Reconciling Theory and Practice for Trusted Deep Learning:
Compositional Regularization in Double Descent

Motivation: Deep learning has revolutionized prediction capabilities in a vast number of data-rich domains. But beneath their advertised success, deep neural networks are brittle, black-box systems sensitive to observation errors and changes in data distribution. These weaknesses can often lead to unexpected failures in production environments. In everyday life, the failure of deep learning systems can be a nuisance: perhaps Siri misinterprets your question, Netflix suggests a movie you don't enjoy, or Google Maps underestimates the time you spend in traffic. But in DoD applications, adversaries can exploit the weaknesses of neural networks to incapacitate drones and autonomous vehicles, disable facial recognition and network security, and send false signals to target detection and tracking systems—potentially jeopardizing American lives and infrastructure. To ensure trusted, robust deployment of deep learning, we must characterize these failure modes by developing an understanding of model generalization and performance in realistic settings.

However, our fundamental understanding of deep learning lags behind its meteoric rise in practical applications. The performance of deep models contradicts classical theory, and we lack a framework—even a heuristic—for predicting the behavior of neural networks with respect to architecture, optimization, and data distribution. One particularly thorny question is the "generalization problem": neural networks achieve low test risk despite having millions more parameters than data. According to current theory, deep learning models in this regime should overfit the data and generalize poorly, but practical experiments show otherwise[1]. To make things worse, recent work has shown that in certain cases bigger models and more data can hurt performance[5], making it impossible to characterize why or when neural networks perform well. **My research goal in graduate school is to reconcile theory and practice in deep learning by developing a foundational understanding of the generalization problem.**

Relevance to the Department of Defense: My research is highly relevant to the Army Research Office (ARO) Computing Sciences program, specifically in Information and Software Assurance via Trusted Learning for Cyber Autonomy and Intelligent Systems via Advanced Learning Theory, Methodology, and Techniques. My research also corresponds to the Air Force Office of Scientific Research (AFOSR) focus on Information and Networks via the Science of Information, Computation, Learning, and Fusion. The ARO aims to "rethink the fundamental aspects of machine learning that leads to its brittleness" and "[establish] a theoretical foundation of machine learning" which characterizes model performance under shifts in data distribution, changing environments, and adversarial attacks. The AFOSR prioritizes "understanding the underpinning of autonomy" with theory which assures "provably guaranteed performance" in high-dimensional problems. **My research connects to these programs because I will develop a theoretical foundation of generalization in deep learning which aligns with empirical observations. Such a theory will enable robust, trusted learning in uncertain environments.**

Background: The bias-variance tradeoff is a central tenet of statistical learning theory which states that overparameterized models are likely to overfit the data; contrary to this notion, modern neural networks exactly fit (*i.e.*, interpolate) the training data yet achieve low test risk. Curiously, models which barely achieve interpolation have high test risk, but better results are attained as model size or epochs increase[1,5]. This leads to unexpected results in practice because model size and training performance are not indicative of generalization; thus, it is impossible for engineers to confidently predict the accuracy of a certain model in general settings. A similar

paradox has been shown to occur in adversarial settings—in particular, an increase in training set size may increase the generalization error of adversarially robust models[3].

This phenomenon of non-monotonicity has been termed "double descent" (Figure 1), and exists in linear regression, decision trees, and deep learning[1,5]. In fact, the generalization curve can have an arbitrary number of peaks, but we rarely observe double descent in practice, let alone triple or higher[2]. To understand why, recent work has shown that optimally-tuned ℓ_2 regularization can mitigate double descent by inducing model-size and sample-wise monotonicity[6]. However, realistic models (*i.e.*, industry-standard deep neural networks which are overparameterized even when trained on massive datasets) still manage to avoid this behavior even without carefully tuned regularization[6]. Thus, **generalization curves are thought to arise from interactions between properties of the data and inductive biases of the model, whose exact characterization is an open question**[2]. Indeed, practical neural networks often combine model-dependent (*e.g.*, ℓ_1 or ℓ_2 penalties) and data-dependent (*e.g.*, batch normalization or dropout) regularization techniques to achieve lower test risk than either alone[7]. I call this synthesis *compositional regularization*, and I hypothesize that it helps avoid double descent in practice.

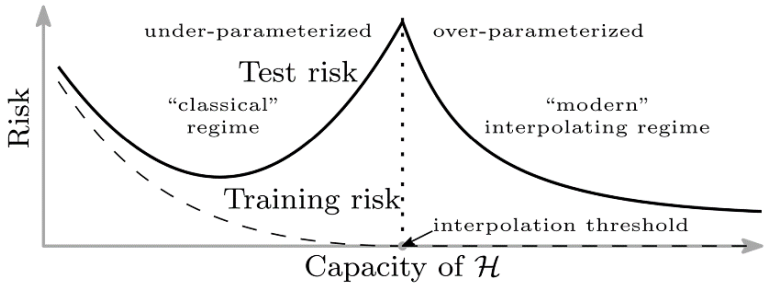

Figure 1: The "double descent" generalization curve. On the left is the classical bias-variance tradeoff, with a peak when models reach interpolation. On the right is the modern interpolating regime, wherein models with zero empirical risk achieve low test risk. Diagram from [1].

