of the objective and moving it to the constraint, we have a new optimization problem

$$\min_{\mathbf{z}, s} \quad \frac{1}{2}\mathbf{z}^\top\mathbf{H}\mathbf{z} \tag{16}$$

$$\text{s.t.} \quad s^2 + \mathbf{z}^\top\left[\begin{smallmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & 0\end{smallmatrix}\right]\mathbf{z} = 1, \tag{17}$$

$$\mathbf{z}^\top\mathbf{z} \leq s^2\eta^2. \tag{18}$$

To validate the equivalence between problem (14) and (16), we only need to check if the optimal value of problem (14) is less than the optimal value of problem (16) when $s = 0$ [13]. Firstly, since $\mathbf{H}$ is not positive semi-definite (which will be shown later), the optimal value of problem (14) is less than zero. Secondly, when $s = 0$, the optimal value of problem (16) is zero, which is apparently larger than the optimal value of problem (14). Therefore, the two problems are equivalent.

To solve problem (16), we substitute s^2 in equation (17) for that in equation (18) and then we have

$$\min_{\mathbf{z}} \quad \frac{1}{2}\mathbf{z}^\top\mathbf{H}\mathbf{z} \tag{19}$$

$$\text{s.t.} \quad \mathbf{z}^\top\left(\mathbf{I} + \eta^2\left[\begin{smallmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & 0\end{smallmatrix}\right]\right)\mathbf{z} \leq \eta^2. \tag{20}$$

Note that $\mathbf{H}$ is not positive semi-definite, hence problem (19) is not a standard convex QCQP problem. However, it is proved that strong duality holds for this problem [14]. So, to solve this problem, we can start by investigating its KKT necessary conditions. The Lagrangian of problem (19) is

$$\mathcal{L}(\mathbf{z}, \lambda) = \frac{1}{2}\mathbf{z}^\top\mathbf{H}\mathbf{z} + \lambda\big(\mathbf{z}^\top(\mathbf{I} + \eta^2\left[\begin{smallmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & 0\end{smallmatrix}\right])\mathbf{z} - \eta^2\big),$$

where λ is the dual variable. According to the KKT conditions, we have

$$(\mathbf{H} + \lambda\mathbf{D})\,\mathbf{z} = \mathbf{0}, \tag{21}$$

$$\frac{1}{2}\mathbf{z}^\top\mathbf{D}\mathbf{z} \leq \eta^2, \tag{22}$$

$$\lambda\left(\frac{1}{2}\mathbf{z}^\top\mathbf{D}\mathbf{z} - \eta^2\right) = 0, \tag{23}$$

$$\lambda \geq 0, \tag{24}$$

where

$$\mathbf{D} = 2\left(\mathbf{I} + \eta^2\begin{bmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}\right). \tag{25}$$

By inspecting the complementary slackness condition (23), we consider two cases based on the value of λ.

Case 1: $\lambda = 0$. In this case, we must have $\mathbf{H}\mathbf{z} = \mathbf{0}$. As a result, the objective value of (19) is zero, which contradicts with the fact the optimal value should be negative. Hence, this case is not possible.

Case 2: $\lambda > 0$. In this case, equality in (22) must hold. According to the stationary condition (21), if the matrix $\mathbf{H} + \lambda\mathbf{D}$ is full rank, we must have $\mathbf{z} = \mathbf{0}$, for which equality in (22) cannot hold. Hence, $\mathbf{H} + \lambda\mathbf{D}$ is not full-rank and we have $\det(\mathbf{H} + \lambda\mathbf{D}) = 0$. As $\mathbf{D}$ is positive definite, we also have

$$\det(\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2} + \lambda\mathbf{I}) = 0. \tag{26}$$

Since $\lambda > 0$, this equality tells us that $-\lambda$ belongs to one of the negative eigenvalues of $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$. In the following, we will show that $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$ has one and only one negative eigenvalue.

