# Evaluating Adversarial Robustness of Linear Models: On Kernels and Sparsity

**Shawn Shan, Ziyu Ye**
Department of Computer Science
University of Chicago
shansixiong, ziyuye@uchicago.edu

## 1 Introduction

Robustness is an essential property for a system which indicates if the system can work reliably and stably. In machine learning systems, we are especially concerned about **adversarial robustness** which considers the robustness of the system when facing adversarial inputs. While there exists rich literature in adversarial attacks in deep neural networks (DNNs) [6], the fundamental principle for adversarial robustness is far from understood.

In this paper, we aim to understand adversarial robustness from the perspective of **linear models**. We believe our work would help fill the gap of current adversarial robustness study in linear models, and such simplified models should provide relevant insights and heuristics for the more complex scenarios with DNNs. Specifically, we are interested in how linear models behaves under adversarial attacks with different **kernels** and **sparsity**. We choose these two factors as we believe they are fundamental in modern linear models and have close connections to practical DNNs.

## 2 Literature Review

**Adversarial robustness of linear models.** [1] first attempts to attack support vector machines (SVMs) by injecting label noise and propose to improve the robustness by kernel matrix correction. [2] proposes a poison attack on SVMs by gradient ascent, and [8] suggests a defense against it by approximating features by a low-rank matrix. Later on, [9] shows that linear classifier are less robust when facing higher-dimensional inputs. [10] claims that an adversarially robust linear classifier requires higher signal-to-noise ratio, and [7] further confirms this by illustrating the tradeoffs of robustness and accuray for linear models.

**The effect of kernels.** To the best of our knowledge, we are **the first** to study the effect of different kernels for linear models under adversarial attacks. Prior work mainly focuses on designing attacks on linear models when kernel is present [1], or designing defenses with a certain kernel [5, 11].

**The effect of sparsity.** [3] discusses that learning a sparse representation of input data helps to improve adversarial robustness. [4] further illustrates the intrinsic relationship of model sparsity and adversarial robustness.

## 3 Problem Statement

We are interested in how (generalized) linear models behave under different settings of activation, kernel function and sparsity. We consider models from the family of **Regularized Kernel Models**.

In evaluating the effect of **kernels**, we consider *kernel SVM* using linear, polynomial, Gaussian or sigmoid kernels; in evaluating the effect of **sparsity**, we consider *regularized linear regression* using $L^1$ regularization (*i.e.* lasso regression) or $L^2$ regularization (*i.e.* ridge regression).

## References


[1] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *Asian conference on machine learning*, pages 97–112, 2011.

[2] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1467–1474, 2012.

[3] Soorya Gopalakrishnan, Zhinus Marzi, Upamanyu Madhow, and Ramtin Pedarsani. Combating adversarial attacks using sparse representations, 2018.

[4] Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse dnns with improved adversarial robustness. *Advances in neural information processing systems*, 31:242–251, 2018.

[5] Yuying Hao, Tuanhui Li, Yong Jiang, Xuanye Cheng, and Li Li. Defending against adversarial examples using defense kernel network. In *BMVC*, page 77, 2019.

[6] Han Xu Yao Ma Hao-Chen, Liu Debayan Deb, Hui Liu Ji-Liang Tang Anil, and K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[7] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. *arXiv preprint arXiv:2002.10477*, 2020.

[8] Chang Liu, Bo Li, Yevgeniy Vorobeychik, and Alina Oprea. Robust linear regression against training data poisoning. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 91–102, 2017.

[9] István Megyeri, István Hegedűs, and Márk Jelasity. Adversarial robustness of linear models: Regularization and dimensionality. 2019.

[10] Xupeng Shi and A Adam Ding. Understanding and quantifying adversarial examples existence in linear classification. *arXiv preprint arXiv:1910.12163*, 2019.

[11] Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi, and Ghassan Hamarneh. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11340–11349, 2019.