1

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

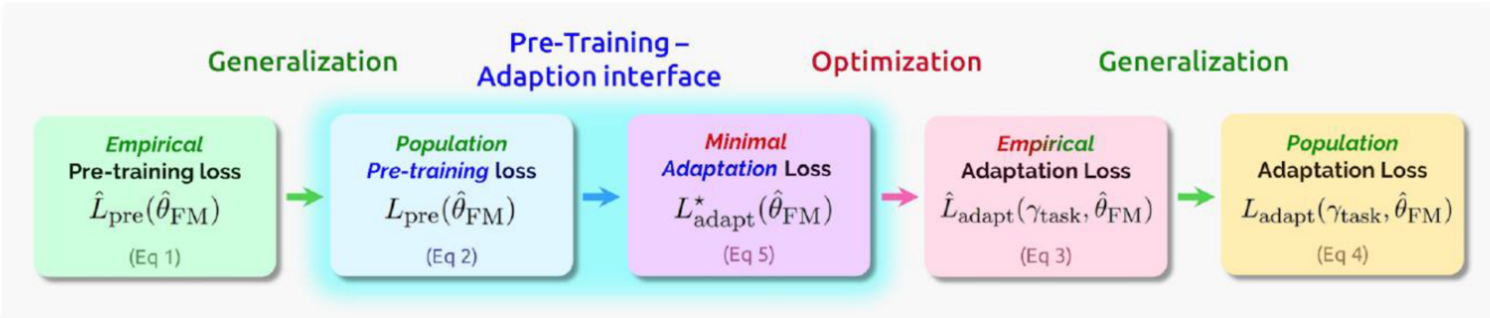


Fig. 22. The analysis of foundation models from pretraining on diverse data to downstream performance on adapted tasks involves capturing the relation between different loss terms as shown above. The main challenge is to analyze the highlighted pretraining-adaptation interface which requires reasoning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

