4.10 Theory

Authors: Aditi Raghunathan, Sang Michael Xie, Ananya Kumar, Niladri Chatterji, Rohan Taori, Tatsunori Hashimoto, Tengyu Ma

Rigorous mathematical theory plays a foundational role in many engineering and science disciplines (e.g., information theory in electrical engineering). We believe that theory of foundation models can be particularly beneficial in guiding technical decisions and innovations because of the huge computational costs associated with experimenting on foundation models. In addition, theoretical insights help elucidate fundamental limitations and explain surprising empirical phenomena. However, the community currently has a limited theoretical understanding of foundation models, despite much recent progress [Arora et al. 2019b; HaoChen et al. 2021a; Wei et al. 2021, 2020b; Zhang and Hashimoto 2021; Saunshi et al. 2020b; Dao et al. 2019; Tosh et al. 2020, 2021; Cai et al. 2021; Lee et al. 2020a; Zimmermann et al. 2021; Bansal et al. 2020; Wang and Isola 2020; Tsai et al. 2020; Tian et al. 2020a,b; Tripuraneni et al. 2020; Du et al. 2020].

Deep neural networks form the backbone of foundation models. Even in the well-studied supervised learning setting, where the train and test scenarios have the same distribution, there are numerous open questions around deep nets such as understanding non-convex optimization, the implicit regularization effect of optimizers, and expressivity. Foundation models raise questions that significantly go beyond the supervised deep learning setting. The core problem in theoretically analyzing foundation models is understanding why training on one distribution with a possibly unsupervised/self-supervised loss leads to good adaptation performance on *different* downstream distributions and tasks.[87]

We will discuss an intuitive modularization to analyze foundation models that lays bare the connections between supervised learning and foundation models, concrete and core technical questions, and some promising theoretical tools to address these questions. These new core questions can provide useful insight into foundation models and can be studied in parallel to supervised deep learning theory. While we focus on analyzing the downstream performance, the proposed modularization and tools could prove useful to analyze other metrics of interest such as robustness to distribution shifts (§4.8: ROBUSTNESS) and security (§4.7: SECURITY).

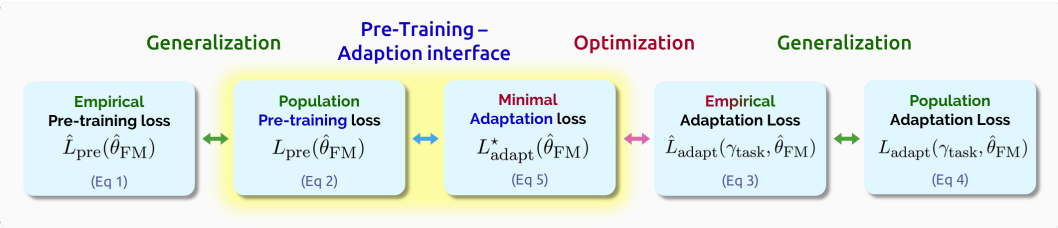

Fig. 22. The analysis of foundation models from pretraining on diverse data to downstream performance on adapted tasks involves capturing the relation between different loss terms as shown above. The main challenge is to analyze the highlighted pretraining-adaptation interface which requires reasoning carefully about the population losses in addition to the model architecture, losses and data distributions of the pretraining and adaptation stages (§4.10.2: THEORY-INTERFACE). Analysis of generalization and optimization largely reduces to their analysis in standard supervised learning.

[87] The theory for foundation models closely relates to, but also goes beyond the theory for transfer learning (which is itself an underexplored area): foundation models are possibly trained with unlabeled data and will be adapted to many or all natural tasks, whereas transfer learning typically studies labeled source tasks and a fixed number of target tasks.

4.10.1 *Theoretical formulations and modularizations.*

Recall that foundation models are trained on a large amount of raw data (§4.2: TRAINING) then adapted to specific tasks (§4.3: ADAPTATION) and therefore can be decomposed naturally into training and adaptation phases. We identify interfaces between them and disentangle parts specific to foundation models from parts that require standard deep learning theory, so that they can be independently worked on. We introduce a modularized analysis framework, which has also been implicitly or explicitly employed in recent works, e.g., Arora et al. [2019b]; HaoChen et al. [2021a]; Wei et al. [2020b]; Tripuraneni et al. [2020]. The crucial component in this modularized analysis turns out to be the *pretrain-adaptation interface*. We first describe the modularization, and discuss why we find this modularization promising and finally some limitations.

We will refer to the training phase explicitly as "pretraining" to distinguish it from the adaptation phase that could also involve training on a few samples from a particular task.

