Regularized Fine-Tuning for Representation Multi-Task Learning: Adaptivity, Minimaxity, and Robustness

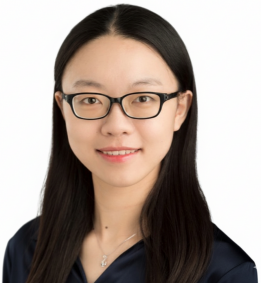
Yang Feng

Department of Biostatistics, New York University

UFL Statistics Winter Workshop
Jan 2026

Joint work with

Ye Tian
(Yale University)

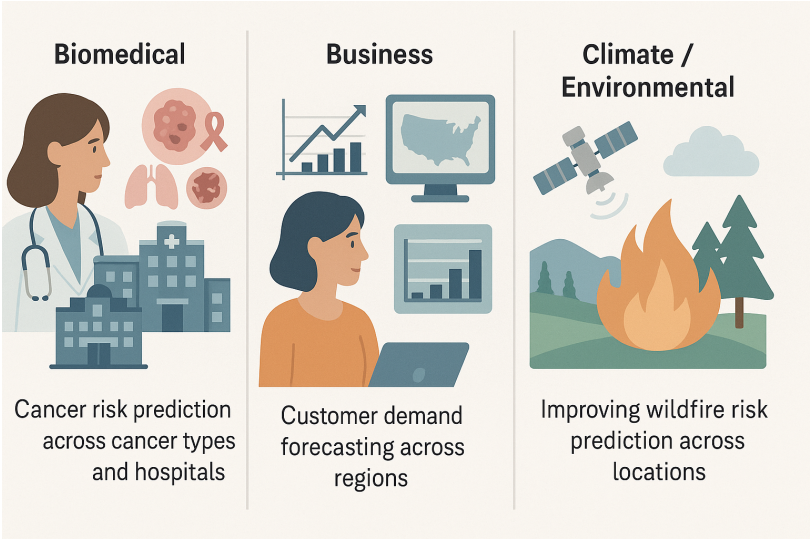


Yuqi Gu
(Columbia University)