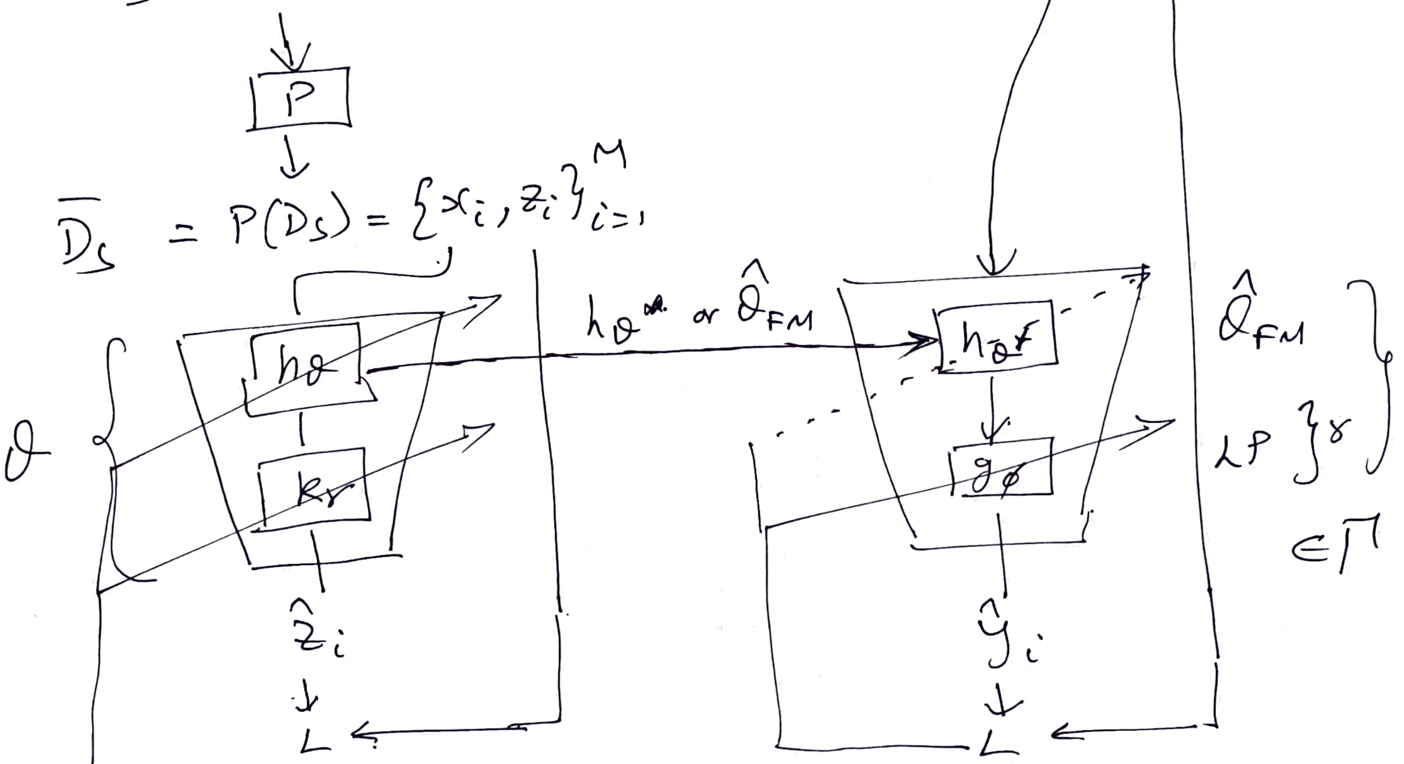
PRE-TRAINING

ADAPTATION [Downstream task]

$D_t = \{x_i^{(t)}, y_i^{(t)}\}_{i=1}^{N}$

$D_s = \{x_i^{(s)}\}_{i=1}^{M}$ $\quad M \gg N$

$\bar{D}_s = P(D_s) = \{x_i, z_i\}_{i=1}^{M}$



$L(\hat{z}_i, z_i) \to \ell_{pre}(x;\theta)$ $\qquad L(\hat{y}_i, y_i) \to \ell_{adapt}(x;\gamma,\hat{\theta}_{FM})$

Pretraining Phase

$\hat{P}_{pre} \sim P_{pre}$

P_{pre} — Distribution of raw domain data [text, speech, images...]

$\hat{P}_{pre}$ → Empirical distribution over a large no. of independent samples from $\underline{P_{pre}}$

Optimization-based Adaptation Phase

$\hat{P}_{task} \sim P_{task}$

P_{task} → Downstream "task" distribution

$\hat{P}_{task}$ → "Few" empirical samples sampled from $\underline{P_{task}}$

$$\ell_{pre}(x;\theta) \triangleq L(z_i, \hat{z}_i)$$

- ERM [Empirical Risk Minimizer]

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

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

$\downarrow$

Derived FOUNDATION MODEL [ERM]

$\boxed{\text{ERM of Task}}$ $\longrightarrow$

adapted from FM $\hat{\theta}_{FM}$

Γ: Space of adapted model parameters

$\subseteq \hat{\theta}_{FM}$

Since different adaptation methods could modify different subsets of the $\hat{\theta}_{FM}$

Recall: h_θ, g_ϕ split / Combination

- ERM [Downstream task]

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

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

Regularization $\| \ C(\gamma, \hat{\theta}_{FM}) \leq C_0$

Expected Risk / Population loss

on the population distribution p_{pre}

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

True, unknown distribution

on the population distribution p_{task}.

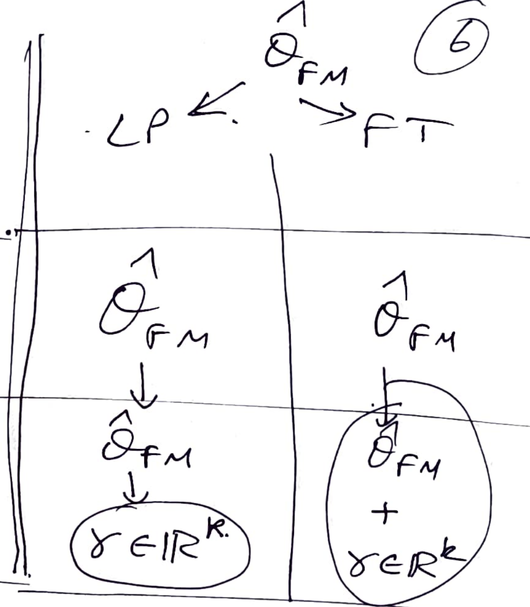
$$L_{adapt}(\gamma, \hat{\theta}_{FM}) = \mathbb{E}_{x \sim p_{task}} \left[\ell_{adapt}(x; \gamma, \hat{\theta}_{FM}) \right]$$

$$\gamma^{*}_{task} = \underset{\gamma \in \Gamma}{\arg\min} \; L_{adapt}(\gamma, \hat{\theta}_{FM})$$

