Benign Overfitting and the Robustness Trade-Off

Kenneth Chow Sreedeekshita Gorugantu Venkata
Artem Kiryukhin Niket Patel Warren Wu

June 7, 2025

Abstract: Benign over-fitting describes the surprising ability of heavily over-parameterized models to interpolate noisy training sets while still generalizing to clean test data. Yet those same models can be derailed by imperceptible, adversarial perturbations. This project asks whether classical ℓ_2-regularization can reconcile that tension—preserving interpolation accuracy while hardening models against adversarial attacks. Guided by these insights from prior works, we generate high-dimensional synthetic classification tasks, sweep the ridge penalty λ, and chart the interplay between standard risk and projected-gradient adversarial risk, and we analyze the affect that the choice of input dimension has on adversarial robustness. Finally, we test the external validity of the trends by training shallow ReLU networks on MNIST with and without weight decay. The study aims to clarify when regularization can simultaneously enable benign over-fitting and adversarial robustness, and when fundamental trade-offs remain unavoidable.

1 Introduction to Benign Overfitting and Adversarial Robustness

Machine learning models today often exhibit an interesting phenomenon known in the literature in deep learning theory as benign overfitting [Bartlett et al., 2020]. This occurs when a model perfectly fits its training data—including any noise or incorrect labels, while still achieving strong generalization performance on unseen test data. This behavior is particularly notable in highly parameterized models trained on datasets containing noisy labels, and is closely related to the double descent phenomenon [Nakkiran et al., 2021]. Despite achieving zero training error by fitting the noise exactly, models can still maintain a relatively low test error. In particular, benign overfitting tends to arise under specific conditions: notably, the dimensionality of the feature space must significantly exceed the number of training samples ($d >> n$), and is often studied in the minimum norm interpolating / max-margin solution.

In parallel to the phenomenon of benign overfitting, another interesting aspect of modern deep learning is adversarial robustness. Adversarial robustness refers to a model's ability to maintain correct predictions even when inputs are deliberately perturbed to make the models predictions incorrect [Szegedy et al., 2014]. Specifically, an adversarial example is created by first order optimization to find a small perturbation to a legitimate input,

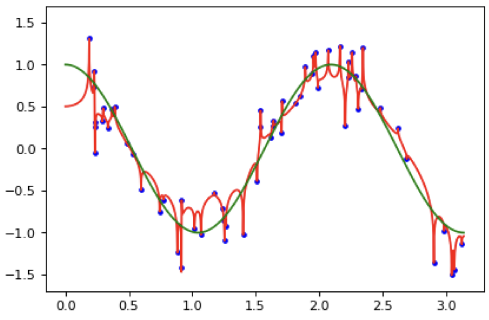


Figure 1: Example of benign overfitting, where a trained function, in red, fits the noisy data-points, despite remaining close to the true signal, in green. The model here is a linear regression with cosine features [Tsigler and Bartlett, 2023].

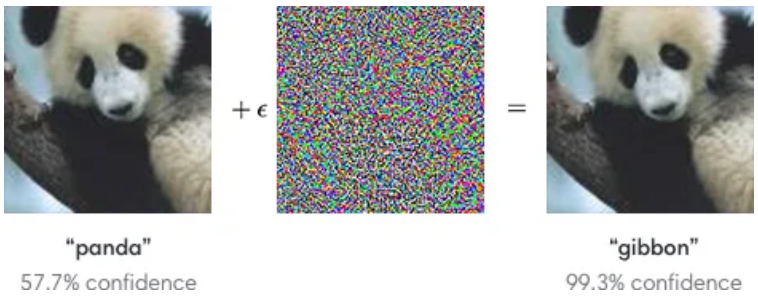


Figure 2: Example of Adversarial Example in Image Classification Goodfellow et al. [2017].

resulting in a misclassification even under very small noise. Robust accuracy measures a model's performance against these worst-case perturbed inputs. Adversarial attacks leverage optimization methods such as Projected Gradient Descent (PGD) or the Fast Gradient Sign Method (FGSM) to identify the perturbation. Achieving robustness often introduces a trade-off with standard accuracy, as robust models typically require larger classification margins or the implementation of explicit defense strategies.