By definition, $\mathbf{D}$ is a block diagonal matrix. Hence, its inverse is also block diagonal. Let us define $\mathbf{D}^{-1/2} =$

$\text{diag}\{\mathbf{G}, g\}$, where $\mathbf{G} = 1/\sqrt{2}(\mathbf{I}+\eta^2\mathbf{A})^{-1/2}$ and $g = 1/\sqrt{2}$. Thus, we have

$$\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2} = \begin{bmatrix} -\mathbf{c}\mathbf{h}^\top - \mathbf{h}\mathbf{c}^\top & g\mathbf{c} \\ g\mathbf{c}^\top & 0 \end{bmatrix},$$

where $\mathbf{c} = \mathbf{G}\mathbf{a}$ and $\mathbf{h} = \mathbf{G}\boldsymbol{\beta}_0$. Define ξ as the eigenvalue of $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$, and compute its eigenvalues by computing the characteristic polynomial:

$$\begin{aligned}&\det\left(\xi\mathbf{I} - \mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}\right) \\ &= \xi^{m-1}\left(\xi^2 + 2\xi\mathbf{c}^\top\mathbf{h} + \mathbf{c}^\top\mathbf{h}\mathbf{h}^\top\mathbf{c} - g^2\mathbf{c}^\top\mathbf{c} - \mathbf{c}^\top\mathbf{c}\mathbf{h}^\top\mathbf{h}\right).\end{aligned}$$

Thus, the eigenvalues of $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$ are $\xi = 0$ ((m-1) multiplications) and $\xi = -\mathbf{c}^\top\mathbf{h} \pm \|\mathbf{c}\|_2\sqrt{g^2 + \mathbf{h}^\top\mathbf{h}}$. Since $\|\mathbf{c}\|_2\sqrt{g^2 + \mathbf{h}^\top\mathbf{h}} > |\mathbf{c}^\top\mathbf{h}|$, the eigenvalues of $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$ satisfy

$$\xi_{m+1} < 0, \quad \xi_m = \xi_{m-1} = \cdots = \xi_2 = 0, \quad \xi_1 > 0.$$

Now, it is clear that $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$ has one and only one negative eigenvalue and one positive eigenvalue respectively. Thus, we have $\lambda = -\xi_{m+1}$. Assume $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_{m+1}$ are two eigenvectors corresponding to eigenvalues ξ_1 and ξ_{m+1}. Through simple calculation, we have

$$\boldsymbol{\nu}_i = k_i[-\frac{\mathbf{c}^\top\mathbf{h} + \xi_i}{\mathbf{c}^\top\mathbf{c}}\mathbf{c}^\top + \mathbf{h}^\top, \frac{g\mathbf{c}^\top}{\xi_i}(-\frac{\mathbf{c}^\top\mathbf{h} + \xi_i}{\mathbf{c}^\top\mathbf{c}}\mathbf{c} + \mathbf{h})]^\top, \tag{27}$$

where $i = 1, m+1$ and scalar k_i is the normalization constant to guarantee the eigenvectors to be of unit length. According to (21), we have

$$(\mathbf{H} + \lambda\mathbf{D})\,\mathbf{z} = \mathbf{D}^{1/2}\left(\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2} + \lambda\mathbf{I}\right)\mathbf{D}^{1/2}\mathbf{z} = \mathbf{0};$$

thus the solution to problem (19) is

$$\mathbf{z}^* = k \cdot \mathbf{D}^{-1/2}\boldsymbol{\nu}_{m+1}. \tag{28}$$

Since $\frac{1}{2}\mathbf{z}^\top\mathbf{D}\mathbf{z} = \eta^2$, we have $k = \sqrt{2}\eta$. Having the expression of the optimal $\mathbf{z}^*$, we can then compute s according to equation (17):

$$s = \pm\sqrt{1 - (\mathbf{z}^*_{1:m})^\top\mathbf{A}\,\mathbf{z}^*_{1:m}}, \tag{29}$$

where $\mathbf{z}^*_{1:m}$ is the vector that comprises the first m elements of $\mathbf{z}^*$. Hence, the corresponding solution to problem (12) is

$$\mathbf{x}_0^* = \mathbf{z}^*_{1:m}/s, \quad y_0^* = \mathbf{z}^*_{m+1}/s. \tag{30}$$