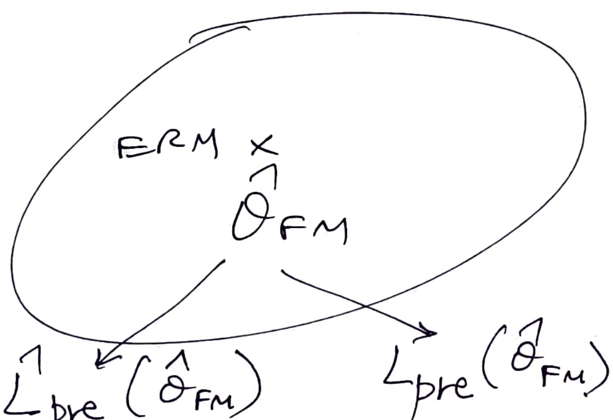
$$C(\gamma, \hat{\theta}_{FM}) \leq C_0$$

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

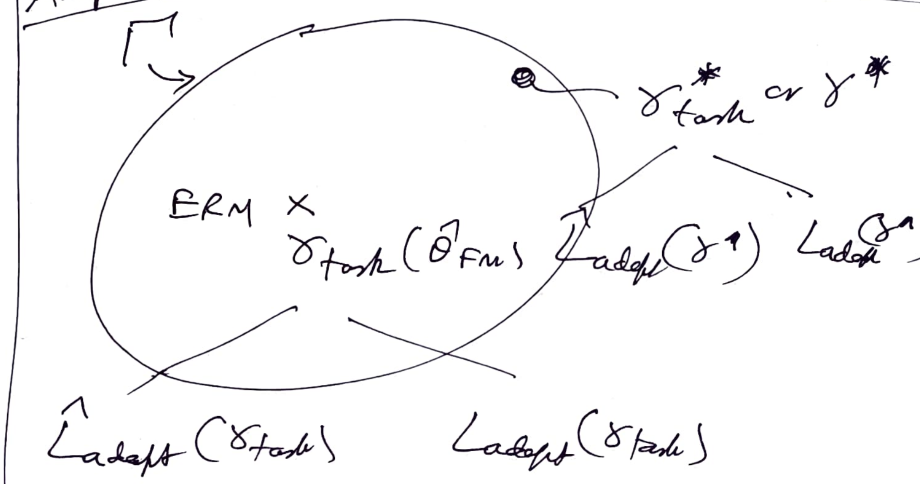
	Empirical Risk (ERM)	Population Loss
Pretraining	$\hat{L}_{pre} / \hat{\theta}_{FM}$ on $\hat{p}_{pre}$	L_{pre} on p_{pre}
Adaptation	$\hat{L}_{adapt} / \gamma_{task}$ on $\hat{p}_{task}$.	$L_{adapt} / \gamma^*_{task}$ on p_{task}