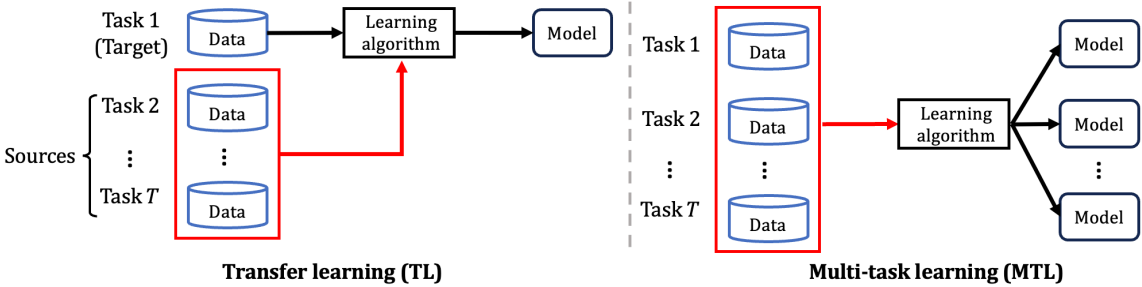
Outline

§1 Introduction

Motivation: Data Integration Across Multiple Sources

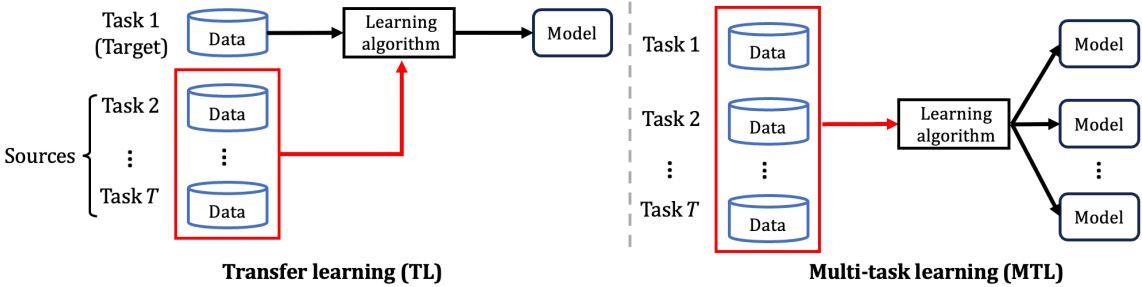


Knowledge transfer



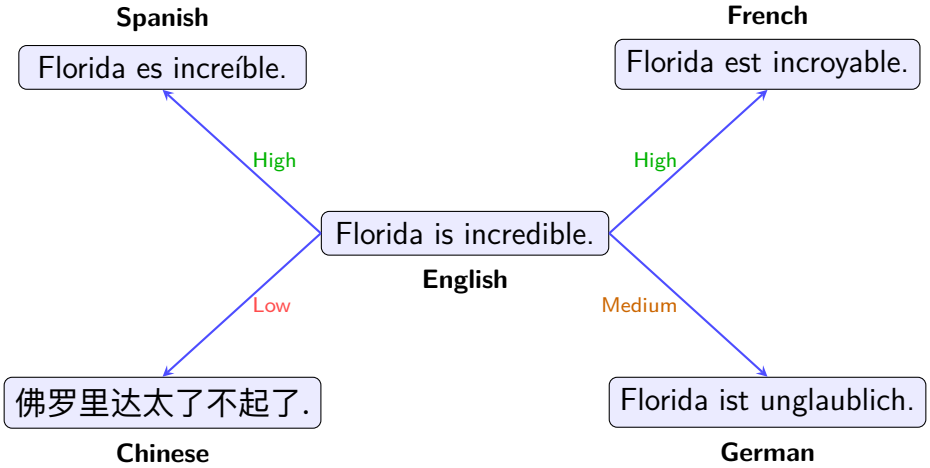
- Borrowing information across tasks can reduce variance and improve generalization.
- However, task heterogeneity or corrupted data (outlier tasks) can degrade performance.
- **Goal:** develop adaptive and robust methods to transfer knowledge effectively under heterogeneity and possible task-level contamination.

Knowledge transfer

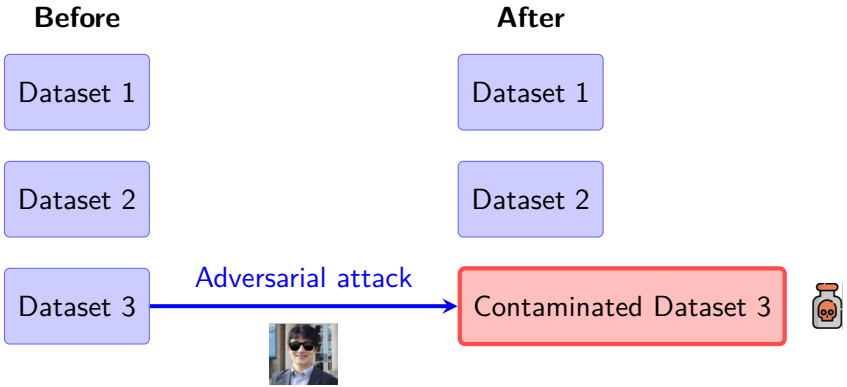


- Borrowing information across tasks can reduce variance and improve generalization.
- However, task heterogeneity or corrupted data (outlier tasks) can degrade performance.
- **Goal:** develop adaptive and robust methods to transfer knowledge effectively under heterogeneity and possible task-level contamination.

Challenge 1: Unknown Similarity

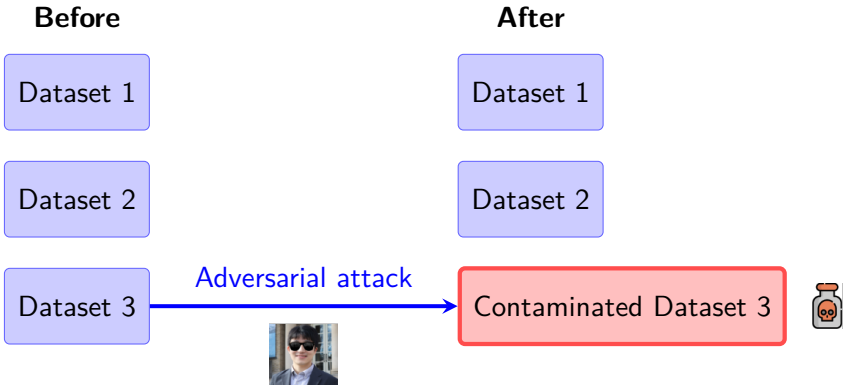


Challenge 2: Data Contamination



- **Today:** how to address them in **representation transfer/multi-task learning**

Challenge 2: Data Contamination



- **Today:** how to address them in **representation transfer/multi-task learning**

Background: What is Representation Learning?

- **Goal:** Learn a compact, informative feature subspace that summarizes complex, high-dimensional data.
- Instead of analyzing each dataset or task separately, we discover a **shared representation** that captures the underlying common structure.
- **Examples:**
  - In **biomedicine**: integrating gene expression and imaging data to predict *cancer risk* across institutions.
  - In **business analytics**: learning customer embeddings that generalize across *regions or product lines*.
  - In **environmental science**: deriving latent features from satellite and sensor data for *climate risk prediction*.
- Such representations enable **transfer and generalization**, allowing related tasks to learn collaboratively and adapt under heterogeneity.

