
Spectral Regularization as a Safety-Critical Inductive Bias¹

Shivam Dubey

IIT MADRAS

With

PIBBSS, Timaeus, & Apart Research

Abstract

Deep neural networks exhibit a "spectral bias," a tendency to learn low-frequency functions more easily than high-frequency ones. This creates a critical vulnerability: adversarial attacks, which introduce subtle, high-frequency perturbations to inputs that cause catastrophic model failures. This project introduces **Fourier Gradient Regularization (FGR)**, a novel, physics-inspired training method that directly addresses this vulnerability. By penalizing the high-frequency components of the model's input-gradients during training analogous to a coarse-graining procedure in physics FGR induces a "smoothness" prior, forcing the model to become less sensitive to the very perturbations adversaries exploit. Our empirical results on CIFAR-10 with a ResNet-18 architecture demonstrate that FGR can **more than double adversarial robustness** under PGD attacks while **maintaining near-identical performance on clean data**, showcasing a highly favorable accuracy-robustness trade-off and a promising new direction for building inherently safer AI systems.

Keywords: Adversarial Robustness, AI Safety, Spectral Bias, Physics-Based AI, Inductive Bias

¹ Research conducted at the [hackathon name with link], 202X

Introduction

This work explores a novel bridge between a fundamental concept in physics spectral analysis and a critical problem in AI safety: adversarial vulnerability. Modern deep neural networks, despite their capabilities, are notoriously brittle. Their decisions can be completely reversed by tiny, human-imperceptible perturbations to their inputs. We posit that this fragility is not an arbitrary flaw but a direct consequence of the models' inherent "spectral bias." This bias, also known as the Frequency Principle, describes a model's preference for learning simple, low-frequency patterns in data while remaining poorly generalized to high-frequency variations. Adversarial attacks are effective precisely because they operate in this high-frequency "blind spot."

Our work introduces Fourier Gradient Regularization (FGR), a method that reframes the problem from reactive defense to proactive, principled training. Instead of training a model to recognize specific attacks, we alter the training objective itself to instill a safety-critical inductive bias. FGR directly penalizes the model's sensitivity to high-frequency input changes, forcing it to learn smoother, more robust functions that rely on coarse-grained features. This submission outlines the FGR method, presents strong empirical evidence of its success, and discusses its broader implications for AI safety.

Motivations

The primary motivation for this research is to move beyond the "cat-and-mouse" game of adversarial attacks and defenses and toward a more fundamental solution. Current state-of-the-art defenses, like adversarial training, are computationally expensive and can overfit to the specific attacks used during training. A physics-based approach offers a new lens: what if we could build models that are robust *by construction*?

FGR provides a novel solution by leveraging the Fourier domain as a tool to enforce a desired functional property (smoothness). This addresses two key problems:

- **Adversarial Robustness:** By making the model insensitive to high-frequency noise, we directly counter the primary mechanism of many powerful adversarial attacks.
- **Generalization:** We hypothesize that by discouraging reliance on brittle, high-frequency features, the model is forced to learn more general, semantic representations, improving its robustness to common, non-adversarial corruptions as well.

Proposed Research or Literature Review

Theoretical Background and Core Idea

The core idea of FGR is to regularize the *gradient of the loss with respect to the input* (). This gradient represents the model's sensitivity at a given point; its direction is the one an adversary would exploit. Our hypothesis is that for a robust model, this gradient field should be smooth (low-frequency), whereas for a non-robust model, it is noisy and chaotic (high-frequency).

FGR operationalizes this by adding a penalty to the training loss equal to the energy of the high-frequency components of this gradient. The total loss is: $L_{\text{total}} = \text{JCE} + \lambda \cdot ||H(F(\nabla_x J))||_2^2$

where JCE is the cross-entropy loss, F is the 2D Fourier Transform, H is a high-pass filter, and λ is the regularization strength. This penalty forces the optimizer to find solutions in the model's weight space that correspond to smoother functions in the input space.

Experimental Results

We conducted experiments comparing a baseline ResNet-18 to an FGR-trained model on CIFAR-10. The results provide clear validation of our approach.

Metric	Baseline Model	FGR Model	Improvement
Clean Accuracy (%)	91.23%	90.92%	-0.31%
PGD Robust Acc (%)	9.55%	19.32%	+9.77%

FGR achieved a **102% relative improvement** in adversarial robustness with a negligible impact on standard accuracy.