We now compute the optimal value of problem (16). Since our objective function is $\frac{1}{2}(\mathbf{z}^*)^\top\mathbf{H}\mathbf{z}^*$, substituting $\mathbf{z}^*$ in (28) leads to the objective value:

$$\eta^2\boldsymbol{\nu}_{m+1}^\top\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}\boldsymbol{\nu}_{m+1}.$$

Algorithm 1 Optimal Adversarial Data Point Design

1: **Input**: the data set, $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, energy budget η, and the index of feature to be attacked.
2: **Steps**:
3: compute $\mathbf{A}$ according to equation (10), compute $\boldsymbol{\beta}_0$ according to (11).
4: compute $\mathbf{H}$ and $\mathbf{D}$ according to (15) and (25), respectively.
5: compute the last eigenvalue, ξ_{m+1}, of $\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}$ and its corresponding eigenvector according to (27).
6: design the adversarial data point, $(y_0, \mathbf{x}_0)$, according to equations (28), (29), and (30).
7: **Output**: return the optimal adversarial data point $(y_0, \mathbf{x}_0)$ and the optimal objective value $\eta^2\xi_{m+1} + (\boldsymbol{\beta}_0)_i$.

Since $\boldsymbol{\nu}_{m+1}^\top\mathbf{D}^{-1/2}\mathbf{H}\mathbf{D}^{-1/2}\boldsymbol{\nu}_{m+1} = \xi_{m+1}$, our optimal objective value is $\eta^2\xi_{m+1}$.

Following similar analysis as above, we can find the optimal $\mathbf{z}^*$ for problem (8), which is $\mathbf{z}^* = \sqrt{2}\eta\mathbf{D}^{-1/2}\boldsymbol{\nu}_1$. Also, we can compute the optimal $\mathbf{x}_0^*$ and y_0^* according to equation (30) and its optimal objective value, which is $\eta^2\xi_1$.

In summary, the optimal values for problems (6) and (8) are $\eta^2\xi_{m+1} + (\beta_0)_i$ and $\eta^2\xi_1 + (\beta_0)_i$ respectively. We have summarized the process to design the optimal adversarial data point in Algorithm 1 with respect to objective (6) and the process with respect to objective (8) can be obtained accordingly. Based on our optimal values of problems (6) and (8), we can further decide the optimal values of problems (4) and (5) as discussed at the beginning of this section.

Moreover, if we use the ridge regression method in linear regression, there is only a slight difference in the matrix $\mathbf{A}$ in problem (12) and the whole analysis remains the same.

4. NUMERICAL EXAMPLES

In this section, we will use numerical examples to demonstrate the results we obtained in this paper.

In the first experiment, we test our algorithms on a synthetic data set. In this experiment, we first generate a 20×10 feature matrix $\mathbf{X}$, where each element of $\mathbf{X}$ is i.i.d. generated according to a standard normal distribution. Then, we generate our response values, $\mathbf{y}$, according to the observation model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{n}$, where each element of $\boldsymbol{\beta}$ is i.i.d. generate according to a standard normal distribution and each entry of $\mathbf{n}$ is i.i.d. generated according to a Gaussian distribution with zero mean and 0.1 variance. On this synthesized data sample, $(\mathbf{y}, \mathbf{X})$, we perform our attacks with two algorithms. The first algorithm is the one described in Algorithm 1. The second algorithm is a random attack strategy. In the random attack strategy, given the energy budget, we randomly generate an adversarial data sample, $(y_0, \mathbf{x}_0)$, with each of its entries being i.i.d. generated according to the standard normal distribution, and normalize its energy to the given energy bud-