Pretraining phase. The pretraining of foundation models often involves a data distribution p_{pre} (e.g., the distribution of natural text) and a *pretraining loss* function $\ell_{\text{pre}}(x;\theta)$ that measures the loss (e.g., language modeling loss in GPT-3) on an input x for a model with parameters $\theta \in \Theta$. Let $\hat{p}_{\text{pre}}$ denote the empirical distribution over a large number of independent samples from p_{pre}.

Pretraining minimizes the loss ℓ_{pre} on $\hat{p}_{\text{pre}}$, which we call *the empirical pretraining loss*, and produces a model $\hat{\theta}_{\text{FM}}$:

$$\widehat{L}_{\text{pre}}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{x\sim\hat{p}_{\text{pre}}}[\ell_{\text{pre}}(x;\theta)], \text{ and } \hat{\theta}_{\text{FM}} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \widehat{L}_{\text{pre}}(\theta). \tag{1}$$

We consider the corresponding loss on the population distribution p_{pre}, called the *population pretraining loss*, as a central concept:

$$L_{\text{pre}}(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{x\sim p_{\text{pre}}}[\ell_{\text{pre}}(x;\theta)]. \tag{2}$$

Optimization-based adaptation phase. We frame adaptation as a general constrained optimization problem that depends on $\hat{\theta}_{\text{FM}}$, abstracting away those adaptation methods that are based on optimizing certain loss functions such as fine-tuning and prompt-tuning (see, e.g., [Houlsby et al. 2019; Li and Liang 2021; Lester et al. 2021], and §4.3: ADAPTATION).

Since different adaptation methods could modify different subsets of the model parameters, we denote the space of adapted model parameters by some Γ. Given a downstream task distribution p_{task} (e.g., question answering in a particular domain) and a few empirical samples $\hat{p}_{\text{task}}$ sampled from p_{task}, we model the adaptation phase as minimizing some *adaptation loss* ℓ_{adapt} on $\hat{p}_{\text{task}}$ w.r.t adapted parameters $\gamma \in \Gamma$:

$$\gamma_{\text{task}}(\hat{\theta}_{\text{FM}}) \stackrel{\text{def}}{=} \arg\min_{\gamma\in\Gamma, C(\gamma;\hat{\theta}_{\text{FM}})\le c_0} \widehat{L}_{\text{adapt}}(\gamma, \hat{\theta}_{\text{FM}}), \tag{3}$$

where $\widehat{L}_{\text{adapt}}(\gamma, \hat{\theta}_{\text{FM}}) \stackrel{\text{def}}{=} \mathbb{E}_{x\sim\hat{p}_{\text{task}}}[\ell_{\text{adapt}}(x;\gamma,\hat{\theta}_{\text{FM}})]$ is the empirical adaptation loss, and $C(\gamma,\hat{\theta}_{\text{FM}}) \le c_0$ is an optional constraint that controls the complexity of the adapted parameters, encompassing both explicit regularization (e.g., model dimensionality and norm) and the implicit regularization of the adaptation process.

We list some common adaptation methods and discuss the corresponding adapted parameter γ and constraints $C(\gamma,\hat{\theta}_{\text{FM}}) \le c_0$.

(1) Linear probing: training a linear classifier on top of the representations from a foundation model. Here $\Gamma = \mathbb{R}^k$ is the set of linear classifiers on the representations of dimensionality k, and $C(\gamma,\hat{\theta}_{\text{FM}})$ could be the ℓ_2 or ℓ_1 norm of γ.

(2) Fine-tuning: optimizing a randomly initialized linear head for a few steps, and all other parameters θ from the initialization of $\hat{\theta}_{\text{FM}}$. Here γ is the concatenation of θ and the linear head. Such a process could correspond to some implicit regularization of γ towards the initialization $\hat{\theta}_{\text{FM}}$ captured by $C(\gamma, \hat{\theta}_{\text{FM}}) \leq c_0$. The exact term $C(\gamma, \hat{\theta}_{\text{FM}})$ would depend on the optimization algorithm used, and such a characterization of the implicit regularization of optimization is an area of active research study [e.g., Gunasekar et al. 2017; Soudry et al. 2018; Gunasekar et al. 2018; Arora et al. 2019a; Blanc et al. 2019; Woodworth et al. 2020; Wei et al. 2020a; HaoChen et al. 2021b; Damian et al. 2021; Kumar et al. 2022, and references therein].[88]

(3) Prompt-tuning: optimizing a small set of continuous task-specific vectors that prepend the task inputs. Here γ is the continuous prompt vectors which often has small dimensionality, and we may optionally have a constraint on the norms of γ.

One obvious limitation to note is that this formulation excludes adaptation methods such as in-context learning [Brown et al. 2020] where there is no "training" (i.e., the minimization of some empirical adaptation loss) during the adaptation phase. We discuss this and other limitations in §4.10.3: THEORY-INCONTEXT.

