Understanding the robustness difference between stochastic gradient descent and adaptive gradient methods

Avery Ma[1,2] Yangchen Pan[3] Amir-massoud Farahmand[1,2]

[1]University of Toronto [2]Vector Institute [3]University of Oxford

Motivation: Understanding the Robustness Difference

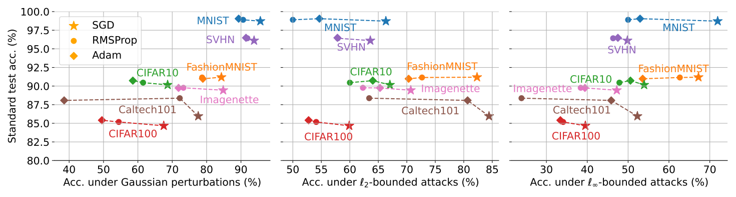


Figure 1: Comparison between models trained using SGD, Adam, and RMSProp. Models trained by different algorithms have similar standard generalization performance, but there is a distinct robustness difference as measured by the test data accuracy under Gaussian noise, ℓ_2 and ℓ_∞ bounded adversarial perturbations.[1]

[1] Croce et al., Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks

Motivation: Understanding the Robustness Difference

More on the motivating example:

1. Audio dataset.
    - Classifying short audio phrases (i.e., numbers, directions, etc.) from the Speech Commands dataset[2].
2. Both CNN and transformer architectures.

⇒ Observing the robustness differences trained by different optimizers, consistently.

[2] Warden et al., A dataset for limited-vocabulary speech recognition.

Main Claim

To optimize the standard training objective, models only need to learn **how to correctly use relevant information in the data**. Their use of irrelevant information in the data, however, is **under-constrained and can lead to solutions sensitive to perturbations**.

Observation I: Irrelevant Frequencies in Natural Signals

To optimize the standard training objective, models only need to learn how to correctly use relevant information in the data (OB I). Their use of irrelevant information in the data, however, is under-constrained and can lead to solutions sensitive to perturbations (OB II).

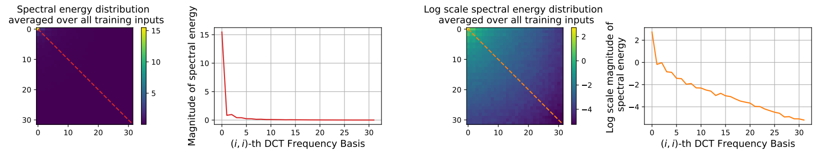
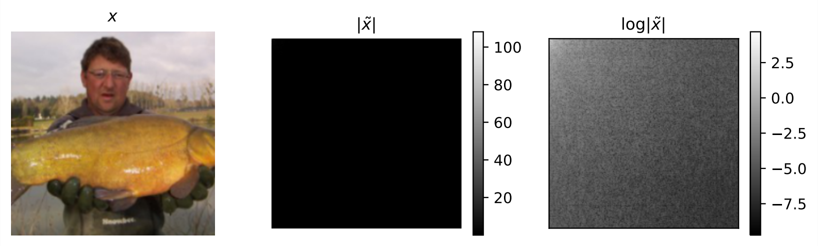


Figure 2: Illustration of the spectral energy distribution in CIFAR100. Distribution of the spectral energy heavily concentrates at low frequencies and decays exponentially towards higher frequencies.

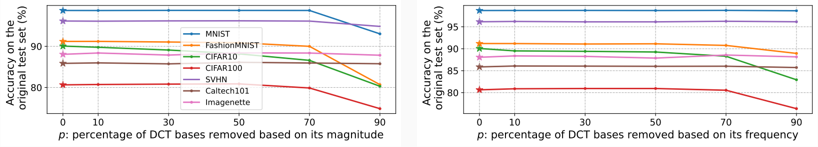
Observation I: Irrelevant Frequencies in Natural Signals



Original Image (left), DCT transformation (mid), Log-scale (right)

Observation I: Irrelevant Frequencies in Natural Signals

To optimize the standard training objective, models only need to learn how to correctly use relevant information in the data (OB I). Their use of irrelevant information in the data, however, is under-constrained and can lead to solutions sensitive to perturbations (OB II).