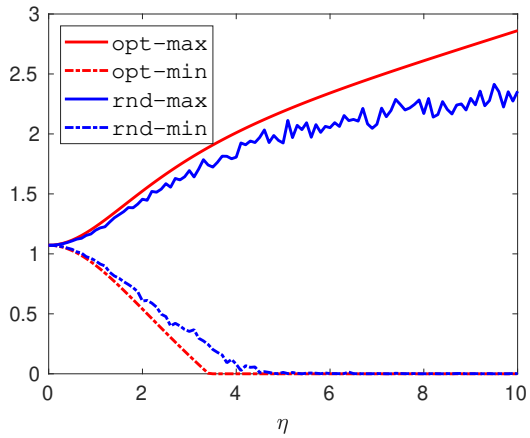

Fig. 1. The value of the fourth regression coefficient after our proposed attacks and after random attacks with different energy budgets.

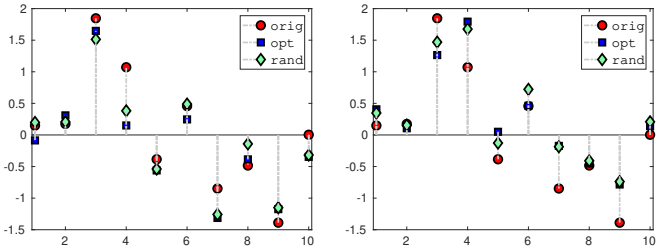

Fig. 2. The regression coefficients before and after our proposed and random attacks on our synthesized data set. The left figure shows the regression coefficients before and after attacking the fourth regression coefficient with objective (4) and the right one shows the regression coefficients before and after attacking the fourth regression coefficient with objective (5)

get. Then, we add this adversarial data point into the original data set and compute the linear regression coefficients. We repeat this process 10000 times and among which we record the data points which lead to the minimal and maximal objective values of (4) and (5), respectively.

Fig. 1 depicts the value of the fourth regression coefficient after our proposed attacks and after the random attacks with different energy budgets. In this figure, the x-axis indicates the energy budget and the y-axis indicates the value of the fourth regression coefficient. 'opt-max', 'opt-min', 'rnd-max, 'rnd-min' denote our proposed strategy with objective (5), our proposed strategy with objective (4), random attacks with objective (5), random attacks with objective (4), respectively. Fig. 2 shows the regression coefficients before and after we attack the fourth regression coefficient with energy budget $\eta = 3$, where 'orig', 'opt', 'rand' denote the original regression coefficients, the regression coefficient after our proposed

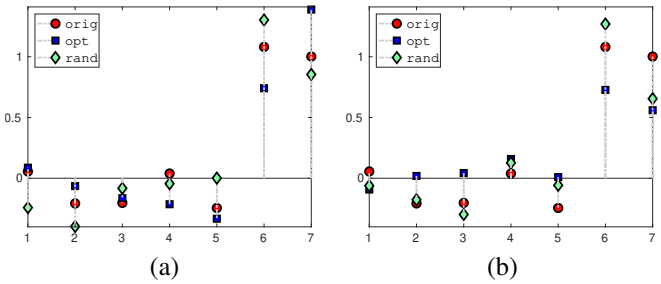

Fig. 3. The regression coefficients under our proposed and random attacks on the Istanbul Stock Exchange data set. The left figure shows the regression coefficients before and after attacking the fourth coefficient with objective (6) and the right one shows the regression coefficients before and after attacking the fifth regression coefficient with objective (4).

attack and the regression coefficients after random attacks, respectively. Fig. 2 (a) shows the regression coefficient with the objective that minimizes the fourth regression coefficient and the right one shows the regression coefficients with the objective that maximizes the magnitude of the fourth regression coefficient. As the two figures demonstrated, our proposed strategy is much more efficient than the random attack strategy. Since the random attach strategy can be seen as an exhaustive search algorithm, it further proves the optimality of our proposed algorithm.