§2 Setup, Algorithms, and Theory

Setup: Representation multi-task learning

- We observe $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task:

$$y_i^{(t)} = g^{(t)}\left(f^{(t)}(\boldsymbol{x}_i^{(t)})\right) + \epsilon_i^{(t)}, \quad i = 1:n, \quad t = 1:T,$$

  where $f^{(t)}$ is the representation and $g^{(t)}$ is the downstream model.

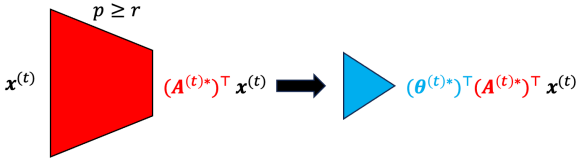
Setup: Representation multi-task learning

- We observe $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task:

$$y_i^{(t)} = g^{(t)}\left(f^{(t)}(\boldsymbol{x}_i^{(t)})\right) + \epsilon_i^{(t)}, \quad i = 1:n, \quad t = 1:T,$$

  where $f^{(t)}$ is the representation and $g^{(t)}$ is the downstream model.

- In this talk:
  - linear representation $f^{(t)}(\boldsymbol{x}) = (\boldsymbol{A}^{(t)*})^\top \boldsymbol{x}$, $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p\times r}$ [1].
  - linear downstream model $g^{(t)}(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{\theta}^{(t)*}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$



- In other words: $y_i^{(t)} = (\boldsymbol{\beta}^{(t)*})^\top \boldsymbol{x}_i^{(t)} + \epsilon_i^{(t)}$, with $\boldsymbol{\beta}_{p\times 1}^{(t)*} = \boldsymbol{A}_{p\times r}^{(t)*} \boldsymbol{\theta}_{r\times 1}^{(t)*}$.

[1] $\mathcal{O}^{p\times r} := \{\boldsymbol{A} \in \mathbb{R}^{p\times r} : \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r\}$

Setup: Representation multi-task learning

- **Unknown similarity:** $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in [T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h, \text{ where } \overline{\boldsymbol{A}} \text{ is an "average" representation.}$$

- **Task-level ϵ-contamination:** Konstantinov and Lampert (2019); Konstantinov et al. (2020)

  The attacker picks $S^c \subseteq [T]$ with $|S^c|/T \leq \epsilon$, for $t \in S^c$, :

$$\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n \xrightarrow{\text{attacker}} \{\widetilde{\boldsymbol{x}}_i^{(t)}, \widetilde{y}_i^{(t)}\}_{i=1}^n$$

  - A stronger "MTL version" of Huber contamination (Huber, 1964)

- **Goal:** Learn uncontaminated $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$, $S = [T] \backslash S^c$

[Hub64] Huber, P. J. (1964). Robust estimation of a location parameter.
[KL19] Konstantinov, N. et al. (2020). On the sample complexity of adversarial multi-source PAC learning.
[KFAL20] Konstantinov, N., & Lampert, C. (2019). Robust learning from untrusted sources.

Setup: Representation multi-task learning

- **Unknown similarity:** $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}} \max_{t\in[T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h, \text{ where } \overline{\boldsymbol{A}} \text{ is an "average" representation.}$$

- **Task-level ϵ-contamination:** Konstantinov and Lampert (2019); Konstantinov et al. (2020)

  The attacker picks $S^c \subseteq [T]$ with $|S^c|/T \leq \epsilon$, for $t \in S^c$, :

$$\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n \xrightarrow{\text{attacker}} \{\widetilde{\boldsymbol{x}}_i^{(t)}, \widetilde{y}_i^{(t)}\}_{i=1}^n$$

  - A stronger "MTL version" of Huber contamination (Huber, 1964)

- **Goal:** Learn uncontaminated $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$, $S = [T]\backslash S^c$

[Hub64] Huber, P. J. (1964). Robust estimation of a location parameter.
[KL19] Konstantinov, N. et al. (2020). On the sample complexity of adversarial multi-source PAC learning.
[KFAL20] Konstantinov, N., & Lampert, C. (2019). Robust learning from untrusted sources.

Setup: Representation multi-task learning

- **Unknown similarity:** $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p \times r}} \max_{t \in [T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h, \text{ where } \overline{\boldsymbol{A}} \text{ is an "average" representation.}$$

- **Task-level ϵ-contamination:** Konstantinov and Lampert (2019); Konstantinov et al. (2020)

  The attacker picks $S^c \subseteq [T]$ with $|S^c|/T \leq \epsilon$, for $t \in S^c$, :

$$\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n \xrightarrow{\text{attacker}} \{\widetilde{\boldsymbol{x}}_i^{(t)}, \widetilde{y}_i^{(t)}\}_{i=1}^n$$

  - A stronger "MTL version" of Huber contamination (Huber, 1964)

- **Goal:** Learn uncontaminated $\{\boldsymbol{\beta}^{(t)*}\}_{t \in S}$, $S = [T] \backslash S^c$

[Hub64] Huber, P. J. (1964). Robust estimation of a location parameter.

