PRE-TRAINING – ADAPTATION INTERFACE: (FM)
Analysis of Foundation Models

4.10 Theory

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

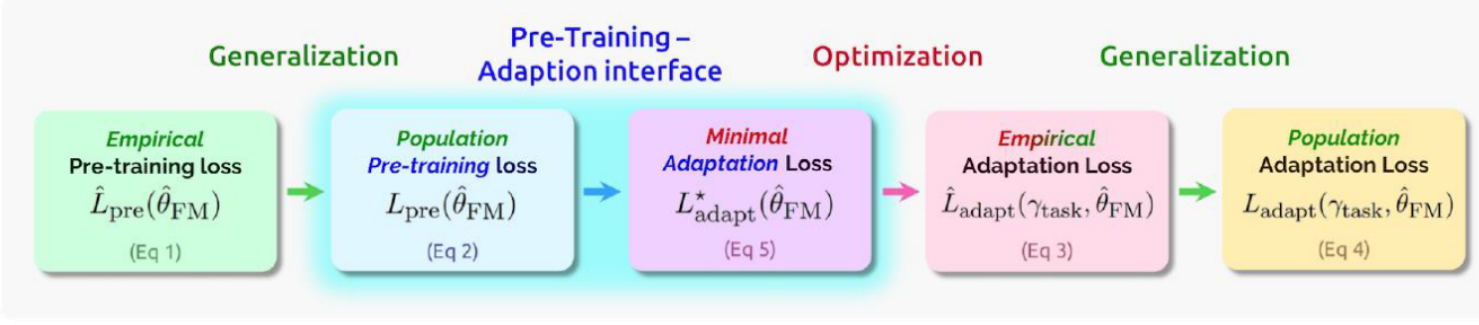
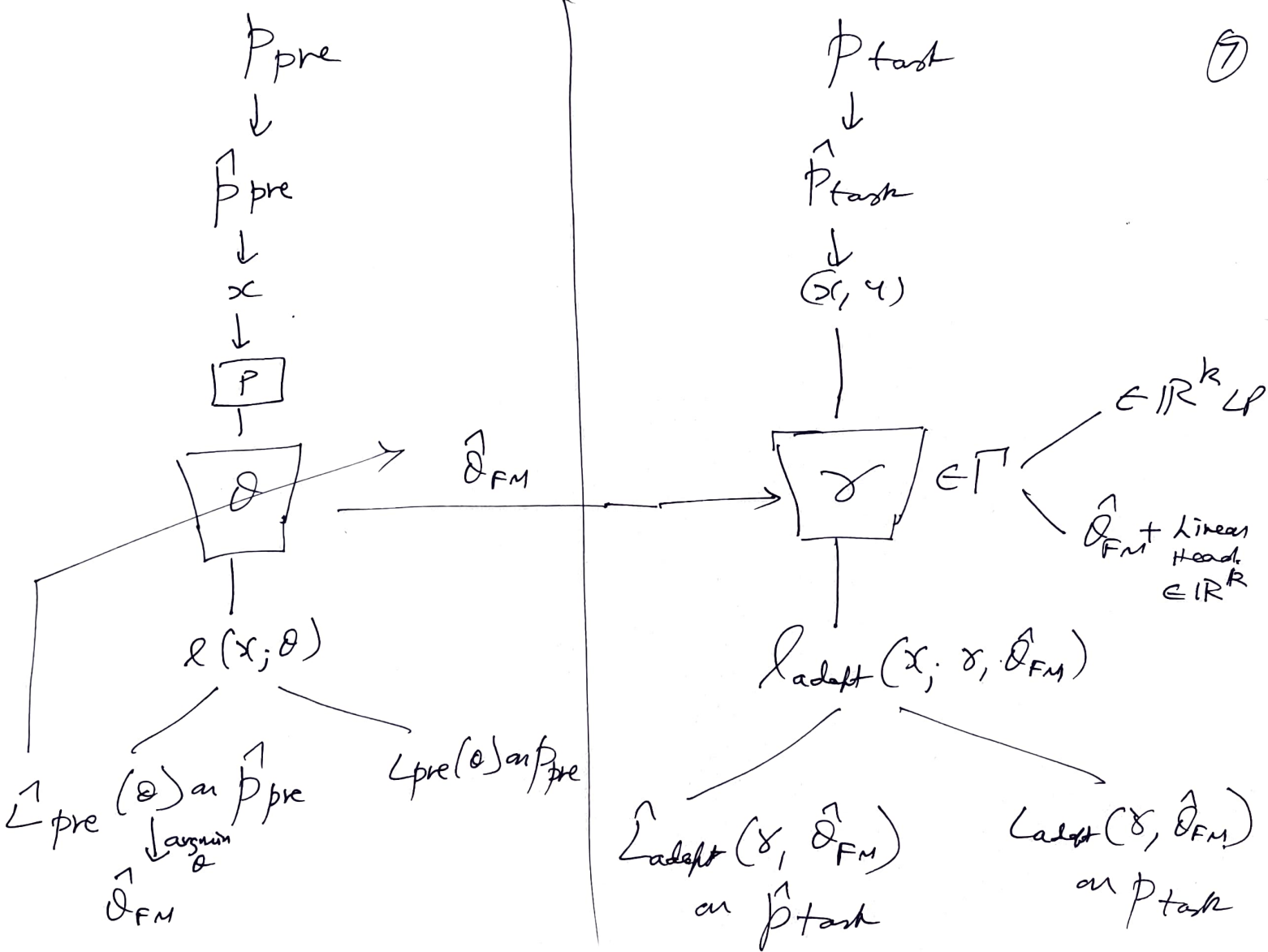
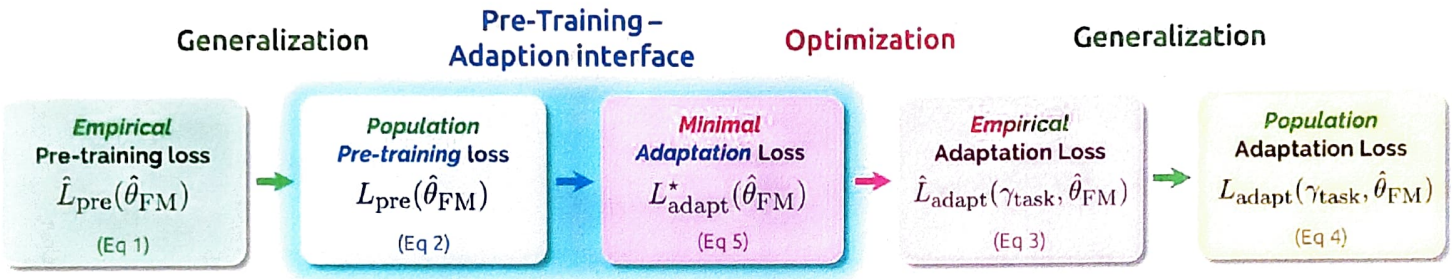


