Theoretical Bounds on the Adversarial Robustness of Reduced-Rank Regression

Soyon Choi and Scott Alfeld

Amherst College, Amherst MA 01002, USA
{sochoi25, salfeld}@amherst.edu

Abstract. We consider multi-target linear regression in the presence of a deployment-time attacker. Specifically, we examine the effects of incorporating a maximum-rank constraint on the learned weight matrix (i.e., performing reduced-rank regression). For a broad class of practically relevant defender and attacker settings, we derive theoretical bounds on the change in adversarial robustness as a function of the rank constraint. In the classical setting—where the learner minimizes Mean Squared Error and the attacker is constrained by an ℓ_2 constraint—we show that adversarial robustness is unaffected by rank reduction. In contrast, under more general and practical settings, rank constraints can dramatically alter robustness. In general, these bounds depend on the eigenvalues of matrix $\mathbf{W}$, which defines defender loss. These bounds and accompanying analysis provide both practical value and further develop a foundational understanding of the robustness of linear methods.

Keywords: Reduced-Rank Regression · Adversarial Learning · Linear Methods.

1 Introduction

Multi-target regression, which models multiple response variables (i.e., targets) at once, is integral to systems in a range of fields including finance, healthcare, and engineering. These machine learning systems may be vulnerable to adversaries that attempt to influence outcomes by injecting carefully crafted inputs. Hence, it is crucial to understand and ensure the robustness of these systems against such adversaries.

Many machine learning systems use linear methods to model their data. When jointly predicting multiple targets, a common method is to assume that the underlying coefficient matrix is of full-rank. A model is learned for each target independently. That is, no additional knowledge is gained by formulating the problem as a multi-variate regression task and learning the coefficients jointly.

To properly address the interconnected nature of a multi-target dataset, practitioners often use specialized multi-target regression methods. These methods are based on the key idea that targets must be tied together during learning. One common strategy is to force the learner to learn an internal representation of the data that implies a dependency between targets by limiting the rank of the

Table 1: Mathematical notation.

Variable	Description	Domain
n	Number of data samples	$\mathbb{N}$
d	Number of input features	$\mathbb{N}$
q	Number of targets	$\mathbb{N}$
r	Rank constraint of regressor	$\mathbb{N}$
c	Upper bound on attacker perturbation	$\mathbb{R}^+$
$\mathbf{x}$	An input vector	$\mathbb{R}^d$
$\mathbf{y}$	An output vector	$\mathbb{R}^q$
$\boldsymbol{\alpha}$	Attack vector to be added to $\mathbf{x}$	$\mathbb{R}^d$
$\mathbf{X}$	Input data matrix	$\mathbb{R}^{n\times d}$
$\mathbf{Y}$	Output data matrix	$\mathbb{R}^{n\times q}$
$\mathbf{M}$	Model coefficient matrix	$\mathbb{R}^{q\times d}$
$\mathbf{W}$	Defining matrix of defender loss	$\{A \in \mathbb{R}^{q\times q} \mid A \succ 0\}$
$\mathbf{C}$	Defining matrix of attacker constraint	$\{A \in \mathbb{R}^{d\times d} \mid A \succ 0\}$
$\mathbf{V}$	Square root of $\mathbf{W}$ (i.e., $\mathbf{W} = \mathbf{V}^\top\mathbf{V}$)	$\{A \in \mathbb{R}^{q\times q} \mid A \succ 0\}$
$\mathbf{G}$	Square root of $\mathbf{C}$ (i.e., $\mathbf{C} = \mathbf{G}^\top\mathbf{G}$)	$\{A \in \mathbb{R}^{d\times d} \mid A \succ 0\}$
$f(\cdot)$	Prediction function	$\mathbb{R} \to \mathbb{R}$

parameter matrix (Reduced-Rank Regression (RRR)) [11], applying a shrinking matrix [19], or adding a special regularization term [18].

Adversarial attacks on single-target linear learners are well-studied in various contexts [1,14]. However, it remains unclear how the addition of *multi-targetness* affects the robustness of linear learners.

The primary contribution of this paper is a series of seven theoretical bounds on the change in model vulnerability to deployment-time attacks when applying a rank constraint in multi-target linear regression (i.e., learning a model using RRR). The first bound applies to what we refer to as the *classical setting*—the scenario where the defender minimizes Means Squared Error (MSE) and the attacker is bound by an ℓ_2 constraint. The rest consider a series of increasingly general settings.

Throughout this manuscirpt we use lower-case letters to denote scalars, bold lower-case letters to denote vectors, and bold upper-case letters to denote matrices. The specific mathematical notation is presented in Table 1.

We highlight the value of these contributions for three overlapping but different communities. For the practitioner, the bounds we present aid in decision making while performing RRR in adversarial settings. For the researcher studying linear predictors, we strengthen the community's understanding of the tie between model robustness and the spectra of the involved matrices. For the broader adversarial learning community, we highlight a weakness of the all-too-common practice of exclusively studying the classical setting. This setting is mathematically elegant and as such, it is the primary focus in much of the adversarial learning literature. However, as we show, the more realistic setting—where the defender minimizes a Mahalanobis loss and the attacker is constrained by an

ellipsoid rather than a sphere—demonstrates behavior that is obfuscated by the simplified special case. Specifically, we demonstrate that applying a rank constraint in multi-target linear regression has *no effect* on adversarial robustness in the classical setting, but can *dramatically change* the robustness in more general and realistic settings. We provide a series of theoretical bounds that range from the classical setting to the most general case where the defender's loss and attacker constraints are defined by Mahalanobis metrics. This serves to develop a foundational understanding of how adversarial robustness is affected by design decisions for more realistic attackers and defenders.

Throughout this manuscript we use the following running example for clarity. While our example is based on economic forecasting, note that our results are not specific to any particular application (economic or otherwise). We use this example only to facilitate understanding.

Running Example. Consider a defender aiming to predict next quarter's profits for target companies T_1, T_2, and T_3. Companies T_1 and T_2 are in the food service industry while T_3 is in retail. To make their predictions, the defender monitors regional shopping trends and use the number of sales of various items as input features. An attacker has limited capability to affect the prices of various raw materials, and seeks to use that capability so as to make the defenders predictions as wrong as possible.

2 Background on Reduced-Rank Regression

Reduced-rank regression (RRR) is a linear regression technique that explicitly ties together targets during learning via a *rank constraint* on the coefficient matrix. Consider the learned prediction function f of a general linear model:

$$f(\mathbf{x}) = \mathbf{M}\mathbf{x} \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{q\times d}$ is the model coefficient matrix and $\mathbf{x} \in \mathbb{R}^{q\times 1}$ is the input vector.