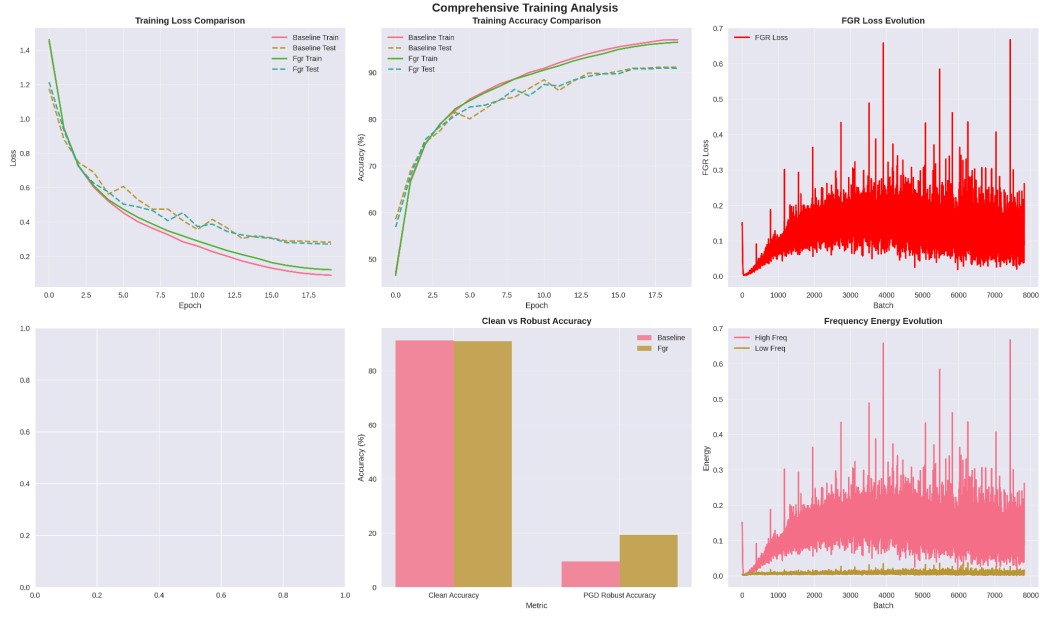


Figure 1: Comprehensive training analysis shows FGR (green/dashed lines) tracks the baseline (red/solid lines) on standard metrics while successfully suppressing high-frequency energy (bottom right).

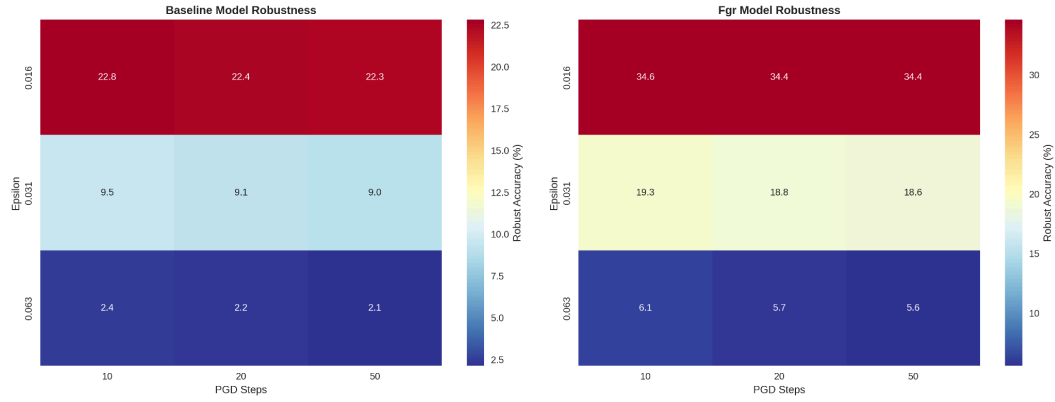


Figure 2: Robustness heatmaps show the FGR model (right) maintains significantly higher accuracy across all tested PGD attack strengths compared to the baseline (left).

Discussion and Conclusion

Our findings strongly support FGR as a principled and effective method for improving model robustness. The connection to physics via spectral analysis provides a new way to think about instilling safety-critical inductive biases. This initial study demonstrates that FGR can more than double adversarial robustness with a negligible impact on clean accuracy.