Two central quantities for the adaptation phase are the *population adaptation loss*

$$L_{\text{adapt}}(\gamma, \hat{\theta}_{\text{FM}}) = \mathbb{E}_{x \sim p_{\text{task}}}[\ell_{\text{adapt}}(x; \gamma, \hat{\theta}_{\text{FM}})] \tag{4}$$

and the *minimal adaptation loss*

$$L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}}) = \min_{\gamma \in \Gamma, C(\gamma; \hat{\theta}_{\text{FM}}) \leq c_0} L_{\text{adapt}}(\gamma, \hat{\theta}_{\text{FM}}) \tag{5}$$

Separate analysis for modularized phases. Existing generalization theory for standard supervised learning aims to show that $\widehat{L}_{\text{pre}} \approx L_{\text{pre}}$ and $\widehat{L}_{\text{adapt}} \approx L_{\text{adapt}}$. Addressing these questions specifically for deep nets is an active research area. We can also leverage the standard learning theory decomposition to bound the final downstream task loss by the excess generalization error and the minimal adaptation loss as follows.

$$L_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}}) \leq \underbrace{L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}})}_{\text{minimal adaptation loss}} + \text{generalization error} \tag{6}$$

where the generalization error captures the closeness between L_{adapt} and $\widehat{L}_{\text{adapt}}$.[89] The decomposition and relationship between these key quantities are shown in Figure 22. The generalization and optimization arrows, as argued above, largely reduce to deep learning theory in the supervised setting. What we are left with is the main challenge with foundation models, which is to understand why the minimal adaptation loss $L^{*}_{\text{adapt}}(\hat{\theta}_{\text{FM}})$ can be small as a result of a small pretraining population loss, which study in §4.10.2: THEORY-INTERFACE.

The work of Arora et al. [2019b] pioneered the pursuit of this question by bounding from above $L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}})$ by $L_{\text{pre}}(\hat{\theta}_{\text{FM}})$ in the context of contrastive learning, and HaoChen et al. [2021a]; Tosh et al. [2020, 2021] relax the data assumptions. Other pretraining methods successfully analyzed

[88] It may not always be feasible to characterize the inductive bias of adaptation via an explicit constraint $C(\gamma, \hat{\theta}_{\text{FM}}) \leq c_0$. The modularization we propose is also applicable in these cases, but for notational simplicity, we focus on the case where implicit regularization can be approximated via an explicit constraint.

[89] More precisely, the generalization error term is the sum of $L_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}}) - \widehat{L}_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}})$ and $\widehat{L}_{\text{adapt}}(\gamma^{\star}_{\text{task}}, \hat{\theta}_{\text{FM}}) - L_{\text{adapt}}(\gamma^{\star}, \hat{\theta}_{\text{FM}}) = \widehat{L}_{\text{adapt}}(\gamma^{\star}_{\text{task}}, \hat{\theta}_{\text{FM}}) - L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}})$, where $\gamma^{\star}_{\text{task}}$ is the minimizer of (5). (6) follows easily be using $\widehat{L}_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}}) \leq \widehat{L}_{\text{adapt}}(\gamma^{\star}_{\text{task}}, \hat{\theta}_{\text{FM}})$.

under this framework (implicitly or explicitly) include pretraining with language models [Wei et al. 2021] or self-supervision [Lee et al. 2020a], with self-training algorithms [Wei et al. 2020b; Cai et al. 2021], and with multiple supervised tasks [Tripuraneni et al. 2020; Du et al. 2020].

4.10.2 Why is the pretraining-adaptation interface interesting?.

As shown in Figure 22, the main missing link beyond standard supervised theory is:

Under what conditions does a small population pretraining loss $L_{pre}(\hat{\theta}_{\text{FM}})$ *imply a small minimal adaptation loss* $L^{\star}_{adapt}(\hat{\theta}_{\text{FM}})$ and why?

The conditions that lead to a successful interface could depend on several quantities such as the pretraining and adaptation distributions, objectives and training methods, as well as the model architecture. This question is beyond the scope of standard generalization theory, but it does narrow us down to a few important factors specific to foundation models, and captures the essence of various important open questions on foundation models as we argue below.

First, we note that this interface deals with population quantities that concern two *different distributions*. Hence, the conditions for a successful interface are likely to involve special properties of the distributions, for example, the diversity of the pretraining distribution and structural shifts between the pretraining and adaptation data. This makes the analysis of the interface challenging (as discussed below in §4.10.4: THEORY-TOOLS) as we need to make careful modeling assumptions about how the two distributions relate to one another. However, this presents the possibility that tools and techniques developed to analyze such interfaces could be useful to understand the effect of distribution shifts and to predict when foundation models can improve robustness.