Fig. 22. The analysis of foundation models from pretraining on diverse data to downstream performance on adapted tasks involves capturing the relation between different loss terms as shown above. The main challenge is to analyze the highlighted pretraining-adaptation interface which requires reasoning 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.


p_{pre}
$\hat{p}_{pre}$
x
P
θ
$\hat{\theta}_{FM}$
$\ell(x;\theta)$
$\mathcal{L}_{pre}(\theta)$ on $\hat{p}_{pre}$
$\mathcal{L}_{pre}(\theta)$ on p_{pre}
$\arg\min_\theta$
$\hat{\theta}_{FM}$
p_{task}
$\hat{p}_{task}$
(x, y)
γ
$\in \Gamma$
$\in \mathbb{R}^k$ LP
$\hat{\theta}_{FM}$ + Linear Head. $\in \mathbb{R}^R$
$\ell_{adapt}(x; \gamma, \hat{\theta}_{FM})$
$\mathcal{L}_{adapt}(\gamma, \hat{\theta}_{FM})$ on $\hat{p}_{task}$
$\mathcal{L}_{adapt}(\gamma, \hat{\theta}_{FM})$ on p_{task}



$$\hat{L}_{pre}(\theta) = \mathbb{E}_{x \sim \hat{p}_{pre}}\left[\ell_{pre}(x;\theta)\right] \quad -(1)$$

$$\hat{\theta}_{FM} = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \hat{L}_{pre}(\theta)$$

$$L_{pre}(\theta) = \mathbb{E}_{x \sim p_{pre}}\left[\ell_{pre}(x;\theta)\right] \quad -(2)$$

$$\hat{L}_{adapt}(\gamma, \hat{\theta}_{FM}) = \mathbb{E}_{x \sim \hat{p}_{task}}\left[\ell_{adapt}(x;\gamma,\hat{\theta}_{FM})\right] \quad -(3)$$

$$\gamma_{task}(\hat{\theta}_{FM}) = \underset{\gamma \in \Gamma}{\operatorname{argmin}}\ \hat{L}_{adapt}(\gamma, \hat{\theta}_{FM})$$