In the second experiment, we test our adversarial attack strategy on a real data set. In this regression task, we use seven international indexes to predict the returns of the Istanbul Stock Exchange [15]. The data set contains 536 data samples, which are the records of the returns of Istanbul Stock Exchange with seven other international indexes starting from Jun. 5, 2009 to Feb. 22, 2011.

We use OLS regression for this task and get the regression coefficients as shown in Fig. 3. In the figure, the x-axis denotes the index of the regression coefficients and the y-axis indicates the value of the regression coefficients. We design our first experiment to attack the fourth regression coefficient and try to make it small by solving problem (6). We use two strategies to attack this coefficient with fixed energy budget by setting $\eta = 0.2$. The first strategy is the one proposed in this paper. As a comparison, we also use a random strategy. In the random one, we randomly generate the adversarial data sample with each entry being i.i.d. generated from a standard normal distribution. Then, we normalize its energy to be η. We repeat this random attack 10000 times and select the one with the smallest value of the fourth regression coefficient.

In the second experiment on this real data set, we intend to make the absolute value of the fifth regression coefficient small. We compare the proposed and random attack strategies to attack the fifth coefficient with fixed energy budget $\eta = 0.1$. Similarly, for the random attacks strategy, we run

10000 times random attacks and select the one with the smallest absolute value of the fifth regression coefficient.

Fig. 3(a) shows the regression coefficients before and after the first experiment and Fig. 3(b) shows the regression coefficients before and after the second experiment. From the figures we can see that our proposed adversarial attack strategy is much more efficient than the random attack strategy. One can also observe that by only adding one adversarial data point, designed using the approach characterized in this paper, one can dramatically change the value of a regression coefficient and hence change the importance of that feature.

5. CONCLUSION

In this paper, we have investigated the adversarial robustness of the linear regression problem. We have characterized the optimal adversarial data sample under the energy constraint. Our closed-form results provide a clear view of the relationship among the adversarial data sample, the original data samples, and the energy budget. It further provides insights into the robustness of linear regression against the adversarial data sample. In the future, it is of interest to consider how to design more than one adversarial data samples and how to defend against such attacks.

6. REFERENCES

[1] X. Yan and X. Su, *Linear regression analysis: theory and computing*. World Scientific, 2009.

[2] J. H. McDonald, *Handbook of biological statistics*. Sparky House Publishing, 2009.

[3] O. E. Barndorff-Nielsen and N. Shephard, "Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics," *Econometrica*, vol. 72, no. 3, pp. 885–925, May 2004.

[4] C. J. ter Braak and S. Juggins, "Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages," in *Proc. International Diatom Symposium*, Renesse, The Netherlands, Aug. 1993, pp. 485–502.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[6] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019.

[7] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *Electronic Imaging*, vol. 2017, no. 19, pp. 70–76, Jan. 2017.

[8] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, Dec. 2014.

[9] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv:1611.01236*, Nov. 2016.

[10] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, Jun. 2018.

[11] S. Mei and X. Zhu, "Using machine teaching to identify optimal training-set attacks on machine learners," in *Proc. AAAI Conference on Artificial Intelligence*, Austin Texas, Jan. 2015, pp. 2871–2877.

[12] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, 2012.

[13] A. Beck and M. Teboulle, "On minimizing quadratically constrained ratio of two quadratic functions," *Journal of Convex Analysis*, vol. 17, no. 3, pp. 789–804, 2010.

[14] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[15] O. Akbilgic, H. Bozdogan, and M. E. Balaban, "A novel hybrid RBF neural networks model as a forecaster," *Statistics and Computing*, vol. 24, no. 3, pp. 365–375, May 2014.