The learner aims to learn the model matrix $\mathbf{M}$ by minimizing the mean loss on the training set $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the matrix of inputs and $\mathbf{Y} \in \mathbb{R}^{n\times q}$ is the matrix of targets. Each row $\mathbf{x}_i$ for $i \in \{1, ..., n\}$ of $\mathbf{X}$ and corresponding row $\mathbf{y}_i$ of $\mathbf{Y}$ forms one training sample $(\mathbf{x}_i, \mathbf{y}_i)$. Let the loss be a generalization of mean squared error (MSE)[1] — the squared Mahalanobis norm. The Mahalanobis norm of a vector $\mathbf{a} \in \mathbb{R}^q$ is defined as follows:

$$||\mathbf{a}||^2_{\mathbf{W}} = \mathbf{a}^\top \mathbf{W}\mathbf{a} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{q\times q}$ is a positive definite matrix.

In ordinary least squares (OLS) estimation, the learner minimizes MSE. We note that OLS is the most commonly studied linear method. However, real-world applications often require models to account for factors beyond uniform

[1] Mean squared error is a special case of Mahalanobis norm, where $\mathbf{W} = \mathcal{I}_q$.

error, such as the relative importance or interaction of prediction targets. These factors are not captured by standard MSE. Hence, we consider a learner that minimizes the Mahalanobis norm, which allows for formalization of robustness as a function of $\mathbf{W}$ (see Section 4). Generalizing to this broader class of learners not only better reflects practical deployment but also provides deeper insight into the adversarial robustness of multi-target regression models.

Full-rank regression (i.e., joint OLS estimation) aims to solve the following optimization problem:

$$\mathbf{M}_{OLS} = \underset{\mathbf{M}\in\mathbb{R}^{q\times d}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} ||f(\mathbf{x}_i) - \mathbf{y}_i||^2_{\mathbf{W}} \tag{3}$$

In RRR, the rank of the coefficient matrix is constrained to some $1 \leq r < q$ during optimization. Hence, the constrained optimization problem of the RRR learner is as follows:

$$\mathbf{M}_r = \underset{\mathbf{M}\in\mathcal{M}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} ||f(\mathbf{x}_i) - \mathbf{y}_i||^2_{\mathbf{W}} \tag{4}$$

$$\mathcal{M} \stackrel{def}{=} \{\mathbf{M} : \mathbf{M} \in \mathbb{R}^{q\times d} \; s.t. \; rank(\mathbf{M}) \leq r\} \tag{5}$$

This rank constraint forces the learner to relate the target dimensions together. In comparison to full-rank regression, RRR resists overfitting to training data. That is, given a noisy dataset, learning a model under a rank constraint often improves the test accuracy of the model.

Running Example. In our running example, the profits of T_1, T_2, and T_3 are not independent. External latent features, such as weather and social trends affect T_1 and T_2 (the companies in food service) together. In other words, the environment is best modeled as one of deficient rank. As such, performing RRR will likely yield a better generalization error than independent OLS.

In this paper, we formalize the effect of constraining the rank of the coefficient matrix on the adversarial robustness of the learned model. We define the adversarial robustness of the learner under the threat model presented in the following section.

3 Threat Model

In this work, we focus on a deployment-time attacker who perturbs inputs at test time, rather than manipulating the training data. The defender learns a model $\mathbf{M}$. Consider an adversary that observes a point $\mathbf{x}$ and perturbs it by adding a vector $\boldsymbol{\alpha}$. The adversary's goal is to maximize the defender's loss on the perturbed point $\mathbf{x} + \boldsymbol{\alpha}$. This constitutes a *repulsive attack*, in which the adversary pushes the model's prediction away from the true target value. We model this interaction in a zero-sum setting, where the adversary's gain is equal to the defender's loss.

The adversary knows the model $\mathbf{M}$ but is constrained by how much they can alter the input. That is, they must pick a perturbation vector $\boldsymbol{\alpha}$ such that the Mahalanobis norm of $\boldsymbol{\alpha}$ is less than a scalar constraint c.

$$||\boldsymbol{\alpha}||_{\mathbf{C}} \leq c \tag{6}$$

where $\mathbf{C} \in \mathbb{R}^{d \times d}$ represents the geometry of the attacker constraint. If $\mathbf{C} = \mathcal{I}_d$, the attacker is bound by the ℓ_2 norm of its perturbation vector. The ℓ_2-bounded attacker is a commonly studied adversarial setting. However, in many real-world applications, an attacker may have unequal control over different input features. In such cases, $\mathbf{C}$ may deviate from the identity, reflecting a non-uniform constraint on perturbations in each feature. Hence, it is useful to consider the general attacker with an arbitrary positive definite matrix $\mathbf{C}$.
The attacker aims to solve the following constrained optimization problem:

$$\boldsymbol{\alpha}^{rep} = \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}} ||f(\mathbf{x}) - f(\mathbf{x} + \boldsymbol{\alpha})||^2_{\mathbf{W}} \tag{7}$$

$$\mathcal{A} \stackrel{def}{=} \{\boldsymbol{\alpha} : ||\boldsymbol{\alpha}||_{\mathbf{C}} \leq c\} \tag{8}$$

Then, by linearity of f, the attacker's optimization problem is:

$$\boldsymbol{\alpha}^{rep} = \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}} ||\mathbf{M}\mathbf{x} - \mathbf{M}(\mathbf{x} + \boldsymbol{\alpha})||^2_{\mathbf{W}} \tag{9}$$

$$= \arg\max_{\boldsymbol{\alpha} \in \mathcal{A}} ||\mathbf{M}\boldsymbol{\alpha}||^2_{\mathbf{W}} \tag{10}$$

That is, *the optimal repulsive attack is independent of the test instance*. This allows us to efficiently compute the exact solution for an optimal repulsive attack on linear regressors [2] and mathematically assess a model's robustness based solely on its parameters (without requiring access to any input data).

We assume that $\mathbf{C}$ and $\mathbf{W}$ are positive definite matrices. Define and let:

$$\mathbf{C} = \mathbf{G}^\top \mathbf{G} \tag{11}$$

$$\mathbf{W} = \mathbf{V}^\top \mathbf{V} \tag{12}$$

By Theorem 2 from Alfeld et al. [2], we have:

$$\boldsymbol{\alpha}^{rep} = c\mathbf{G}^{-1} s_1 \tag{13}$$

where s_1 is the right singular vector corresponding to the largest singular value of $\mathbf{V}\mathbf{M}\mathbf{G}^{-1}$ and the corresponding induced defender loss is the spectral norm of $\mathbf{V}\mathbf{M}\mathbf{G}^{-1}$:

$$||\mathbf{M}\boldsymbol{\alpha}^{rep}||^2_{\mathbf{W}} = ||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2 \tag{14}$$

That is, for any linear regressor with prediction function f, the optimal repulsive attack can be efficiently computed as a function of $\mathbf{V}$, $\mathbf{M}$, and $\mathbf{G}$, where $\mathbf{V}$ captures the loss of the defender, $\mathbf{M}$ is the model matrix such that $f(\mathbf{x}) = \mathbf{M}\mathbf{x}$, and $\mathbf{G}$ captures the constraints of the attacker.