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

$$\gamma^{*}_{task} = \underset{\gamma \in \Gamma,\ C(\gamma,\hat{\theta}_{FM}) \le c_0}{\operatorname{argmin}}\ L_{adapt}(\gamma, \hat{\theta}_{FM})$$

$$L^{*}_{adapt}(\hat{\theta}_{FM}) = L_{adapt}(\gamma^{*}_{task}, \hat{\theta}_{FM}) \quad -(5)$$

PRETRAINING | ADAPTATION

Generalization | Pre-Training – Adaption interface | Optimization | Generalization

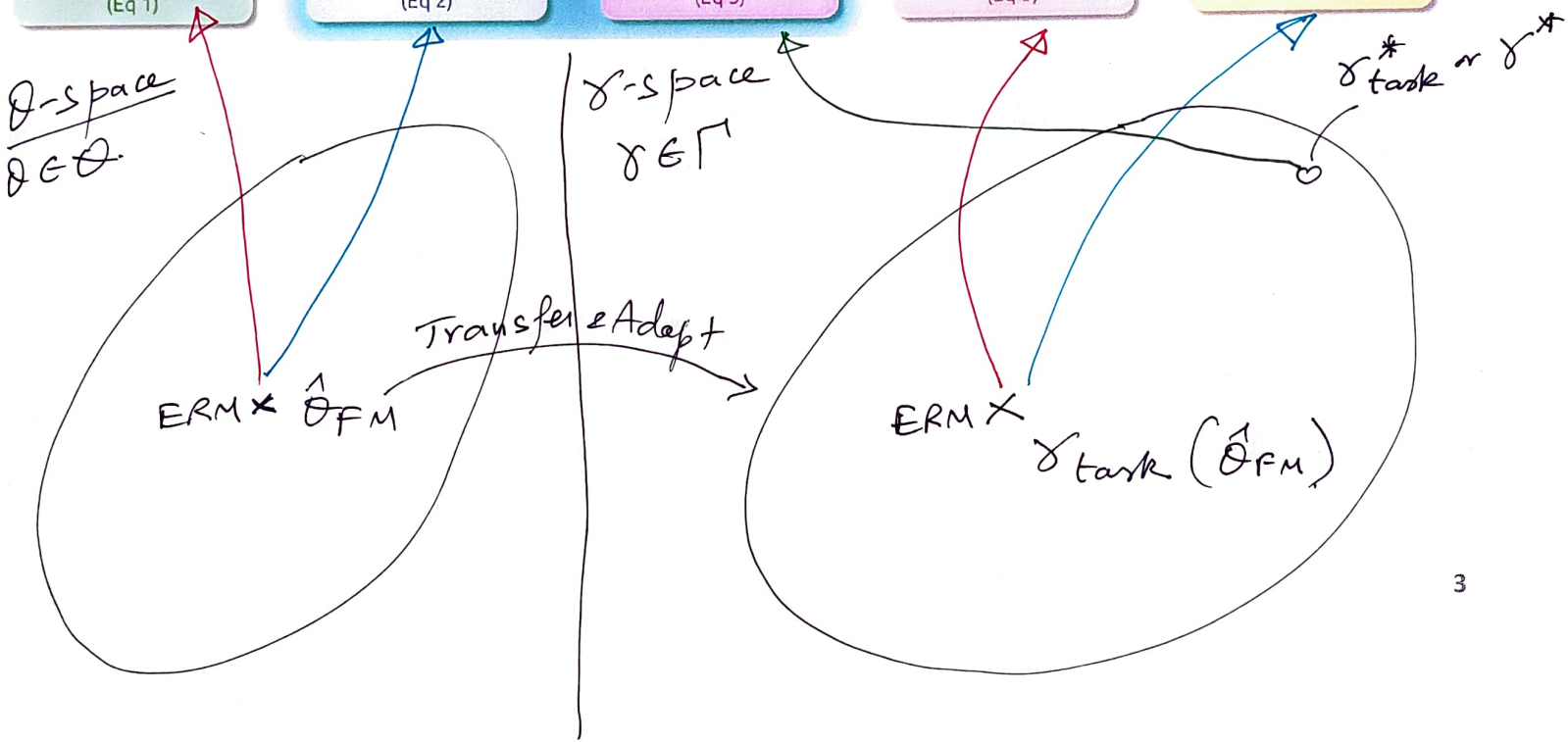
Empirical Pre-training loss
$\hat{L}_{\text{pre}}(\hat{\theta}_{\text{FM}})$
(Eq 1)

Population *Pre-training* loss
$L_{\text{pre}}(\hat{\theta}_{\text{FM}})$
(Eq 2)