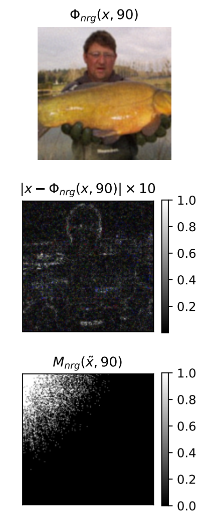


a. Parts of the signal with **low spectral energy** is irrelevant.

b. Parts of the signal with **high-frequency basis** is irrelevant.

Figure 4: Irrelevant frequencies exist in the natural data. Accuracy on the original test set remains high when the training inputs are augmented by removing parts of the signal with a) low spectrum energy and b) high frequencies.

Observation I: Irrelevant Frequencies in Natural Signals

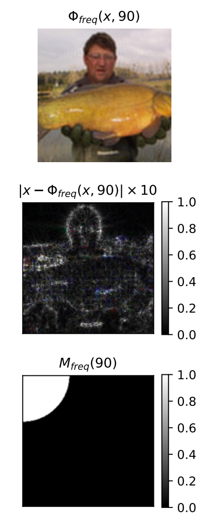


$\Phi_{nrg}(x, 90)$

$|x - \Phi_{nrg}(x, 90)| \times 10$

$M_{nrg}(\tilde{x}, 90)$

low energy removed



$\Phi_{freq}(x, 90)$

$|x - \Phi_{freq}(x, 90)| \times 10$

$M_{freq}(90)$

high freq. removed

Observation II: Model Robustness along Irrelevant Frequencies

To optimize the standard training objective, models only need to learn how to correctly use relevant information in the data (OB I). Their use of irrelevant information in the data, however, is under-constrained and can lead to solutions sensitive to perturbations (OB II).

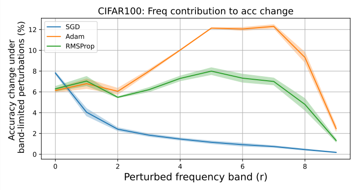


Figure 6: The effect of band-limited Gaussian perturbations on the model. Perturbations from the lowest band have a similar effect on all the models. On the other hand, models' responses vary significantly when the perturbation focuses on higher frequency bands.

⇒ This suggest that the robustness differences comes from the different responses towards perturbations along those irrelevant frequencies.

Theoretical Analysis with Linear Models

Summarizing the Linear Analysis Result

Setup: In linear regression, we compare the standard and adversarial risk of the GD and signGD solutions.

Data: A synthetic dataset that mimics the characteristics of a natural dataset: contains irrelevant frequencies.

Summarizing the Linear Analysis Result

We are interested in the standard risk:

$$\begin{aligned}\mathcal{R}_{\mathsf{s}}(w) &\triangleq \mathbb{E}\Big[\ell(X, Y; w)\Big] \\ &= \mathbb{E}\Big[|w^T X - Y|^2\Big]\end{aligned}$$

and the adversarial risk under ℓ_2 bounded perturbations:

$$\begin{aligned}\mathcal{R}_{\mathsf{a}}(w) &\triangleq \mathbb{E}\Big[\max_{||\Delta x||_2 \leq \epsilon} \ell(X + \Delta x, Y; w)\Big] \\ &= \mathbb{E}\Big[\max_{||\Delta x||_2 \leq \epsilon} |w^T(X + \Delta x) - Y|^2\Big].\end{aligned}$$

Linear Analysis: Three Main Results

1. Irrelevant information leads to multiple standard risk minimizers. For an arbitrary minimizer w^* from $\mathcal{W}^*$, we can obtain its the adversarial risk as

$$\mathcal{R}_{\mathsf{a}}(w^*) = \frac{\epsilon^2}{2}||w^*||_2^2.$$

$\Rightarrow$ Given that the linear model is a standard risk minimizer, its robustness against ℓ_2-norm bounded perturbations is proportional to its ℓ_2 norm squared.

Linear Analysis: Three Main Results

2. With a sufficiently small learning rate η, the standard risk of GD and signGD can be both close to 0.
$\Rightarrow$ Providing an explanation to the similar standard generalization performance observed on neural networks.