Second, the population losses and possibly the conditions of a successful interface depend on the *model architecture*. This raises the challenge of opening up the black-box of the neural nets. What does a small pretraining loss on a particular distribution tell us about the properties of the intermediate layers? Such analyses would also guide us in designing new adaptation methods that more carefully exploit different intermediate representations.

Third, *few-shot learning* or the sample efficiency of adaptation can be captured through the constraint on the complexity measure $C(\gamma, \hat{\theta}_{\text{FM}}) < c_0$ in the minimal adaptation loss. We need to formally characterize these complexity measures (e.g., by understanding the implicit regularization effect of the adaptation process) and further understand why a small population pretraining loss would imply a low-complexity adaptation parameters γ_{task}. A satisfactory answer to this question would likely allow us to improve the sample-efficiency of downstream adaptation.

Finally, and importantly, critical components of the interface are the choice of the *pretraining and adaptation losses*. We want to understand how to best combine the pretraining and adaptation objectives for successful adaptation. It is possible that the pretraining objective that best guarantees successful adaptation differs from what is explicitly minimized during the pretraining process — the interface above allows one to use any surrogate population objective on the pretraining distribution. In addition, new surrogate objectives that provably lead to good adaptation across a broad set of tasks could shed light on the fundamental aspects that make foundation models successful.

To summarize, the interface precludes the issue of generalization and allows us to formally reason about the interaction between several important quantities of the pretraining and adaptation phases that can guide practice in important ways.

4.10.3 Challenge: analysis of in-context learning and other emergent behavior.

GPT-3 [Brown et al. 2020] demonstrates the power of in-context learning, an adaptation method that does not need any parameter optimization. In the adaptation phase, the pretrained language foundation model takes in a prompt — a sequence of tokens that concatenates input-output examples

from the task — followed by a test example and simply generates the label of the test example by conditioning on the sequence seen thus far (prompt plus test example). In other words, there is no explicit training or change to the model parameters. What is the mechanism by which the model "learns" from the different examples by simply executing with the examples as inputs? The previous modularization does not directly apply because we do not obtain new model parameters during adaptation, but rather we only use the generative capabilities of the foundation model by executing on structurally-designed inputs. However, the idea of separating pretraining with infinite data and pretraining with finite data can still be useful. For example, a recent work starts with the assumption of infinite pretraining data and sufficient model expressivity to study in-context learning [Xie et al. 2021c]. These assumptions reduce the characterization of in-context learning to a matter of analyzing the pretraining distribution conditioned on in-context learning prompts, which are drawn from a different distribution than the pretraining data. In particular, Xie et al. [2021c] proposes that in-context learning emerges from long-term coherence structure in the pretraining distribution, which is described by a latent variable generative model with coherence structure. More broadly, while the modularization proposed in this section provides a nice framework to gain useful theoretical insights into foundation models, it is possible that some emergent behavior like in-context learning and other capabilities yet to be discovered would require going beyond the modularization, e.g., by opening the black box of the architecture.

4.10.4 Challenge: appropriate data assumptions and mathematical tools.

Understanding the interface between pretraining and adaptation phases requires a more careful study of data distributions than in traditional supervised learning. This is because the pretraining and task adaptation distributions are inherently different. By definition, foundation models are trained on raw data that is typically extremely diverse and task-agnostic, while the adaptation data depends heavily on the task. Similarly, in-context learning emerges as a result of learning to generate data that looks like the pretraining distribution, and thereby understanding in-context learning requires careful modeling of the pretraining data. Hence answering the central questions around foundation models requires realistic and interpretable assumptions that are also amenable to analysis. Recent works either assume certain properties of the population data, e.g., the expansion property in HaoChen et al. [2021a]; Wei et al. [2020b], or that the population data is generated from latent variable models with some structure [Saunshi et al. 2020a; Wei et al. 2021; Arora et al. 2016; Lee et al. 2020a; Zhang and Hashimoto 2020; Tosh et al. 2021].

We generally lack mathematical tools for relating properties of foundation models to the structure in the population data distribution. HaoChen et al. [2021a] applies spectral graph theory to leverage the inner-class connectivity in the population distribution. More precise characterization of $\hat{\theta}_{\text{FM}}$ via probabilistic and analytical derivations is possible for latent variable models, but so far restricted to relatively simple ones. The community will significantly benefit from more systematic and general mathematical tools to address this question.

It is also highly desirable to define simple toy cases so that theoreticians can precisely compare the strengths of various tools and analyses. For example, HaoChen et al. [2021a] and Wei et al. [2020b] consider the mixture of manifolds problem which might potentially be a good simplified test bed for vision applications. We need more interesting test beds for discrete domains such as NLP. We believe that tractable theoretical models which capture relevant properties of real datasets are a crucial step towards placing foundation models on solid theoretical footing.