Under this threat model, the attacker aims to **maximize the increase in the defender's loss due to the perturbation**. We note that this can be thought of as a "worst-case attacker", as it causes the maximum change in defender loss that any attractive attacker (i.e., an attacker that aims to change the output to their preferred value) could cause. Therefore, we denote **"attacker happiness"** as the induced defender loss (the spectral norm of $\mathbf{VMG}^{-1}$). To evaluate the adversarial robustness of RRR, we investigate the change in attacker happiness as an effect of constraining the rank of the coefficient matrix to be $\leq r$. In particular, we present upper and lower bounds on how much the attacker happiness can differ between full-rank and reduced-rank regression, given r.

4 Bounds on Robustness

In this section, we present various upper and lower bounds on how much the induced defender loss, or "attacker happiness", is affected by constraining the rank of the defender's weight matrix.

A defender employing a RRR learner must choose the appropriate r for its task. It is useful to understand how its choice of r may affect the vulnerability of the learned model.

We first consider the case which we call the *classical setting*: the defender learns models by minimizing MSE, and the attacker is bounded by spherical constraints. In this case, $\mathbf{W} = \mathcal{I}_q$ and $\mathbf{C} = \mathcal{I}_d$.

4.1 Case 1. The classical setting

Theorem 1. *If $\mathbf{W} = \mathcal{I}_q$ and $\mathbf{C} = \mathcal{I}_d$, constraining the rank of the coefficient matrix to r* ***does not affect*** *attacker happiness.*

Proof. Consider an unconstrained linear learner that learns the model coefficient matrix $\mathbf{M}$. Without loss of generality, we assume that $\min(d, q) = q$. Then, $rank(\mathbf{M}) \leq q$.

The **spectral norm** of $\mathbf{A} \in \mathbb{R}^{a \times b}$ is defined as:

$$||\mathbf{A}||_2 = \sigma_{\max}(\mathbf{A}) \tag{15}$$

where $\sigma_{\max}$ denotes the maximum singular value of $\mathbf{A}$.

The low-rank approximation of a matrix $\mathbf{A}$ can be computed using singular value decomposition (SVD). The SVD of $\mathbf{A}$ is a factorization of $\mathbf{A}$ into a rotation, followed by a rescaling, followed by another rotation [9].

$$\mathbf{A} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^\top \tag{16}$$

where $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{a \times b}$ is a sorted diagonal matrix of the singular values of $\mathbf{A}$ in descending order and $\tilde{\mathbf{U}} \in \mathbb{R}^{a \times a}, \tilde{\mathbf{V}} \in \mathbb{R}^{b \times b}$ are the matrices in which the columns are the corresponding left and right singular vectors, respectively[2].

[2] Standard notation for singular value decomposition is $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. In this paper, we use $\tilde{\mathbf{U}}$, $\tilde{\boldsymbol{\Sigma}}$, and $\tilde{\mathbf{V}}$ to avoid overloading $\mathbf{V}$.

We use this decomposition to reconstruct a low-rank approximation of the original matrix $\mathbf{M}$.

$$\mathbf{M} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^\top \tag{17}$$

Recall that $\mathbf{M} \in \mathbb{R}^{q\times d}$. Then, $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{q\times d}$. To reconstruct a rank r matrix, where $r < q$, we keep the top r rows of $\tilde{\boldsymbol{\Sigma}}$ and zero out the rest to create $\tilde{\boldsymbol{\Sigma}}'$. Then, we compute:

$$\mathbf{M}' = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'\tilde{\mathbf{V}}^\top \tag{18}$$

$\mathbf{M}'$ is a rank r approximation of the model weight matrix $\mathbf{M}$, where $r < q$. That is, $\mathbf{M}'$ is a rank r approximation of the weight matrix learned by a full-rank regressor.

By definition of SVD, the maximum singular value of the model $\mathbf{M}$ is preserved in its approximation $\mathbf{M}'$. That is,

$$||\mathbf{M}||_2 = ||\mathbf{M}'||_2 \tag{19}$$

Let the model $\mathbf{M}_r$ denote the learned coefficient matrix by RRR with rank constraint r.

Theorem 2.2 of Bernstein [15] states that minimizing equation (4) (i.e., RRR) is equivalent to minimizing equation (3) (i.e., OLS) then performing low-rank approximation via the SVD. That is, $\mathbf{M}' = \mathbf{M}_r$. It follows that:

$$||\mathbf{M}||_2 = ||\mathbf{M}'||_2 \tag{20}$$

$$= ||\mathbf{M}_r||_2 \tag{21}$$

Recall that attacker happiness for linear models is $||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2$. When $\mathbf{W} = \mathcal{I}_q$ and $\mathbf{C} = \mathcal{I}_d$, we have $\mathbf{V} = \mathcal{I}_q, \mathbf{G} = \mathcal{I}_d$. Then, attacker happiness is simply the spectral norm of the weight matrix $\mathbf{M}$.

$$\textit{Atkr.happ.}(\mathbf{M}, \mathcal{I}_q, \mathcal{I}_d) = ||\mathbf{M}||_2 \tag{22}$$

Since $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$ by equation (21), we have:

$$h_r = \frac{\textit{Atkr.happ.}(\mathbf{M}, \mathbf{W}, \mathbf{C})}{\textit{Atkr.happ.}(\mathbf{M}_r, \mathbf{W}, \mathbf{C})} \tag{23}$$

$$= \frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2} \tag{24}$$

$$= \frac{||\mathbf{M}||_2}{||\mathbf{M}||_2} \tag{25}$$

$$= 1 \tag{26}$$

Thus, attacker happiness is *unaffected* by reducing the rank of the coefficient matrix from $\min(d, q)$ to $1 \leq r < \min(d, q)$. □

Recall that RRR is a method of performing multi-target linear regression that, in comparison to full-rank regression, improves the test accuracy of the model. Often, improving the performance of a model may reduce adversarial

robustness. However, Theorem 1 states that any optimal linear learner with a maximum-rank constraint of r, where $1 \leq r \leq \min(d, q)$, will learn a model with equal adversarial robustness. In this case, **there is no decrease in robustness** for an increase in performance. This bound provides the ideal model robustness guarantee for any linear learner minimizing MSE.

4.2 Robustness in practical settings

While Theorem 1 covers the most commonly studied adversarial setting, exploring alternative defender–attacker configurations can offer further insight. Moreover, these settings, though less studied, are also commonly encountered in practice. In addition, they provide a more general lens through which to understand the effects of rank constraint on model robustness.