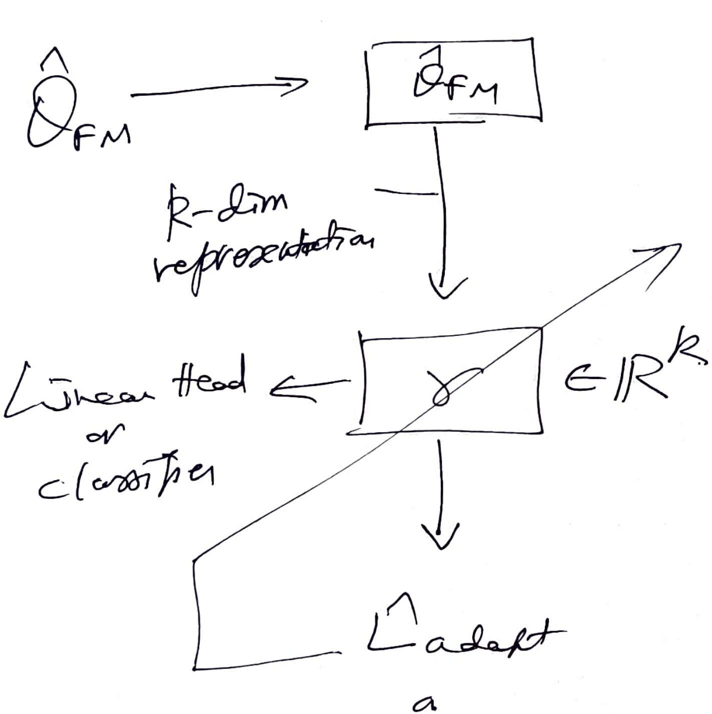
Pretraining θ-space



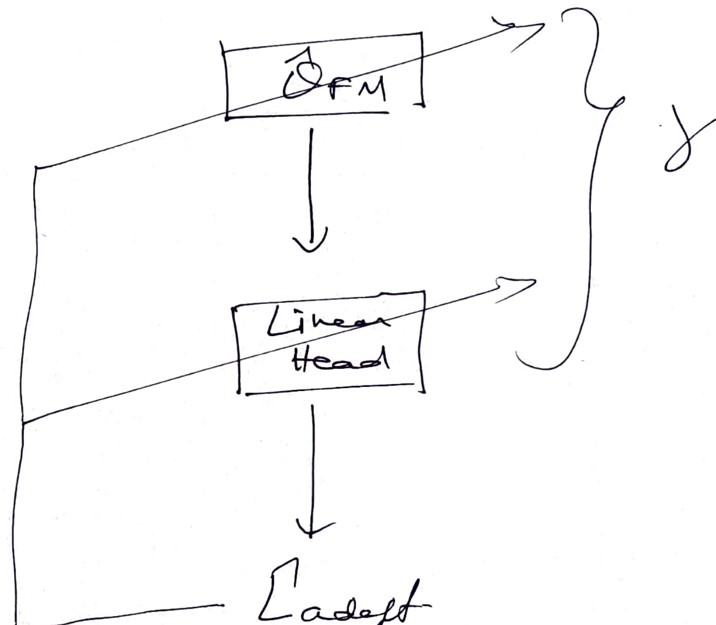
Adaptation

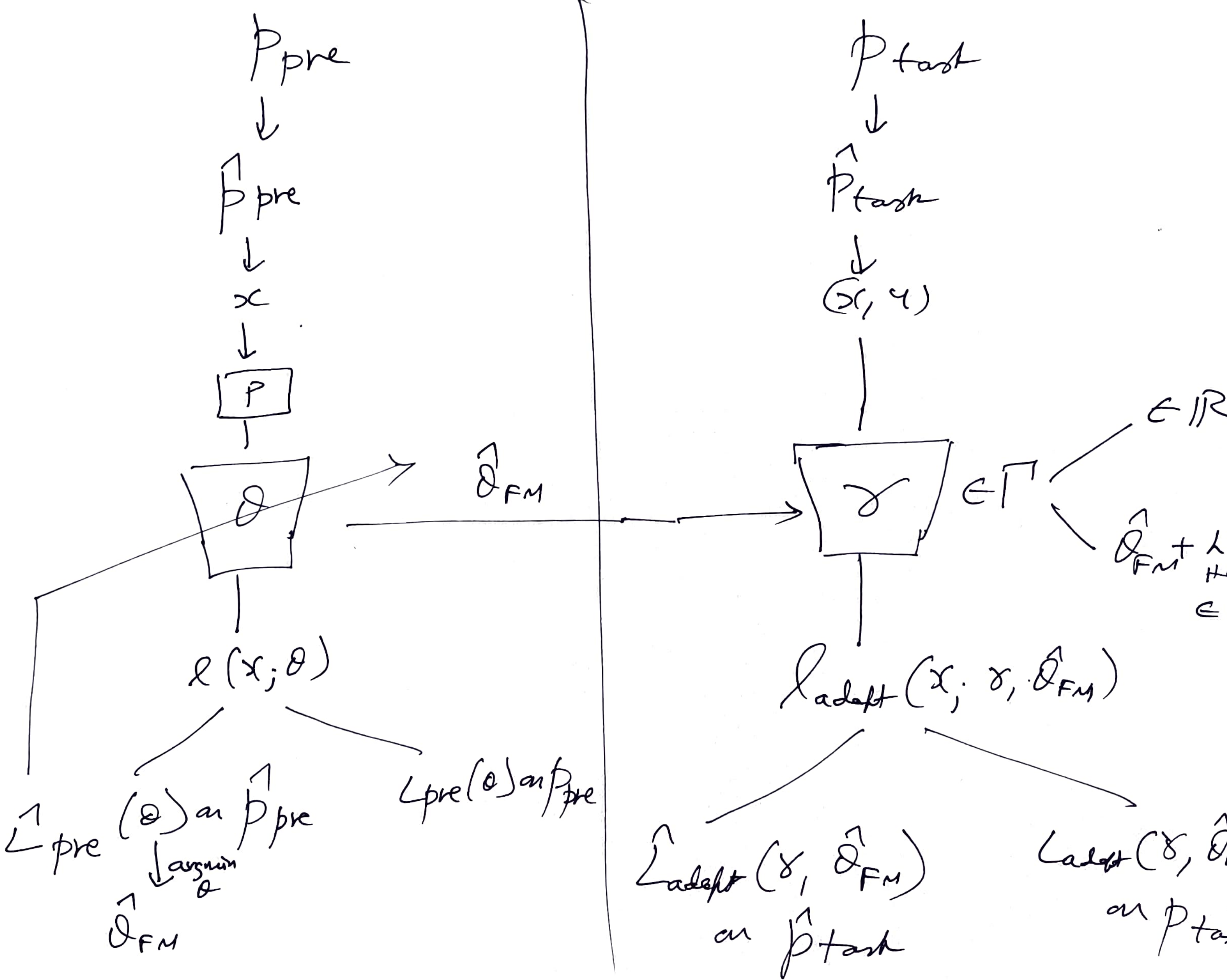


Linear Probing



Fine Tuning



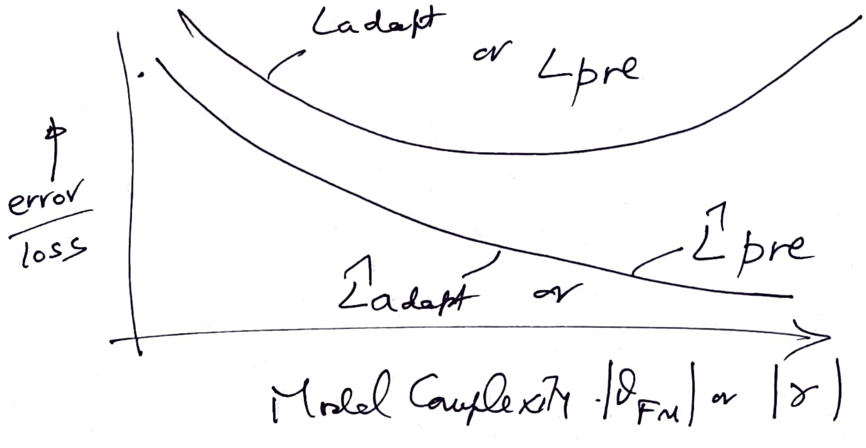

P_{pre}
$\hat{p}_{pre}$
x
p
θ
$\hat{\theta}_{FM}$
$\ell(x;\theta)$
$\hat{L}_{pre}(\theta)$ on $\hat{p}_{pre}$
$\arg\min_\theta$
$\hat{\theta}_{FM}$
$L_{pre}(\theta)$ on P_{pre}
P_{task}
$\hat{P}_{task}$
(x, y)
γ
$\in \Gamma$
$\in \mathbb{R}^k \ll p$
$\hat{\theta}_{FM}$ + Linear Head $\in \mathbb{R}^k$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

Modularized Phases - Analysis

SLT Concerns how

$$\hat{L}_{pre} \approx L_{pre}$$

$$\hat{L}_{adapt} \approx L_{adapt}$$



The Generalization Error

$$= L_{pre} - \hat{L}_{pre}$$

Standard Error Decomposition

Generalization Error

$$= \left[L_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) - \hat{L}_{adapt}(\gamma_{task}, \hat{\theta}_{FM}) \right] \quad (1)$$