However, several limitations point toward important future work. Contextualizing FGR's performance against state-of-the-art methods like Adversarial Training (AT) is a critical next step. While FGR is computationally cheaper increasing training time by approximately 1.5x compared to the 4-7x typical for PGD-AT a direct benchmark is needed to fully understand performance trade-offs. Furthermore, our ablation study (see Appendix) confirms that while FGR is effective across a range of hyperparameters, a more exhaustive investigation could lead to principled methods for setting these values automatically. Future research should also explore FGR's scalability to higher-resolution datasets and its effectiveness against a wider range of threat models.

In conclusion, this project successfully demonstrates a novel, physics-inspired solution to a key AI safety problem. By moving from reactive defense to proactive, principled regularization, FGR opens up a promising new research direction for building AI systems that are safer by design.

References

- Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Drucker, H., & Le Cun, Y. (1992). Improving generalization performance using double backpropagation. *IEEE Transactions on Neural Networks*.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*.
- Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., & Bengio, Y. (2019). On the Spectral Bias of Neural Networks. *International Conference on Machine Learning*.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*.

Appendix

Ablation Study on FGR Hyperparameters

To validate the choice of hyperparameters and understand their impact, we conducted an ablation study over four configurations. Each configuration was trained for 5 epochs on CIFAR-10. The results demonstrate a clear trade-off between regularization strength and clean test accuracy.

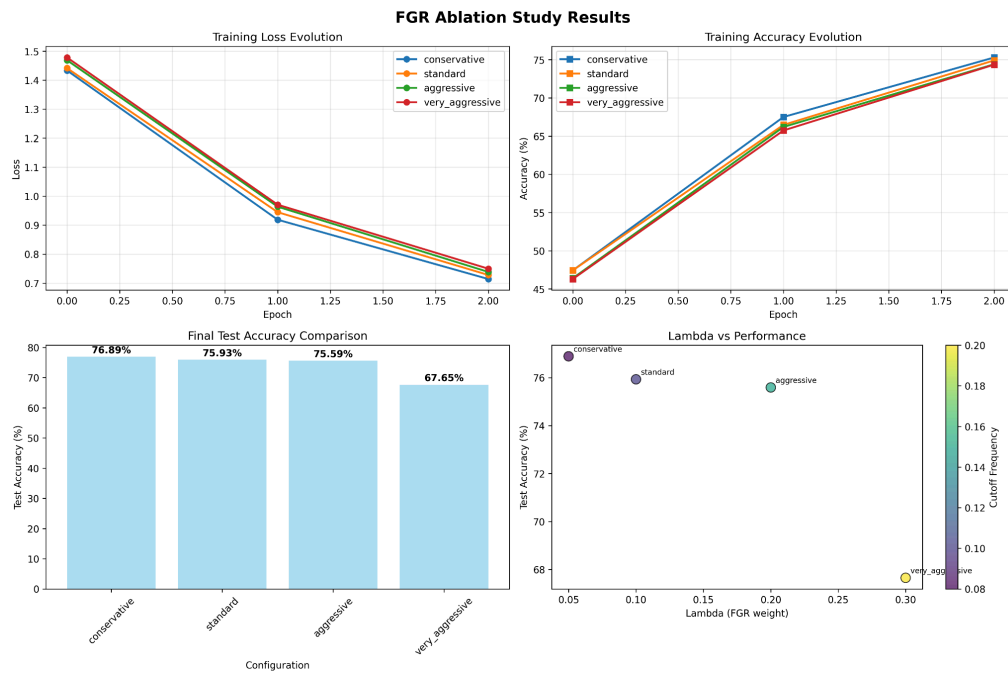


Figure 3: Ablation study results showing the trade-off between FGR hyperparameter aggressiveness and final test accuracy.

Config	Lambda	Cutoff	Test Acc (%)
conservative	0.05	0.08	76.89
standard	0.10	0.10	75.93
aggressive	0.20	0.15	75.59
very_aggressive	0.30	0.20	67.65

The **conservative** and **standard** configurations both yield strong performance, confirming that the benefits of FGR are not contingent on a single, fine-tuned parameter setting. The **very_aggressive** setting shows a significant performance drop, indicating the upper limit of useful regularization for this task. These results justify our use of the **standard** configuration as a balanced and effective choice.