The interaction between benign benign overfitting and adversarial robustness is of particular interest, as the two phenomena can conflict fundamentally. Benign overfitting favors solutions that precisely match noisy training data, while adversarial robustness necessitates margin-rich solutions capable of absorbing input perturbations. Recent work by Hao and Zhang [2024] demonstrate theoretically that benign overfitting negatively impacts adversarial robustness specifically in regression. This results make sense intuitively, as if a model fits noise in the training data, then it is likely that the decision boundary can be rough in spots and it will be easier to find adversarial examples.

Building upon these insights, we seek to expand the understanding of the relationship between benign overfitting and adversarial robustness in classification, exploring these dynamics in logistic regression and simple neural network architectures. Since benign overfitting is often associated with high-dimensional settings where the number of features greatly exceeds the number of samples, we also aim to investigate how changes in input dimensionality affect both generalization and robustness. This investigation aims to provide further clarity regarding how benign overfitting and the norm of the weights might influence model

robustness.

2 Interpreting Closed-Form Expressions for Adversarial Risk

We plan on interpreting the closed-form expressions for adversarial risk in ridge linear regression Hao and Zhang [2024]. We will explore how ridge regression estimators will behave in certain conditions and also explore how the solution will change according to the norm.

2.1 Defining an Adversarial Attack

We first define what an adversarial attack is. An adversarial attack's main goal is to disturb a machine learning model from learning, specifically increasing the loss function by perturbing input data. This can be defined as $x_{\text{adv}} = x + \delta$, where the constraint can be defined as an ℓ_p-norm (e.g., $\|\delta\|_p \leq \epsilon$). This will cause the model to make incorrect predictions on data.

2.2 Loss Functions

We can begin by diving deep into how regression and classification models are impacted.

Linear Regression Loss The linear regression loss function can be defined as follows:

For n samples $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, the loss is:

$$L_{\text{MSE}}(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

Furthermore, we can expand this in terms of ridge regression, where we add the ℓ_2 penalty term onto the weight:

$$J(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\|w\|_2^2$$

λ is defined as the regularization parameter.

Logistic Regression Loss Next, we will look at the loss for logistic regression. In this case, the Binary Cross-Entropy is given as follows:

For $y_i \in \{0, 1\}$, the loss is:

$$L_b(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\frac{1}{1+e^{-w^T x_i}}\right) + (1-y_i)\log\left(1 - \frac{1}{1+e^{-w^T x_i}}\right)\right]$$

2.3 Maximizing Adversarial Risk

The main goal is to understand how to maximize the adversarial risk, where adversarial risk can be defined as the expected loss in the worst-case scenario perturbations.

We can write the adversarial risk as R_{adv} Hao and Zhang [2024]:

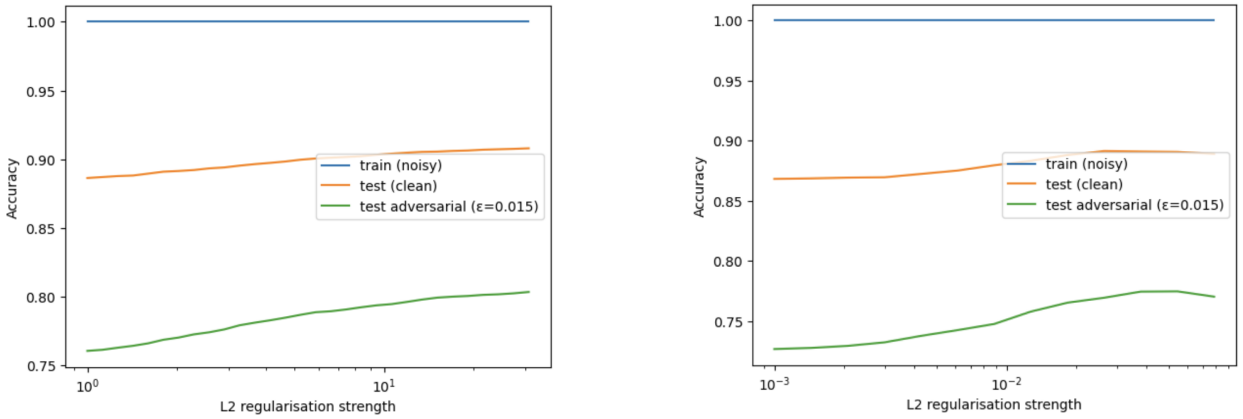
$$R_{\text{adv}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|\delta\|_p \leq \epsilon} L(f(x+\delta;\theta), y)\right]$$

Where $\mathcal{D}$ represents the distribution of data. In summary, the formula helps find the most harmful modification to the data given a perturbation δ.

