Explaining and Predicting Fine-Tuning in Large Models via Linearization

Parameter-Efficient Fine-Tuning (PEFT) is a family of methods developed to address the computational burden of adapting large language models (LLMs). These methods cut down on training costs by updating only certain layers or using low-rank adjustments, while keeping the model effective at adapting to new tasks. Despite their practical success, the underlying dynamics that explain why certain design choices work remain poorly understood, which makes principled exploration of the algorithmic space challenging. In this work, we study PEFT through the lens of linearization. We observe that fine-tuned models are often implicitly encouraged to remain close to their pretrained initialization, and by making this proximity explicit through an l_2-regularization in parameter space, we show that fine-tuning dynamics can be characterized as learning with the positive-definite neural tangent kernel (NTK).

In this paper, we introduce *linearized* fine-tuning to understand better how large models adapt, framing the process through Neural Tangent Kernel (NTK) regression. Adding linearization reduces fine-tuning to a form of NTK regression, which lets us anticipate how different design choices will affect performance by looking at the properties of the kernel. This perspective is not overly restrictive, since most parameter-efficient methods already encourage the fine-tuned model to stay close to its pretrained state. However, what has been missing in prior works is a way to measure how well the fine-tuning dynamics align with this linear approximation. Earlier approaches primarily relied on the assumption that models remain close enough to initialization for a Taylor expansion around the pretrained parameters to hold. Motivated by this, we introduce an explicit inductive bias that allows us to quantify how much linearity is preserved during fine-tuning, and we provide a theoretical upper bound on the distance between the fine-tuned model and its linearized counterpart. This bound, in turn, justifies using linearization to predict fine-tuning performance. Also, earlier theory has shown that the NTK remains constant during training for sufficiently wide networks, a regime referred to as *lazy* training, which further supports our use of this framework.

We demonstrate that fine-tuning large language models is not always well captured by a linearized approximation. However, we induce a regime of lazy training by introducing an explicit inductive bias in the form of a weight decay toward the original pretrained parameters. In this setting, the fine-tuning process can be approximated by the linearized model. During training, the model's behavior can be described by the neural tangent kernel (NTK), which stays unchanged as optimization progresses. This stability makes it possible to study how well a fine-tuned model will generalize by looking directly at the spectral properties of the NTK. Building on this idea, we show that adding an explicit bias toward the pretrained model naturally produces a linearized form of fine-tuning, keeping the final solution close to its linear approximation. Extending this view, we recast fine-tuning as NTK regression and use the eigenvalue decomposition of the kernel to derive predictive bounds on the empirical risk of fine-tuning outcomes prior to training. We further provide new spectral perturbation results that characterize how the NTK changes when different parameter subsets are chosen for adaptation; to our knowledge, this is the first analysis of layer selection through NTK spectrum, offering principled guidance for PEFT design, saving valuable time and computational resources.

Finally, we validate these theoretical insights through extensive Low-Rank Adaptation (LoRA) experiments. In particular, we show that the NTK condition number at initialization is a reliable predictor of downstream performance, enabling practitioners to anticipate fine-tuning effectiveness before training even begins. The experiments clearly demonstrate an inverse relationship between the NTK condition number and the model's evaluation accuracy. This finding suggests that by examining the NTK matrix prior to training, we can identify tasks that are well-conditioned, where a lower condition number corresponds to lower training and evaluation loss. While our empirical focus is on LoRA, the tools we introduce broadly apply to other PEFT approaches.

Although our theory is developed in the setting of squared loss and gradient descent, our experiments show that the insights also carry over to more practical cases, such as cross-entropy loss and optimizers like AdamW. While our analysis focuses on explicitly regularized fine-tuning, extending it to standard, unregularized methods is an important and promising direction for future work. Our results highlight linearization as a powerful way to understand and improve PEFT, offering theoretical foundations and practical tools for making LLM adaptation more efficient and effective.