Recall that we want to understand how the choice of r may affect the vulnerability of the learned model. Hence, we aim to bound the ratio of attacker happiness for $\mathbf{M}$ to the attacker happiness for $\mathbf{M}_r$.

$$h_r = \frac{\mathit{Atkr.happ.}(\mathbf{M}, \mathbf{W}, \mathbf{C})}{\mathit{Atkr.happ.}(\mathbf{M}_r, \mathbf{W}, \mathbf{C})} \tag{27}$$

Bounds on h_r capture how much the attacker happiness may change as a result of the defender reducing its rank to r. Suppose $\mathit{Atkr.happ.}(\mathbf{M}, \mathbf{W}, \mathbf{C}) = a$, where $a \geq 0$ is some scalar value. If the bounds $\frac{1}{3} \leq h_r \leq 2$ hold, then it follows that $\frac{1}{2}a \leq \mathit{Atkr.happ.}(\mathbf{M}_r, \mathbf{W}, \mathbf{C}) \leq 3a$. Knowing these bounds on h_r allows the defender to, given the loss function it minimizes, assess the implications of choosing a rank constraint r on the vulnerability of the learned model.

We consider the following four adversarial settings, ranging from least to most general. Case 1 has already been discussed; for each of the additional three cases, we provide upper and lower bounds on h_r.

- **Case 1.** Defender loss is MSE:
  $\mathbf{W} = \mathcal{I}_q$, $\mathbf{C} = \mathcal{I}_d$
  This covers the classical setting. In this case, $h_r = 1$ (see section 4.1).

- **Case 2.** Defender loss is a weighted sum of targets, and the attacker is ℓ_2-bound.
  $\mathbf{W} = (w_{ij}),\ \forall i, j \in \{1, 2, ..., q\},\ i \neq j \implies w_{ij} = 0$, $\mathbf{C} = \mathcal{I}_d$
  This is a generalization of Case 1, where the defender predicts q targets and weights the importance of the targets differently (i.e., W is a diagonal matrix).

- **Case 3.** Defender loss is Mahalanobis squared norm, and the attacker is ℓ_2-bound.
  $\mathbf{W}$ is any arbitrary positive definite matrix, $\mathbf{C} = \mathcal{I}_d$
  This is a generalization of Case 2, where we consider the general defender minimizing squared Mahalanobis norm.

– **Case 4.** The general case.
  $\mathbf{W}$ and $\mathbf{C}$ are arbitrary positive definite matrices.

Case 2 represents a common and practically relevant setting. It is a generalization of Case 1, as it considers a larger subset of $\mathbf{W}$, including $\mathcal{I}_q$. Case 3 considers a general defender paired with a standard (i.e., ℓ_2-bounded) attacker, resulting in a broader robustness bound. This is a generalization of Case 2, as it considers all possible $\mathbf{W}$. Case 4 is the most general case, allowing for arbitrary defender and attacker constraints, and yields the bound that requires minimal assumptions.

4.3 Case 2. Defender considers a weighted sum of targets

Consider the case in which the defender cares more about its performance on certain targets than others. That is, the defender's loss is weighted MSE, where $\mathbf{W} = diag(\mathbf{w}), \mathbf{w} \succeq 0$. For $i \in \{1, ..., q\}$ $\mathbf{w}_i$ is the weight that represents how much the defender cares about its performance on the i-th target. The attacker is constrained by ℓ_2 norm (i.e., $\mathbf{C} = \mathcal{I}_d$). In this case, $\mathbf{W}$ is a *diagonal matrix*.

Running Example. In our running example, suppose the predictions of the companies' profits are informing investment and risk management decisions of a client. The client holds large equity positions in T_1 and T_2, making accurate forecasts for these companies particularly critical. As a result, prediction errors for T_1 and T_2 carry greater consequences than errors for T_3. This unequal weighting of targets is not captured when the defender's loss is MSE (i.e., when $\mathbf{W} = \mathcal{I}_q$).

Theorem 2. *If* $\mathbf{W}$ *is a diagonal matrix and* $\mathbf{C}$ *is the identity matrix* ($\mathbf{C} = \mathcal{I}_d$)*, the following bounds on* h_r *hold:*

$$\frac{\sigma_{q-r+1}(\mathbf{V})}{||\mathbf{V}||_2} \leq h_r \leq \frac{||\mathbf{V}||_2}{\sigma_{q-r+1}(\mathbf{V})}$$

where $\sigma_{q-r+1}(\mathbf{V})$ *denotes the* $(q-r+1)$*-th largest singular value of* $\mathbf{V}$ *(i.e., the* r*-th smallest singular value of* $\mathbf{V}$*).*

Proof. We aim to bound h_r, which is defined in equation (27) and repeated below:

$$h_r = \frac{Atkr.happ.(\mathbf{M}, \mathbf{W}, \mathbf{C})}{Atkr.happ.(\mathbf{M}_r, \mathbf{W}, \mathbf{C})}$$

Recall the definition of attacker happiness for linear models.

$$h_r = \frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} \tag{28}$$

Since $\mathbf{C} = \mathcal{I}_d$, we have $\mathbf{G}^{-1} = \mathcal{I}_d$.

$$h_r = \frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \tag{29}$$

We first prove the upper bound on h_r. By the submultiplicativity of the spectral norm[3], we have:

$$||\mathbf{VM}||_2 \leq ||\mathbf{V}||_2||\mathbf{M}||_2 \tag{30}$$

Consider the SVD of $\mathbf{M}$ and $\mathbf{M}_r$:

$$\mathbf{M} = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{V}}^\top \tag{31}$$

$$\mathbf{M}_r = \tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'\tilde{\mathbf{V}}^\top \tag{32}$$

where $\tilde{\boldsymbol{\Sigma}}'$ consists of the top r rows of $\tilde{\boldsymbol{\Sigma}}$.

We note that $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are orthogonal rotation matrices by definition of the SVD [9]. Then, the singular values of $\mathbf{VM}_r$ are determined only by $\mathbf{V}, \tilde{\mathbf{U}}$, and $\tilde{\boldsymbol{\Sigma}}'$. Hence, without loss of generality, we let $\tilde{\mathbf{V}} = \mathcal{I}_q$.

$$||\mathbf{VM}_r||_2 = ||\mathbf{V}\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'||_2 \tag{33}$$

By assumption, $\mathbf{V}$ is a diagonal matrix.

$$\mathbf{V} = \begin{bmatrix} v_1 & & \\ & \ddots & \\ & & v_q \end{bmatrix}$$

Then, $\mathbf{V}$ is a rescaling matrix, and the elements $\{v_1, ..., v_q\}$ of $\mathbf{V}$ are the singular values of $\mathbf{V}$.

Recall that $\tilde{\mathbf{U}}$ is a orthogonal rotation matrix. Therefore, the singular values of $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$ are equal to those of $\tilde{\boldsymbol{\Sigma}}'$, but the singular vectors of $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$ are determined by the columns of $\tilde{\mathbf{U}}$ and, importantly, may not be axis-aligned[4]. Consider the case in which $\tilde{\mathbf{U}}$ is a non-axis-aligned rotation matrix, i.e., a matrix such that the columns of $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$ are non-axis-aligned. Then, applying $\mathbf{V}$ will rescale $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$ along non-singular vectors. Clearly, this results in a smaller change in the singular values than when $\mathbf{V}$ directly rescales along the singular vectors of $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$. Hence, without loss of generality, let $\tilde{\mathbf{U}}$ be a permutation matrix. Then, $\tilde{\mathbf{U}}\tilde{\boldsymbol{\Sigma}}'$ is a diagonal matrix of the singular values of $\mathbf{M}'$ in some order, determined by $\tilde{\mathbf{U}}$.