Minimal *Adaptation* **Loss**
$L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}})$
(Eq 5)

Empirical Adaptation Loss
$\hat{L}_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}})$
(Eq 3)

Population Adaptation Loss
$L_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}})$
(Eq 4)



PRE-TRAINING $\xrightarrow{\hat{\theta}_{FM}}$ ADAPTATION

Generalization | Pre-Training – Adaption interface | Optimization | Generalization

Empirical Pre-training loss $\hat{L}_{\text{pre}}(\hat{\theta}_{\text{FM}})$ (Eq 1)

Population Pre-training loss $L_{\text{pre}}(\hat{\theta}_{\text{FM}})$ (Eq 2)

Minimal Adaptation Loss $L^{\star}_{\text{adapt}}(\hat{\theta}_{\text{FM}})$ (Eq 5)

Empirical Adaptation Loss $\hat{L}_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}})$ (Eq 3)

Population Adaptation Loss $L_{\text{adapt}}(\gamma_{\text{task}}, \hat{\theta}_{\text{FM}})$ (Eq 4)

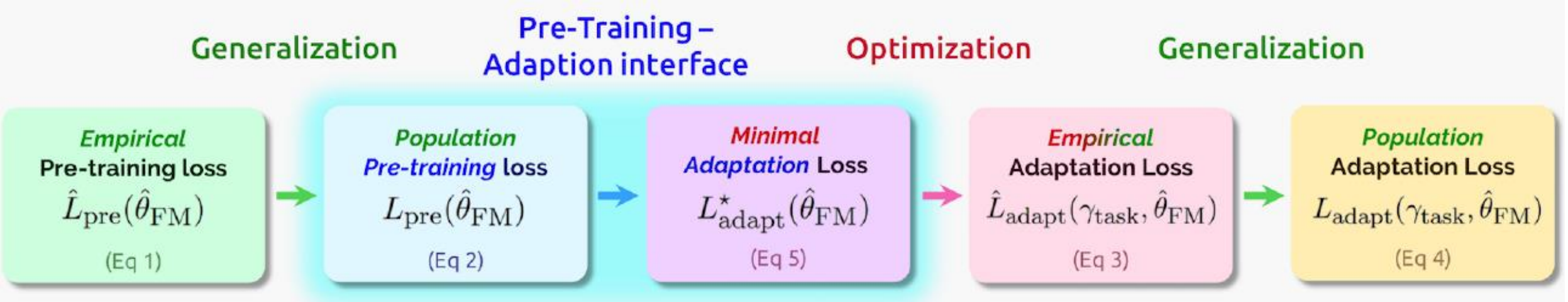
$L_a(\gamma_{task}) \leq L_a^* +$ Generalization Error.

$L_a(\gamma^*_{task})$ Minimal Adaptation Loss

To be replaced by

$L_{pre}(\hat{\theta}_{FM})$

Smaller $L_{pre}(\hat{\theta}_{FM}) \Rightarrow$ Smaller $L_a(\gamma_{task})$



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?

Conditions / Factors affecting / influencing the statement / analysis

1. Pretraining – Adaptation Interface
   – Two different population quantities / distributions

Pretraining → Task.

- How do the two distributions relate to each other.
- Effect of distribution shifts ≈ Structural shifts

2. Model Architecture.
   - Pretraining Distribution impact on Intermediate Representations in DFN (e.g. No, Ky split)

3. Few shot Learning in Downstream Supervised Task.

Small "Population Pretraining Law"

Low Complexity
γ_{task} (eg LP)

Sample efficiency
[Low, small $\hat{\gamma}_{task}$]

4- IMPORTANT : Choice of ℓ_{pre} & ℓ_{adapt}.

Minimization of ℓ_{pre} using a particular ℓ_{pre} [Best suited for a given PRETEXT task]

↓

Different "surrogate" objectives

Need not be optimal for

Downstream ℓ_{adapt} on a particular ℓ_{adapt}

→

Good adaptation on a wide range of Downstream tasks.

A Theoretical Analysis of Contrastive Unsupervised Representation Learning

Sanjeev Arora [1 2] **Hrishikesh Khandeparkar** [1] **Mikhail Khodak** [3] **Orestis Plevrakis** [1] **Nikunj Saunshi** [1]

[1]Princeton University, Princeton, New Jersey, USA. [2]Institute for Advanced Study, Princeton, New Jersey, USA. [3]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence to: Orestis Plevrakis <orestisp@cs.princeton.edu>, Nikunj Saunshi <nsaunshi@cs.princeton.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Thank you !!