Reconciling Theory & Practice in Deep Learning: Compositional Regularization in Double Descent

Keywords: generalization, regularization, overparameterization

Research Goal: 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. **My research goal is to reconcile theory and practice in deep learning by performing mathematical investigation and empirical analysis in tandem.** I aim to uncover universal phenomena that challenge conventional knowledge, then develop theory to explain and predict these behaviors. I will advance the *science* of deep learning in a way beneficial to theorists, engineers, and the public. One aspect of this research is the "generalization problem": neural networks achieve low test risk despite having millions more parameters than data. This proposal describes my novel approach to the generalization problem based on interactions between regularization techniques; my research will advance scientific understanding of deep learning and translate to measurable impact for practitioners and end users.

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[1,4]. This phenomenon of non-monotonicity has been termed "double descent" (Figure 1), and exists in linear regression, decision trees, and deep learning[1,4]. 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]. Optimally-tuned ℓ_2 regularization can mitigate double descent by inducing model-size and sample-wise monotonicity; 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[5]. 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*.

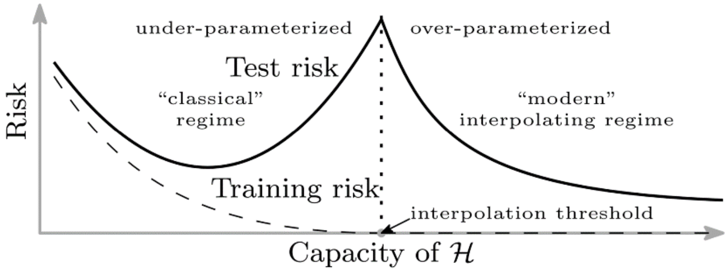

Figure 1: The "double descent" generalization curve. On the left is the classical bias-variance tradeoff, with a peak when models reach interpolation. On the right is the modern interpolating regime, wherein models with zero empirical risk achieve low test risk. Diagram from [1].

Proposal: 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 induce model-size or sample-wise monotonicity in overparameterized neural networks 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.

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[3] (for dropout) or Lipschitz augmentation strategy[8] (for batch normalization) implies asymptotic model-size or sample-wise monotonicity in two-layer feedforward neural networks. I hypothesize that data-dependent regularization strategies will require optimal tuning (*e.g.,* of dropout probability) to imply monotonicity. Then, I will develop a **theoretical framework for analyzing**

compositional regularization, combining the model-dependent monotonicity framework[5] 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. I will measure the results by the theoretical guarantees I obtain.

In my empirical study, I will first investigate **whether compositional regularization universally lowers test risk** of realistic models. I will evaluate several 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. I hypothesize that compositional regularization will consistently outperform single regularization and experience diminishing returns as the number of applied techniques increases. 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. I hypothesize that optimal model-dependent regularization and optimal data-dependent regularization alone each induce model-size and sample-wise monotonicity, but their composition need not be optimally tuned. I will determine the results by plotting model size against epoch, and both quantities against test risk on held-out datasets[5].

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 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 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 value Stanford's **proximity to industry and government research laboratories**, as my computationally-intensive research would greatly benefit from partnerships with my colleagues at Sandia and Google X.

Intellectual Merit: My proposal contributes to **advancing the fundamental understanding of generalization in deep learning**, with the potential to reconcile classical statistical learning theory with modern regimes. This would represent a significant step towards a **cohesive theory of deep learning** which includes a characterization of training and generalization behavior. My research would enable future work in developing compositional regularization techniques, such as meta-learning for optimizing regularization. Other disciplines would benefit from the augmented training stability of large models, increasing the use of machine learning as a tool for discovery in areas including physical simulations and drug design.

Broader Impacts: The generalization problem is a fundamental question in machine learning which affects any application context where accurate inference is desired. My proposal will increase understanding of double descent and enable practitioners to predict model behavior in overparameterized regimes. This will result in a more principled application of deep learning, **increasing the efficiency, accuracy, and robustness of practical neural networks.** A major concern is whether models remain fair and equitable when deployed at scale. Interpolating models can learn spurious correlations which harm minority groups, but recent work established that regularization has potential to mitigate this[6]. My proposal will develop theory and application of compositional regularization, generating principled techniques which could **provably reduce spurious correlations and improve test performance on minority groups.**

References: **[1]** Belkin *et al.* "Reconciling modern machine learning practice and the bias-variance tradeoff". PNAS 2019. **[2]** Chen *et al.* "Multiple descent: design your own generalization curve". arXiv 2020. **[3]** Mou *et al.* "Dropout training, data-dependent regularization, and generalization bounds". ICML 2018. **[4]** Nakkiran *et al.* "Deep double descent: where bigger models and more data hurt". ICLR 2020. **[5]** Nakkiran *et al*. "Optimal regularization can mitigate double descent". arXiv 2020. **[6]** Sagawa *et al.* "An investigation of why overparameterization exacerbates spurious correlations". ICML 2020. **[7]** Srivastava *et al.* "Dropout: a simple way to prevent neural networks from overfitting". JMLR 2014. **[8]** Wei and Ma. "Data-dependent sample complexity of deep neural networks via Lipschitz augmentation". NeurIPS 2019.