The minimum $||\mathbf{VM}_r||_2$ occurs when the smallest values of $\mathbf{V}$ align with the largest singular values of $\mathbf{M}_r$.[5] We can choose $\tilde{\mathbf{U}}$ to be the permutation matrix such that this alignment occurs. Therefore, the following holds:

$$||\mathbf{VM}_r||_2 \geq \sigma_{q-r+1}(\mathbf{V})||\mathbf{M}_r||_2 \tag{34}$$

[3] For any two matrices $\mathbf{A}, \mathbf{B}$, $||\mathbf{AB}||_2 \leq ||\mathbf{A}||_2||\mathbf{B}||_2$.

[4] A set of vectors is said to be axis-aligned when each vector has at most one non-zero element.

[5] In the general setting (i.e., if $\mathbf{C}$ is an arbitrary $d \times d$ matrix), this optimization problem is the following: Given vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, find the permutation matrices $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{F}$ such that the maximum of the component-wise product of $\mathbf{Dx}$, $\mathbf{Ey}$, and $\mathbf{Fz}$ is minimized. Note that this is NP-hard to compute [10].

where $\sigma_{q-r+1}(\mathbf{V})$ denotes the $(q-r+1)$-th largest singular value of $\mathbf{V}$. Note that since $\mathbf{V} \in \mathbb{R}^{q\times q}$, the $(q-r+1)$-th largest singular value of $\mathbf{V}$ is the r-th smallest singular value of $\mathbf{V}$.

Finally, putting equations (30) and (34) together, we have:

$$\frac{||\mathbf{VM}||_2}{||\mathbf{VM}_r||_2} \leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2}{\sigma_{q-r+1}(\mathbf{V})||\mathbf{M}_r||_2} \tag{35}$$

Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$ by equation (21). Thus, we have:

$$h_r = \frac{||\mathbf{VM}||_2}{||\mathbf{VM}_r||_2} \tag{36}$$

$$\leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2}{\sigma_{q-r+1}(\mathbf{V})||\mathbf{M}_r||_2} \tag{37}$$

$$= \frac{||\mathbf{V}||_2}{\sigma_{q-r+1}(\mathbf{V})} \tag{38}$$

which upper bounds h_r as desired. □

We include the proof of the lower bound in Appendix A.1 for clarity. Note that this proof closely parallels the proof of the upper bound with minor differences.

In Case 2, the defender is a linear learner that considers weighted MSE. When the learner is constrained to rank r, the change in attacker happiness is bounded as a function of r. This bound is independent of $\mathbf{M}$; that is, it holds for any arbitrary linear model, and can be computed as long as $\mathbf{W}$ is known.

4.4 Case 3. General defender

While we have covered the two most common defenders in the area of multi-target regression, there may be scenarios in which $\mathbf{W}$ is some non-identity, non-diagonal matrix. Hence, we now generalize to the case where $\mathbf{W}$ is any positive definite matrix $\mathbf{W} \in \mathbb{R}^{q\times q}$, i.e., where the defender is any linear learner minimizing squared Mahalanobis norm. The attacker is constrained by ℓ_2 norm (i.e., $\mathbf{C} = \mathcal{I}_d$).

A non-identity, non-diagonal $\mathbf{W}$ is able to capture the preference for *error consistency*. If the model incorrectly predicts a lower value for one target but a higher value for another target, there is an inconsistency in the error of the model. This inconsistency may, in some applications, be more or less desirable than an error in which the model incorrectly predicts higher (or lower) values for both targets.

Running Example. In our running example, suppose the predictions of the companies' profits are used as inputs to a broader economic forecasting model. If T_1 and T_2 are both over- or both under-estimated, then the broader economic model will infer a growth/decline in the food service industry. If, instead, T_1's profit is over-estimated while T_2's is under-estimated (or vice versa), no such additive error will occur. This fact—that anti-correlated errors are better than correlated errors—is not captured when the defender's loss is decomposable into the sum of independent errors (when $\mathbf{W}$ is diagonal).

Theorem 3. *If* $\mathbf{W}$ *is any* $q \times q$ *matrix and* $\mathbf{C}$ *is the identity matrix* ($\mathbf{C} = \mathcal{I}_d$), h_r *is bounded by* $\mathbf{V}$*'s condition number* $\kappa(\mathbf{V})$ *and its reciprocal.*

$$\frac{1}{\kappa(\mathbf{V})} \leq h_r \leq \kappa(\mathbf{V})$$

Proof. We aim to bound h_r, which is defined in equation (27) and repeated below:

$$h_r = \frac{\mathit{Atkr.happ.}(\mathbf{M}, \mathbf{W}, \mathbf{C})}{\mathit{Atkr.happ.}(\mathbf{M}_r, \mathbf{W}, \mathbf{C})}$$

Recall the definition of attacker happiness for linear models.

$$h_r = \frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} \tag{39}$$

Since $\mathbf{C} = \mathcal{I}_d$, we have $\mathbf{G}^{-1} = \mathcal{I}_d$.

$$h_r = \frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \tag{40}$$

We first prove the upper bound on h_r. By the submultiplicativity of the spectral norm, we have:

$$||\mathbf{V}\mathbf{M}||_2 \leq ||\mathbf{V}||_2||\mathbf{M}||_2 \tag{41}$$

By Corollary 11.6.6 of Bernstein [3],

$$\sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M}_r) \leq \sigma_{\max}(\mathbf{V}\mathbf{M}_r) \tag{42}$$

By definition of spectral norm,

$$||\mathbf{V}\mathbf{M}_r||_2 \geq \sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2 \tag{43}$$

Combining equations (41) and (43), we have:

$$\frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2} \tag{44}$$

Note that the definition of the condition number of a matrix $\mathbf{A}$ is:

$$\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

Hence, we have:

$$h_r = \frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \tag{45}$$

$$\leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2} \tag{46}$$

$$= \frac{\sigma_{\max}(\mathbf{V})||\mathbf{M}||_2}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2} \tag{47}$$

$$= \kappa(\mathbf{V})\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2} \tag{48}$$

Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$ by equation (21). Therefore,

$$h_r \leq \kappa(\mathbf{V}) \tag{49}$$

which upper bounds h_r as desired. □

We include the proof of the lower bound in Appendix A.2 for clarity. Note that this proof closely parallels the proof of the upper bound with minor differences.

In Case 3, the defender is a linear learner that considers mean squared Mahalanobis norm with *any* $\mathbf{W}$, i.e., it considers the general defender. When the learner is constrained to rank r, the change in attacker happiness is bounded by bounds that are independent of $\mathbf{M}$ and r.