Linear Analysis: Three Main Results

3. Consider a three dimensional input space, we demonstrate that ratio between the two adversarial risks is always greater than 1:

$$\frac{\mathcal{R}_{\mathsf{a}}(\boldsymbol{w}^{\mathsf{signGD}})}{\mathcal{R}_{\mathsf{a}}(\boldsymbol{w}^{\mathsf{GD}})} > 1 + C,$$

where $C > 0$ and its value depends on initialization, and the magnitude of the data covariance.

$\Rightarrow$ The linear model obtained through GD is **always more robust** against ℓ_2-bounded perturbations in comparison to the model obtained from signGD.

Can we show something similar to $\mathcal{R}_{\mathbf{a}}(w^*) = \frac{\epsilon^2}{2}||w^*||_2^2$ on neural networks?

Product of Weight Norm of NN Upper-bounds its Lipschitzness

Consider the form $f(x) = (\phi_l \circ \phi_{l-1} \circ \ldots \circ \phi_1)(x)$, where each ϕ_i is a linear operation, an activation function, or pooling operation.

Denoting the Lipschitz constant of function f as $L(f)$, we can establish an upper bound[3] on the Lipschitz constant for the entire feed-forward neural network using

$$L(f) \leq \prod_{i=1}^{l} L(\phi_i).$$

The approximation to the Lipschitzness of various components of the network are typically functions of the weight norm. Examples: Linear operations: $\|W\|_p$; Skip-connections: $\|W\|_p + 1$.
$\Rightarrow$ Smaller weight norm implies smaller Lipschitz upper bound. Less vulnerable to perturbations.

[3] Any value of L satisfying the Lipschitz condition is considered a valid Lipschitz constant. For the sake of clarity, we will refer to the smallest (optimal) Lipschitz constant as L

Product of Weight Norm of NN Upper-bounds its Lipschitzness

Table 1: Comparing the upper bound on the Lipschitz constant and the averaged robust accuracy of neural networks trained by SGD, Adam, and RMSProp.

	Dataset	MNIST	Fashion	CIFAR10	CIFAR100	SVHN	Caltech101	Imagenette
$\prod_{i=1}^{l} L(\phi_i)$	SGD	**3.80**	**3.83**	**26.81**	**40.41**	**22.65**	**18.53**	**23.99**
	Adam	5.75	8.12	28.70	41.87	30.45	26.20	28.55
	RMSProp	6.21	5.11	37.75	41.71	28.31	45.84	27.11
Averaged Robust Acc.	SGD	**77.97%**	**77.95%**	**63.21%**	**55.65%**	**69.08%**	**71.42%**	**67.59%**
	Adam	65.64%	67.60%	57.71%	45.25%	65.60%	55.03%	58.86%
	RMSProp	63.54%	71.34%	56.47%	47.55%	65.37%	53.16%	57.98%

- SGD-trained neural networks have considerably smaller Lipschitz constants across all datasets.
- Explaining the better robustness to perturbations than those trained with adaptive gradient methods as shown in the motivating example.

Summary

Motivation

Large robustness difference between models trained by SGD and adaptive gradient methods, despite similar standard generalization.

Claims + empirical observations

To optimize the standard training objective, models only need to learn how to correctly use relevant information in the data. Their use of irrelevant information in the data, however, is under-constrained and can lead to solutions sensitive to perturbations.

Theoretical analysis with linear models

We show the asymptotic solution found by signGD is always more vulnerable.

Back to neural networks

We demonstrate a connection between Lipschitz upper bound and model robustness.

Practical Implications

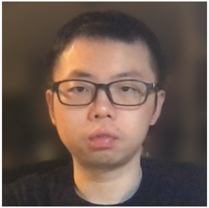
This work

- highlights the importance of optimizer selection in achieving both generalization and robustness.
- guides the development of optimization strategies that maintain high accuracy while being resilient to input perturbations.
    - Possible remedy: Initialization, Regularization (i.e., Adam/RMSProp+L2 or AdamW[4]).

[4] Loshchilov et al., Decoupled Weight Decay Regularization, ICLR 2019

Collaborators

Avery Ma



Yangchen Pan

Amir-massoud Farahmand

Thank you!