Proposal: Understanding why double descent does not occur in practice is the first step in fully characterizing the generalization of deep neural networks. A framework that explains this phenomenon would provide theoretically principled techniques for assessing generalization based on model- and data-dependent factors, a foundational contribution to predicting the performance and robustness of deep models. To understand why realistic models do not exhibit double descent, **I propose a joint theoretical and empirical investigation** to answer the following research question: *Does compositional regularization mitigate double descent in overparameterized neural networks by inducing monotonicity without careful tuning?* First, I will analyze whether data-dependent regularization must be optimally tuned to induce monotonicity. Then, I will rigorously determine whether compositional regularization is a consistently beneficial technique with respect to generalization. Finally, I will develop a theoretical framework for compositional regularization in simple nonlinear models and empirically probe the performance of different combinations and tunings of regularization techniques in general settings, including under distribution shift.

Methods: On the theoretical side, I will begin with a **characterization of data-dependent regularization** in nonlinear models. I will extend recent work on theory of dropout and batch normalization to analysis of monotonicity. In particular, I will seek to understand if the distribution-free model perturbation framework[4] (for dropout) or Lipschitz augmentation strategy[8] (for batch normalization) implies asymptotic model-size or sample-wise monotonicity in simple feedforward and convolutional neural networks. By varying model tuning parameters (*e.g.*, dropout probability) and data distribution, I will obtain different guarantees on model monotonicity. I will also analyze the performance of data-dependent regularization under distribution shift by introducing a noise parameter. I hypothesize that data-dependent regularization strategies will require optimal tuning with respect to the input data to imply

monotonicity. I predict that this weakness will cause such techniques to fail to mitigate double descent under distribution shift, because the optimal tuning differs between training and testing.

Next, I will develop a **theoretical framework for analyzing compositional regularization**, combining the model-dependent monotonicity framework[6] with my analysis of data-dependent regularization. I will begin with a simple linear model and progress to more general scenarios if my analysis holds. I hypothesize that the effects of compositional regularization will sufficiently constrain the weight, hidden layer, and interlayer Jacobian norms such that monotonicity is induced without optimal tuning (though it is likely that *some* amount of tuning is needed). Similar to the data-dependent case, I will evaluate performance under distribution shift via a noise parameter. I predict that compositional regularization will outperform data-dependent regularization in this case by implying monotonicity when the distribution shift is not too excessive. This would explain why deep neural networks do not experience double descent in realistic settings even when the test distribution is not perfectly captured by the training set.

In my empirical study, I will first investigate **whether compositional regularization universally lowers test risk** of realistic models. This is important to establish the validity of compositional regularization as a hypothesis for avoiding double descent in practice. I will compare no, single, and compositional regularization on overparameterized models with various inductive biases including feedforward networks, ResNets, and Transformers on industry-standard datasets such as CIFAR-10/100, TIMIT, and WMT ’14 English-French. While minor experiments have been performed[7], this will be (to the best of my knowledge) the first massive-scale empirical analysis of compositional regularization in deep learning. I hypothesize that compositional regularization will consistently outperform single regularization and experience diminishing returns as the number of applied techniques increases. I will determine empirical results by plotting model size against epoch, and both quantities against test risk on held-out datasets[5].

Then, I will perform an **ablation study of different compositions** (*e.g.*, of ℓ_2 penalties, batch normalization, and dropout) with various amounts of tuning to isolate the effects of each technique and identify their contribution to the generalization of each model. This experiment has two aims: it will both assess the impact of tuning on compositional regularization effectiveness and provide insight into the interactions between model- and data-dependent regularization. I will also repeat these experiments with injected noise in the test datasets to simulate distribution shift, evaluating the effect of different compositions on model robustness. I hypothesize that optimal model-dependent regularization and optimal data-dependent regularization alone each induce monotonicity, but (as in realistic settings) their composition need not be optimally tuned.

Qualifications and Resources: I am uniquely positioned to perform this research because of my significant background in both the theory and practice of machine learning. At USC, I studied the theoretical and empirical **impact of regularization techniques on generalization** of interpolating models, and I am currently investigating fundamental questions in the optimization theory of linear functions. My deep learning internships at the Air Force Research Laboratory, Sandia National Laboratories, and Google X have resulted in **two submitted publications** with a third in preparation, and my code has been transitioned to the official TensorFlow GitHub repository. I also joined a Google Brain group focused on generalization phenomena in deep learning. I would most like to conduct the proposed research in collaboration with **Prof. Tengyu Ma at Stanford University**, whose theoretical expertise in data-dependent regularization double descent would accelerate my research. I also value Stanford’s **proximity to industry and government research laboratories**, as my computationally-intensive research would greatly benefit from partnerships with my colleagues at the Air Force, Sandia, and Google.

References

[1] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. “Reconciling modern machine learning practice and the bias-variance tradeoff”. PNAS 2019.
[2] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. “Multiple descent: design your own generalization curve”. arXiv 2020.
[3] Yifei Min, Lin Chen, and Amin Karbasi. “The curious case of adversarially robust models: more data can help, double descend, or hurt generalization”. arXiv 2020.
[4] Wenlong Mou, Yuchen Zhou, Jun Gao, and Liwei Wang. “Dropout training, data-dependent regularization, and generalization bounds”. ICML 2018.
[5] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. “Deep double descent: where bigger models and more data hurt”. ICLR 2020.
[6] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. “Optimal regularization can mitigate double descent”. arXiv 2020.
[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. JMLR 2014.
[8] Colin Wei and Tengyu Ma. “Data-dependent sample complexity of deep neural networks via Lipschitz augmentation”. NeurIPS 2019.