4.5 Case 4. The general case

Cases 1, 2, and 3 all assume the presence of an ℓ_2-bounded attacker, i.e., we assume that $\mathbf{C} = \mathcal{I}_d$. However, there may be scenarios in which $\mathbf{C}$ is some non-identity, non-diagonal matrix. We now generalize to the case where, in addition to $\mathbf{W}$ being any matrix $\mathbf{W} \in \mathbb{R}^{q\times q}$, $\mathbf{C}$ is now any positive definite matrix $\mathbf{C} \in \mathbb{R}^{d\times d}$.

Running Example. In our running example, the attacker has limited capability to manipulate prices of various raw materials. Note, however, that these prices are not the features that the learner uses to predict the companies' profits. Instead, the prices indirectly affect the features. As such, the attacker is not constrained by an ℓ_2 ball as in the classical setting. Because manipulating the price of raw materials will have joint effects on the input features (the prices of goods), the attacker is better modeled by being constrained by a general ellipsoid. For example, increasing the price of cloth may decrease the number of aprons and shirts sold while simultaneously increasing the amount of laundry detergent sold.

Theorem 4. *If* $\mathbf{W} \in \mathbb{R}^{q\times q}$ *and* $\mathbf{C} \in \mathbb{R}^{d\times d}$, h_r *is bounded by the product of the condition numbers of* $\mathbf{V}$ *and* $\mathbf{G}$ *and its reciprocal.*

$$\frac{1}{\kappa(\mathbf{V})\kappa(\mathbf{G})} \leq h_r \leq \kappa(\mathbf{V})\kappa(\mathbf{G})$$

Proof. We aim to bound h_r, which is defined in (27) and repeated below:

$$h_r = \frac{\mathit{Atkr.happ.}(\mathbf{M}, \mathbf{W}, \mathbf{C})}{\mathit{Atkr.happ.}(\mathbf{M}_r, \mathbf{W}, \mathbf{C})}$$

Recall the definition of attacker happiness for linear models.

$$h_r = \frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} \tag{50}$$

We first prove the upper bound on h_r. By the submultiplicativity of the spectral norm, we have:

$$||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2 \leq ||\mathbf{V}||_2||\mathbf{M}||_2||\mathbf{G}^{-1}||_2 \tag{51}$$

By Corollary 11.6.6 of Bernstein [3],

$$\sigma_{\max}(\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}) \geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M}_r\mathbf{G}^{-1}) \tag{52}$$

and by Corollary 11.6.7 of Bernstein [3],

$$\sigma_{\max}(\mathbf{M}_r\mathbf{G}^{-1}) \geq \sigma_{\max}(\mathbf{M}_r)\sigma_{\min}(\mathbf{G}^{-1}) \tag{53}$$

Combining (52) and (53), we have:

$$\begin{aligned}\sigma_{\max}(\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}) &\geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M}_r\mathbf{G}^{-1}) &(54)\\ &\geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M}_r)\sigma_{\min}(\mathbf{G}^{-1}) &(55)\end{aligned}$$

By definition of spectral norm,

$$||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2 \geq \sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2\sigma_{\min}(\mathbf{G}^{-1}) \tag{56}$$

Putting (51) and (56) together, we have:

$$\frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} \leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2||\mathbf{G}^{-1}||_2}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2\sigma_{\min}(\mathbf{G}^{-1})} \tag{57}$$

Note that the condition number $\kappa(\mathbf{A})$ is defined as $\frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$. Therefore:

$$\begin{aligned}h_r &= \frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} &(58)\\ &\leq \frac{||\mathbf{V}||_2||\mathbf{M}||_2||\mathbf{G}^{-1}||_2}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2\sigma_{\min}(\mathbf{G}^{-1})} &(59)\\ &= \frac{\sigma_{\max}(\mathbf{V})||\mathbf{M}||_2\sigma_{\max}(\mathbf{G}^{-1})}{\sigma_{\min}(\mathbf{V})||\mathbf{M}_r||_2\sigma_{\min}(\mathbf{G}^{-1})} &(60)\\ &= \kappa(\mathbf{V})\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2}\kappa(\mathbf{G}^{-1}) &(61)\\ & &(62)\end{aligned}$$

For any invertible matrix $\mathbf{A}$, $\kappa(\mathbf{A}) = \kappa(\mathbf{A}^{-1})$. Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$. Therefore, we have:

$$\begin{aligned}h_r &= \frac{||\mathbf{V}\mathbf{M}\mathbf{G}^{-1}||_2}{||\mathbf{V}\mathbf{M}_r\mathbf{G}^{-1}||_2} &(63)\\ &\leq \kappa(\mathbf{V})\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2}\kappa(\mathbf{G}^{-1}) &(64)\\ &= \kappa(\mathbf{V})\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2}\kappa(\mathbf{G}) &(65)\\ &= \kappa(\mathbf{V})\kappa(\mathbf{G}) &(66)\end{aligned}$$

Thus, we have

$$h_r \leq \kappa(\mathbf{V})\kappa(\mathbf{G}) \tag{67}$$

which upper bounds h_r as desired. □

We include the proof of the lower bound in Appendix A.3 for clarity. Note that this proof closely parallels the proof of the upper bound with minor differences.

5 Discussion

In this paper we considered a series of scenarios—from the classical setting where the defender minimizes MSE and the attacker is constrained by an ℓ_2 bound to the realistic setting where both defender and attacker use general Mahalanobis metrics—and shed light on the effects of rank restrictions on adversarial robustness for linear regression. We proved seven theoretical bounds on the change in adversarial robustness for multi-target linear regressions when incorporating a maximum rank constraint on the model.

These bounds provide a tool for practitioners making design decisions about learners when adversaries are at play. In Cases 1, 3, and 4, the bounds are independent of r. In Case 1, this is powerful—no matter what width r the defender chooses, the defender knows that it will not affect the vulnerability of the model. Hence, it is able to select the constraint on model rank solely based on model performance (without risking a decrease in robustness).

In Cases 3 and 4, the defender only knows the maximum amount of change that a rank constraint could cause on the vulnerability of the model. That is, no matter what constraint the defender chooses, the upper bound on the change in robustness remains the same value. We note that Theorems 3 and 4 are looser bounds than 1 and 2. However, they are also the most general, as they apply to any defender (and, in Case 4, any attacker). Only the bounds in Theorem 2 depend on the rank constraint—specifically, the upper (lower) bound decreases (increases) as r approaches q. The more constrained the learner is, the more it may affect the vulnerability of your model. Knowing these bounds allows the defender to assess the implications of choosing a rank r on the vulnerability of the model. In addition, these insights contribute to a deeper understanding of how model compression techniques (e.g., distillation, weight pruning, or low-rank approximation) may compromise robustness.

In addition, our presented bounds further illuminate the connection between adversarial robustness and matrix spectra for linear models.