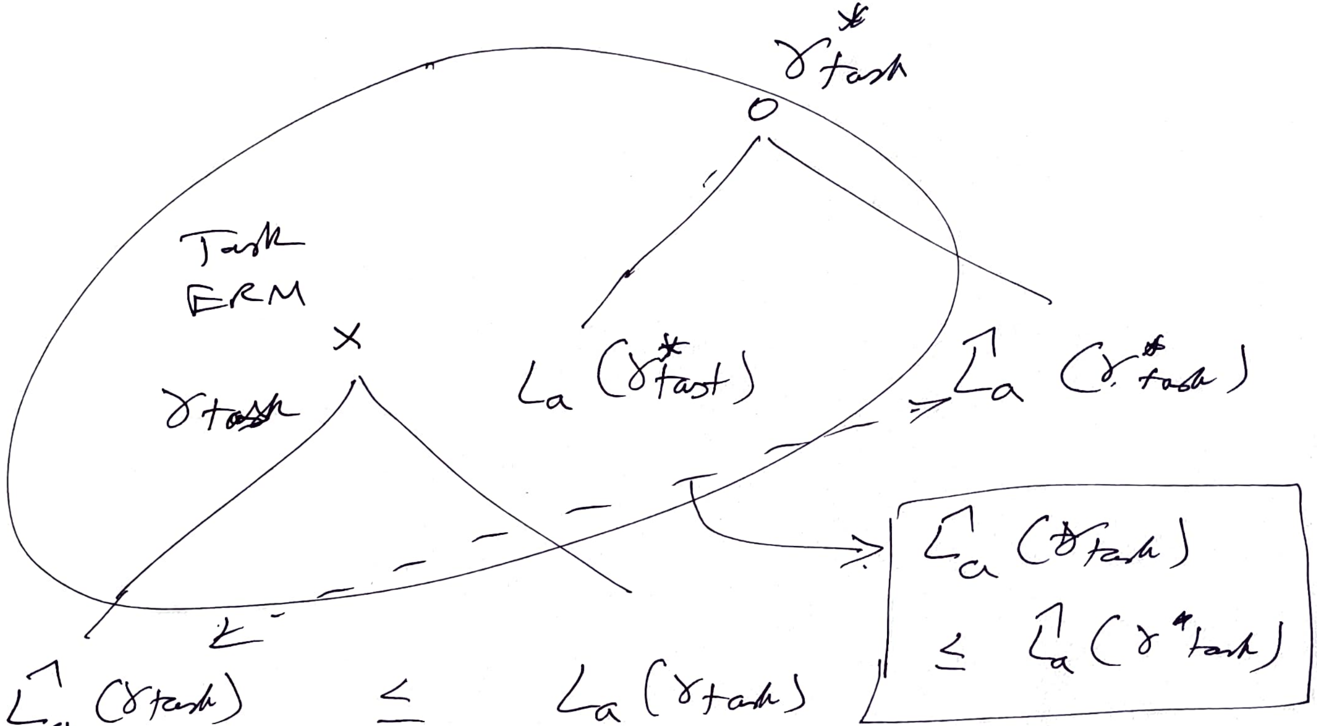
$$+ \left[\hat{L}_{adapt}(\gamma^{*}_{task}, \hat{\theta}_{FM}) - L_{adapt}(\gamma^{*}_{task}, \hat{\theta}_{FM}) \right]$$

Setting L_{adapt} as L_a & dropping $\hat{\theta}_{FM}$ in (1) for simplicity

$$GE = \left[L_a(\gamma_{task}) - \hat{L}_a(\gamma_{task}) \right] \quad (2)$$

$$+ \left[\hat{L}_a(\gamma^{*}_{task}) - L_a(\gamma^{*}_{task}) \right]$$

$$GE = \left[L_a(\gamma_{task}) - \hat{L}_a(\gamma_{task}) \right] + \left[\hat{L}_a(\gamma^*_{task}) - L_a(\gamma^*_{task}) \right]$$



$$GE = \left[L_a(\gamma_{task}) - \hat{L}_a(\gamma_{task}) \right] + \left[\hat{L}_a(\gamma^*_{task}) - L_a^* \right]$$

$$L_a^* = L_a(\gamma^*_{task})$$

$\Downarrow$

$\boxed{\text{Minimal Adaptation Loss}}$

Using $\hat{L}_a(\gamma_{task}) \leq \hat{L}_a(\gamma^*_{task})$

Best loss model on Empirical Risk Minimization ie on $\hat{p}_{task}$

Best loss model from Expected Risk Minimization ie on p_{task}

Different Loss Surfaces

— γ^*_{task} optimized on p_{task} loss is NOT optimal for loss surface $\hat{p}_{task} \sim p_{task}$

(12)

$$GE = \left[L_a(\gamma_{task}) - \hat{L}_a(\gamma_{task}) \right] + \left[\hat{L}_a(\gamma^*_{task}) - L^*_a \right]$$