By understanding how adversarial attacks and risk work, this allows us to understand how the regularization term can influence the model's vulnerability.

3 Experiments on Synthetic Data

A common technique to improve models is to introduce regularization of the norm of the weights. We begin with a controlled synthetic study to probe how ℓ_2-regularization (weight decay) influences the trade-off between standard generalization and adversarial robustness. We consider a synthetic dataset consisting of a simple classification task with two classes, where each class is a gaussian. In the training dataset, we randomly flip 10 percent of labels. We specifically choose to study this in a setting where we see Benign Overfitting, and we demonstrate that adversarial robustness improves with higher levels of regularization in Figure 3.



(a) Logistic Regression

(b) Shallow MLP

Figure 3: On the same data, we see here evidence that in both logistic regression and in shallow MLPs for classification, increasing the ℓ_2 regularization parameter leads to stronger performance on adversarial datasets.

4 Impact of Input Dimensionality on Generalization and Robustness

To investigate how increasing input dimensionality affects both generalization and adversarial robustness, we conduct controlled experiments using synthetic binary classification data. We train L2-regularized logistic regression models across input dimensions ranging from 10 to 5000, keeping the number of samples fixed at 1000.

4.1 Methodology

Each dataset is generated with all features being informative and includes slight label noise (flip probability = 0.01) to mimic real-world imperfections. We compute the standard training and test accuracy, as well as the weight norm and adversaral accuracy computing with PGD.

4.2 Observations

The results, shown in Figures 4 and 5, support the key observations:

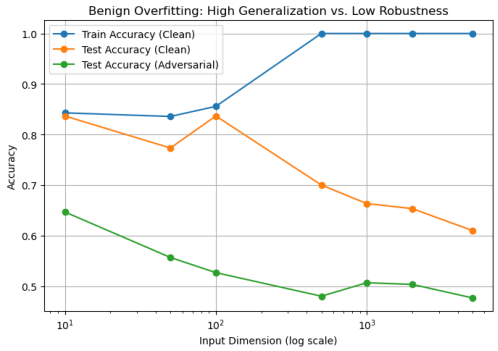


Figure 4: Effect of input dimensionality on model performance. As input dimension increases, training accuracy approaches 100%, but test and adversarial accuracy deteriorate—exposing a trade-off between benign overfitting and robustness.

- **Training accuracy** improves with dimension and reaches 100% when $d \geq 500$, illustrating classic benign overfitting behavior.
- **Test accuracy**, while initially high (84% at $d = 100$), gradually declines to around 61% at $d = 5000$. This indicates degradation in generalization as dimensionality increases, despite zero training error.
- **Adversarial accuracy** drops more sharply, from 65% at $d = 10$ to as low as 47% at $d = 5000$, highlighting the vulnerability of these models to small perturbations even when generalization on clean data is acceptable.