A defining characteristic of these bounds is their dependency on the singular values of $\mathbf{V}$. All four bounds are a function of the singular values of $\mathbf{V}$, which depends on $\mathbf{W}$ (where the defender loss is $||\cdot||^2_{\mathbf{W}}$). That is, they are specific to the defender and independent of the input x. In general, the possible change in attacker happiness is defined by the **range of the singular values of V.** For instance, in Case 2, if the largest singular value $\sigma_1(\mathbf{V})$ is far from the second largest singular value $\sigma_2(\mathbf{V})$, i.e., $\mathbf{W}^\top\mathbf{W}$ has a large spectral gap, simply constraining the rank of the model to $r = q - 1$ can result in a large change in attacker happiness.

Finally, our investigation highlights an important lesson regarding adversarial learning. The classical setting is mathematically elegant and well studied, but is

(a) unrealistic for a broad range of real-world tasks and (b) fails to capture the full nuance of how design decisions affect model robustness.

6 Related Works

Deployment-time attacks against classifiers are often called "evasion" attacks [20] and date back to Dalvi et al. [6] and Lowd and Meek [12]. One canonical example is spam detection [13]. Here, an attacker aims to slip past the spam filter of the defender by perturbing the original message so as to be classified as legitimate by the defender. Such attacks against deep learning methods have been extensively studied [4,7,5].

Deployment-time attacks against regression models, however, remain understudied. Grosshans et al. [8] consider attacks on regression methods under a different threat model, and use game-theoretic approaches to derive attacks. More closely related to the work presented herein, Alfeld et al [1] develop deployment-time attacks against autoregressive forecasters, where the attacker aims to perturb past values so as to change the defender's forecast of the future. In later work, the authors derive optimal defenses against linear models by efficiently computing the optimal repulsive attack [2]. We utilize their notation and rely on their Theorem 2 (defining the optimal attack) in our proofs.

Outside of adversarial contexts, there have been various methods of performing multi-target linear regression. Filtered Canonical Y-variate Regression (FICYREG) leverages canonical variates and canonical correlations to formulate an optimization problem that forces the explicit learning of relationships between targets [19]. Simila and Tikka [18] adds a regularization term that results in an optimization problem equivalent to minimizing MSE subject to a *sparsity constraint* (i.e., by solving the ℓ_2-SVS problem). Reduced-rank regression (RRR), is a method that constrains the rank of the parameter matrix in the optimization problem [11]. While the focus in this paper has been RRR, a promising avenue for future work is to perform similar analysis of the adversarial robustness of other multi-target regression methods.

Prior work has empirically demonstrated that reduced-rank models are often less robust than their full-rank counterpart because the low-rank structure between the input features and targets is easily distorted by outliers in the data. As such, a method of defense that specifically addresses this weakness for RRR was proposed, in which sparse mean-shift parameterization is utilized during training [17]. Our work complements prior work by forming a theoretical foundation supporting the observed empirical behaviors.

The robustness of reduced-rank models have also been studied in other contexts. As deep neural networks (DNNs) continue to grow in size (i.e., number of parameters), the task of reducing training and inference costs has become increasingly important. Model compression techniques, such as low-rank approximation, are commonly used to address this challenge. However, low-rank approximation often degrades adversarial robustness. Savostianova et al. [16] analyzes this trade-off by discussing robustness in terms of the *local* condition number of

the DNN. In their threat model, the attacker is constrained to a perturbation vector with norm $||\boldsymbol{\alpha}|| \leq \epsilon||\mathbf{x}||$, where $\epsilon \in \mathbb{R}$ is a fixed small multiplicative factor. This threat model provides an alternative setting to which our work can be applied and provides an interesting avenue for future work.

7 Conclusion

In this paper we derived seven theoretical bounds on the change in adversarial robustness for multi-target linear regression when incorporating a maximum rank constraint on the model. These bounds are directly applicable by practitioners and cover a range of common defender and attacker settings. The bounds and their proofs lend insights into the connections between adversarial robustness and the spectra of matrices defining defender loss ($\mathbf{W}$) and attacker capability ($\mathbf{C}$). Most prominently, we find that in the classical setting (the learner minimizes Mean Squared Error and the attacker is constrained by an ℓ_2 constraint), adversarial robustness is *unaffected* by rank reduction. Under more general and practical settings, rank constraints can dramatically affect robustness, the extent of which is bounded by a function of the eigenvalues of $\mathbf{W}$. Through studying reduced-rank regression, we strengthen a foundational understanding of the adversarial robustness of low-rank models and highlight the importance of examining non-classical yet practically relevant adversarial settings.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

A Lower Bounds on Adversarial Robustness

A.1 Proof: Theorem 2 Lower Bound

Proof. By Proposition 11.6.4 of Bernstein [3], for $i \in \{1, ..., q\}$,

$$\sigma_i(\mathbf{V})\sigma_{\max}(\mathbf{M}) \leq \sigma_i(\mathbf{VM}) \tag{68}$$

where $\sigma_i(\cdot)$ is the i-th largest singular value.
Since $1 \leq r \leq q$, we have that $1 \leq (q - r + 1) \leq q$. Hence, the following holds:

$$\sigma_{q-r+1}(\mathbf{V})\sigma_{\max}(\mathbf{M}) \leq \sigma_{q-r+1}(\mathbf{VM}) \tag{69}$$

By definition of the spectral norm, we have:

$$\begin{aligned} \sigma_{q-r+1}(\mathbf{V})||\mathbf{M}||_2 &= \sigma_{q-r+1}(\mathbf{V})\sigma_{\max}(\mathbf{M}) && (70)\\ &\leq \sigma_{q-r+1}(\mathbf{VM}) && (71)\\ &\leq \sigma_{max}(\mathbf{VM}) && (72)\\ &= ||\mathbf{VM}||_2 && (73) \end{aligned}$$

By the submultiplicativity of the spectral norm,

$$||\mathbf{V}\mathbf{M}_r||_2 \leq ||\mathbf{V}||_2||\mathbf{M}_r||_2 \tag{74}$$

Combining equations (73) and (74), we have:

$$h_r = \frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \tag{75}$$

$$\geq \frac{\sigma_{q-r+1}(\mathbf{V})||\mathbf{M}||_2}{||\mathbf{V}||_2||\mathbf{M}_r||_2} \tag{76}$$

Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$ by equation (21). Therefore,

$$h_r \geq \frac{\sigma_{q-r+1}(\mathbf{V})}{||\mathbf{V}||_2} \tag{77}$$

which lower bounds h_r as desired. □

A.2 Proof: Theorem 3 Lower Bound

Proof. By Corollary 11.6.6 of Bernstein [3], we have:

$$\sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M}) \leq \sigma_{\max}(\mathbf{V}\mathbf{M}) \tag{78}$$

By the definition of spectral norm,

$$\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2 \leq ||\mathbf{V}\mathbf{M}||_2 \tag{79}$$

By the submultiplicativity of the spectral norm, we have:

$$||\mathbf{V}\mathbf{M}_r||_2 \leq ||\mathbf{V}||_2||\mathbf{M}_r||_2 \tag{80}$$