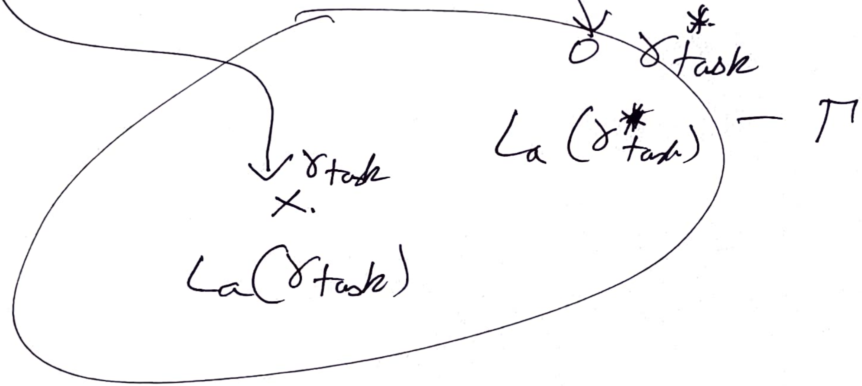
Using $\hat{L}_a(\gamma_{task}) \leq \hat{L}_a(\gamma^*_{task})$

we have

$$\boxed{L_a(\gamma_{task}, \hat{\theta}_{FM}) \leq L^*_a(\hat{\theta}_{FM}) + \text{Generalization Error}}$$

(Excess)

$$L_a(\gamma_{task}, \hat{\theta}_{FM}) \le L_a^*(\hat{\theta}_{FM}) + \text{Excess Generalization Error}$$



ie Population Loss (ERM$_{task}$ Model) $\le$ Population Loss (Best Model in Γ)

+ GE

Kind of Main Results shown

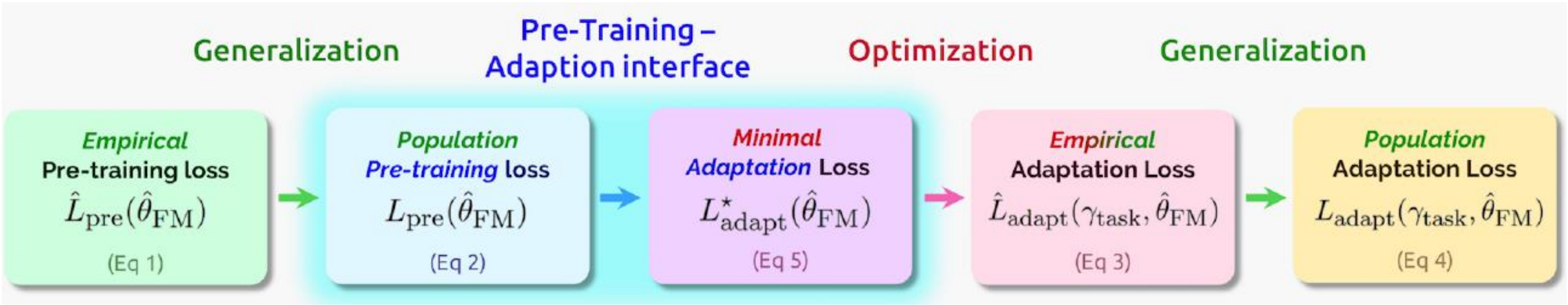
Population Loss (ERM Task Model) $\leq$ Population Loss (Best Model in $\mathcal{H}$)

+ GE

$\Downarrow$

To use $\mathcal{L}_{pre}(\hat{\theta}_{FM})$

i.e. we can Bound the Performance of downstream adapted Model ERM γ task [Popln Loss a P task] by PRETRAINING Population Loss of FM



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?

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.

Conditions | Factors affecting / influencing the statement / analysis

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

Pretraining

Task.

- How do ~~these~~ two distributions relate to each other.
- Effect of distribution shifts & structural shifts

2. Model Architecture.
   - Pretraining Distribution $\xrightarrow{\text{impact on}}$ Intermediate Representations in $\overrightarrow{Q}_{FN}$ (e.g. No, K_s split)

3. Few Shot Learning in Downstream Supervised Task.

Small "Population Pretraining Loss"

Low Complexity γ_{task} (eg LP)

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

4. <u>IMPORTANT</u> : Choice of l_{pre} & l_{adapt}.

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

$\xrightarrow{\text{Need not be optimal for}}$

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

$\downarrow$

Different "surrogate" objectives

$\longrightarrow$

Good adaptation on a wide range of Downstream tasks.

Thank you !!