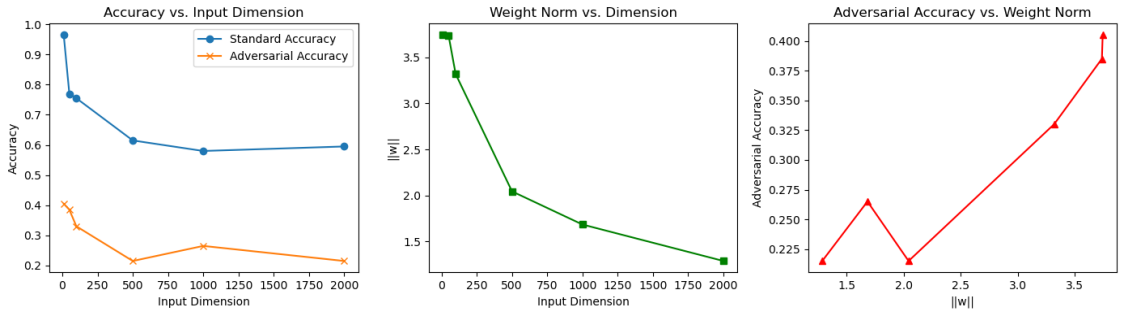


Figure 5: (1) Accuracy (clean and adversarial) declines with higher input dimension despite perfect training accuracy. (2) Weight norm decreases with input dimension, suggesting implicit regularization. (3) Adversarial accuracy shows non-monotonic relation with weight norm, indicating that smaller norms don't guarantee robustness.

4.3 Insights

These findings reinforce that benign overfitting—characterized by perfect training performance despite noisy data—can coexist with moderate test performance. However, robustness to adversarial attacks deteriorates substantially with increasing input dimension, even as the weight norm decreases. This supports prior theoretical findings that smaller norms do not necessarily imply better adversarial robustness and that additional mechanisms (e.g., adversarial training or margin-maximization) may be required in high-dimensional regimes.

5 Experiments on MNIST

In this section, we investigated if the trends observed in synthetic data in the previous sections of this paper were observable in an empirical dataset. The dataset we used was the MNIST LeCun and Cortes [2010] data which contains 70,000 handwritten digits, each made of 28×28 pixels.

5.1 Preparing Data

To accurately recreate the experiment to investigate if increasing the L-2 Regularization parameter improved accuracy of benign overfit models, we first needed to introduce label noise into the MNIST dataset. Recall, benign overfit models fit to noisy training data. In our case, since MNIST has no noise on its own, we needed to introduce label noise ourselves. To do so we split the dataset into test and train sets, after which we changed the correct labels to random ones for 15% of the data for the training set.

5.2 Model Architecture

Our model architecture consisted of a single hidden layer with 2000 activation units and an output layer that predicts a digit zero through nine. There were several components that influenced our model design. First, we wanted to use a single hidden layer in order to increase

the likelihood of overfitting. Additionally, we wanted to create an over-parametrized model as it is a vital component of benign overfitting Bartlett et al. [2020]. With our input size being vectors of 784 scalar values, a 2000 unit hidden layer seemed like a reasonable choice.

5.3 Testing the Models on Adversarial Data

In order to test our models, we introduced the Fast Gradient Sign Method (FGSM) attack, similar to the Projected Gradient Method that, for computational efficiency, uses only one iteration. Like the Projected Gradient Descent Method, this attack introduces small perturbations into the image that maximizes loss while being imperceptible to the human eye.

The procedure of the FGSM is as follows, let X be the input data, y the true label, and L the loss function of the model with parameters W, b. Compute the gradient of L with respect to x, $\nabla_x L(W, b, x, y)$. Then, the perturbation δ is given by $\delta = \varepsilon * sign(\nabla_x L(W, b, x, y))$, where the sign is determined element-wise. This projects onto the ℓ_∞ ball of radius ϵ. Each model with the regularization parameters [0, 0.001, 0.01, 0.1, 0.25, 1] was tested on the perturbed testing dataset of 1000 images with the FGSM attack applied.