Combining equations (79) and (80), we have:

$$\frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \geq \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2}{||\mathbf{V}||_2||\mathbf{M}_r||_2} \tag{81}$$

By definition of the condition number,

$$h_r = \frac{||\mathbf{V}\mathbf{M}||_2}{||\mathbf{V}\mathbf{M}_r||_2} \tag{82}$$

$$\geq \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2}{||\mathbf{V}||_2||\mathbf{M}_r||_2} \tag{83}$$

$$= \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2}{\sigma_{\max}(\mathbf{V})||\mathbf{M}_r||_2} \tag{84}$$

$$= \frac{1}{\kappa(\mathbf{V})}\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2} \tag{85}$$

Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$ by equation (21). Then,

$$h_r \geq \frac{1}{\kappa(\mathbf{V})} \tag{86}$$

which lower bounds h_r as desired. □

A.3 Proof: Theorem 4 Lower Bound

Proof. By Corollary 11.6.6 of Bernstein [3],

$$\sigma_{\max}(\mathbf{VMG}^{-1}) \geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{MG}^{-1}) \tag{87}$$

and by Corollary 11.6.7 of Bernstein [3],

$$\sigma_{\max}(\mathbf{MG}^{-1}) \geq \sigma_{\max}(\mathbf{M})\sigma_{\min}(\mathbf{G}^{-1}) \tag{88}$$

Combining (87) and (88), we have:

$$\sigma_{\max}(\mathbf{VMG}^{-1}) \geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{MG}^{-1}) \tag{89}$$

$$\geq \sigma_{\min}(\mathbf{V})\sigma_{\max}(\mathbf{M})\sigma_{\min}(\mathbf{G}^{-1}) \tag{90}$$

By definition of spectral norm,

$$||\mathbf{VMG}^{-1}||_2 \geq \sigma_{\min}(\mathbf{V})||\mathbf{M}||_2\sigma_{\min}(\mathbf{G}^{-1}) \tag{91}$$

By the submultiplicativity of the spectral norm,

$$||\mathbf{VM}_r\mathbf{G}^{-1}||_2 \leq ||\mathbf{V}||_2||\mathbf{M}_r||_2||\mathbf{G}^{-1}||_2 \tag{92}$$

Putting (91) and (92) together, we have:

$$\frac{||\mathbf{VMG}^{-1}||_2}{||\mathbf{VM}_r\mathbf{G}^{-1}||_2} \geq \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2\sigma_{\min}(\mathbf{G}^{-1})}{||\mathbf{V}||_2||\mathbf{M}_r||_2||\mathbf{G}^{-1}||_2} \tag{93}$$

By the definition of the spectral norm and the condition number $\kappa(\cdot)$:

$$h_r = \frac{||\mathbf{VMG}^{-1}||_2}{||\mathbf{VM}_r\mathbf{G}^{-1}||_2} \tag{94}$$

$$\geq \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2\sigma_{\min}(\mathbf{G}^{-1})}{||\mathbf{V}||_2||\mathbf{M}_r||_2||\mathbf{G}^{-1}||_2} \tag{95}$$

$$= \frac{\sigma_{\min}(\mathbf{V})||\mathbf{M}||_2\sigma_{\min}(\mathbf{G}^{-1})}{\sigma_{\max}(\mathbf{V})||\mathbf{M}_r||_2\sigma_{\max}(\mathbf{G}^{-1})} \tag{96}$$

$$= \frac{||\mathbf{M}||_2}{\kappa(\mathbf{V})||\mathbf{M}_r||_2\kappa(\mathbf{G}^{-1})} \tag{97}$$

$$\tag{98}$$

For any invertible matrix $\mathbf{A}$, $\kappa(\mathbf{A}) = \kappa(\mathbf{A}^{-1})$. Recall that $||\mathbf{M}||_2 = ||\mathbf{M}_r||_2$. Therefore, we have:

$$h_r = \frac{||\mathbf{VMG}^{-1}||_2}{||\mathbf{VM}_r\mathbf{G}^{-1}||_2} \tag{99}$$

$$\geq \frac{1}{\kappa(\mathbf{V})}\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2}\frac{1}{\kappa(\mathbf{G}^{-1})} \tag{100}$$

$$= \frac{1}{\kappa(\mathbf{V})}\frac{||\mathbf{M}||_2}{||\mathbf{M}_r||_2}\frac{1}{\kappa(\mathbf{G})} \tag{101}$$

$$= \frac{1}{\kappa(\mathbf{V})\kappa(\mathbf{G})} \tag{102}$$

which lower bounds h_r as desired. □

References

1. Alfeld, S., Zhu, X., Barford, P.: Data poisoning attacks against autoregressive models. Proceedings of the AAAI Conference on Artificial Intelligence (Feb 2016)
2. Alfeld, S., Zhu, X., Barford, P.: Explicit defense actions against test-set attacks. Proceedings of the AAAI Conference on Artificial Intelligence (Feb 2017)
3. Bernstein, D.S.: Scalar, Vector, and Matrix Mathematics. Princeton University Press (2018)
4. Biggio, B., Roli, F.: Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition p. 317–331 (Dec 2018)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks (2017)
6. Dalvi, N.N., Domingos, P.M., Mausam, Sanghai, S.K., Verma, D.: Adversarial classification. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (2004)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. Proceedings of the International Conference on Learning Representations (2014)
8. Großhans, M., Sawade, C., Brückner, M., Scheffer, T.: Bayesian games for adversarial regression problems. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research (2013)
9. Horn, R.A., Johnson, C.R.: Matrix Analysis (2nd edition). Cambridge University Press (2013)
10. Hsu, W.L.: Approximation algorithms for the assembly line crew scheduling problem. Mathematics of Operations Research **9**(3) (1984)
11. Izenman, A.J.: Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis **5**(2), 248–264 (1975)
12. Lowd, D., Meek, C.: Adversarial learning. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Association for Computing Machinery (2005)
13. Lowd, D., Meek, C.: Good word attacks on statistical spam filters. Proceedings of the International Conference on Email and Anti-Spam (01 2005)
14. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition **abs/1511.04599** (2015)
15. Reinsel, G.C., Velu, R.P., Chen, K.: Multivariate Reduced-Rank Regression: Theory, Methods and Applications. Springer (1998)
16. Savostianova, D., Zangrando, E., Ceruti, G., Tudisco, F.: Robust low-rank training via approximate orthonormal constraints. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
17. She, Y., Chen, K.: Robust reduced-rank regression. Biometrika (07 2017)
18. Similä, T., Tikka, J.: Input selection and shrinkage in multiresponse linear regression. Computational Statistics & Data Analysis (09 2007)
19. Van Der Merwe, A., Zidek, J.V.: Multivariate regression analysis and canonical variates. Canadian Journal of Statistics (1980)
20. Vorobeychik, Y., Kantarcioglu, M.: Adversarial Machine Learning. Morgan Claypool Publishers (2018)