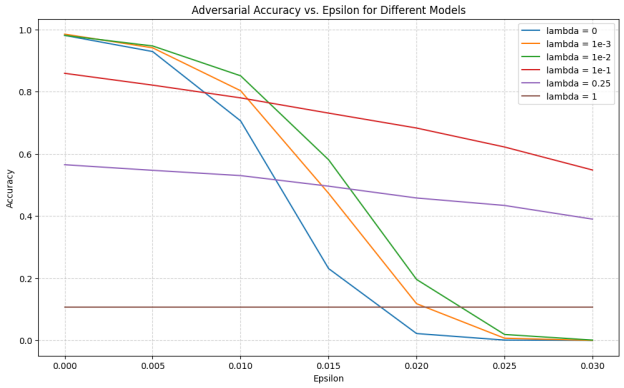


Figure 6: Accuracies of different models on adversarial examples

5.4 Takeaways

We see from the graph that the model with $\lambda = 0.1$ regularization parameter demonstrated the best adversarial robustness as ε increased despite only reaching 0.85 accuracy on the unperturbed training dataset. We also find that the models with $\lambda = 0.1$ and $\lambda = 0.25$ regularization parameters demonstrated worse accuracy on unperturbed or slightly perturbed datasets, but surpasses the models with lower λ values as the perturbation ε becomes stronger. Overall, we observe that models with nonzero regularization parameter demonstrated better adversarial accuracy across the board. This supports our findings from earlier that adversarial accuracy improves with stronger regularization, up to the point where the regularization is too strong and the model fails to make accurate predictions whatsoever.

The experimental results demonstrate that applying L2-norm regularization, particularly with larger values of λ (e.g. $\lambda = 1$), significantly reduces the degradation in model accuracy

when confronted with adversarial examples. While increasing λ generally improves robustness to adversarial perturbations, it also leads to a decrease in accuracy on the clean training data. This suggests a trade-off where improved robustness comes at the cost of some performance on the original data, and the model's ability to accurately predict unperturbed inputs is reduced.

6 Conclusion

This work set out to better understand the trade-off between benign overfitting and adversarial robustness—two phenomena that, while emblematic of modern machine learning, often seem fundamentally at odds. Benign overfitting describes a regime where heavily overparameterized models interpolate noisy training data yet still generalize surprisingly well. However, this same sensitivity to training data can leave models exposed to adversarial perturbations. Our central question was whether classical ℓ_2 regularization could serve as a bridge, preserving interpolation ability while simultaneously improving robustness to adversarial attacks.

Across both controlled synthetic settings and empirical evaluation on MNIST, our findings point to a nuanced relationship. As we increased the strength of ℓ_2 regularization, we observed a consistent trend: models became more resilient to adversarial examples, even if their clean-data accuracy declined marginally. This aligns with the intuition that regularization promotes smoother, lower-norm solutions, which are less sensitive to perturbations. However, our results also illustrate the limits of this approach. In high-dimensional regimes, benign overfitting occurs readily—with training accuracy reaching 100%—but this comes at the cost of declining test and adversarial performance. Crucially, we found that even as weight norms decreased with increasing dimension, robustness did not improve monotonically. This challenges the assumption that smaller norm alone is a reliable proxy for robustness, especially in the presence of label noise or highly expressive models.

Our MNIST experiments further support these findings. Injecting label noise to simulate the overfitting conditions of synthetic data, we trained overparameterized shallow neural networks with varying degrees of weight decay. We found that moderate levels of regularization substantially improved robustness to FGSM adversarial attacks. These models, although not achieving the highest accuracy on clean data, exhibited significantly better performance under adversarial perturbations—underscoring that a small sacrifice in clean accuracy may be necessary to gain robustness in practice.

Altogether, this study highlights the delicate balancing act required to achieve both benign overfitting and adversarial robustness. While ℓ_2 regularization provides a promising and computationally simple lever to mitigate adversarial vulnerability, it does not fully resolve the tension. High-dimensional interpolation remains inherently fragile, and robustness demands more than norm control alone. Future directions might explore hybrid approaches, such as combining regularization with margin maximization or adversarial training, to better navigate this trade-off. Ultimately, our results suggest that benign overfitting and robustness are not mutually exclusive—but achieving both requires a careful understanding of when, and how, to regularize.

Code

All code can be found at: https://drive.google.com/drive/folders/1MxfzzSDtNsTiqFHUKQcFFTWKMAT?usp=sharing.

References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression, 2020. 1, 5.2

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021. 1

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023. 1

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199. 1

Ian Goodfellow, Nicolas Papernot, Sandy Huang, Yan Duan, Pieter Abbeel, and Jack Clark. Attacking machine learning with adversarial examples. OpenAI Blog, February 2017. URL https://openai.com/index/attacking-machine-learning-with-adversarial-examples/. Accessed: 2025-06-08. 2

S. Hao and C. Zhang. The surprising harmfulness of benign overfitting for adversarial robustness, 2024. 1, 2, 2.3

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/. 5