[KL19] Konstantinov, N. et al. (2020). On the sample complexity of adversarial multi-source PAC learning.

[KFAL20] Konstantinov, N., & Lampert, C. (2019). Robust learning from untrusted sources.

Setup: Representation multi-task learning

- **Unknown similarity:** $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}} \max_{t\in[T]} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^\top - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^\top\|_2 \leq h, \text{ where } \overline{\boldsymbol{A}} \text{ is an "average" representation.}$$

- **Task-level ϵ-contamination:** Konstantinov and Lampert (2019); Konstantinov et al. (2020)

  The attacker picks $S^c \subseteq [T]$ with $|S^c|/T \leq \epsilon$, for $t \in S^c$, :

$$\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n \xrightarrow{\text{attacker}} \{\widetilde{\boldsymbol{x}}_i^{(t)}, \widetilde{y}_i^{(t)}\}_{i=1}^n$$

  - A stronger "MTL version" of Huber contamination (Huber, 1964)

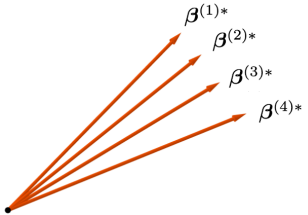
- **Goal:** Learn uncontaminated $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$, $S = [T]\backslash S^c$

[Hub64] Huber, P. J. (1964). Robust estimation of a location parameter.

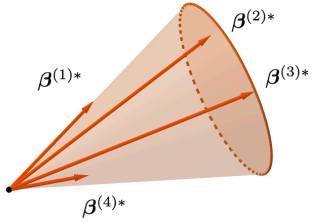
[KL19] Konstantinov, N. et al. (2020). On the sample complexity of adversarial multi-source PAC learning.

[KFAL20] Konstantinov, N., & Lampert, C. (2019). Robust learning from untrusted sources.

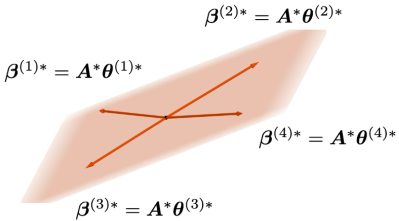
Related works: other parametric knowledge transfer



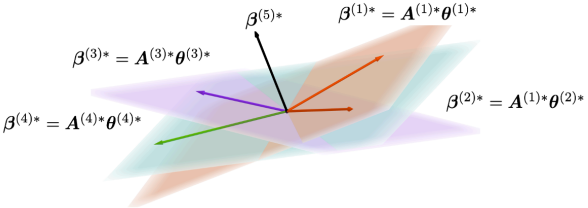
(a) Distance-based similarity



(b) Angle-based similarity

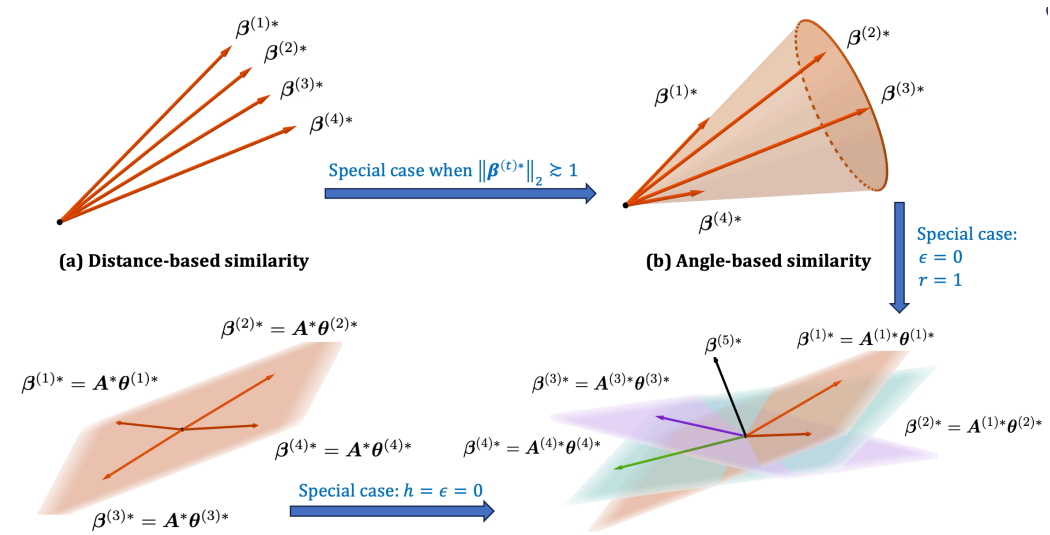


(c) Sharing a representation

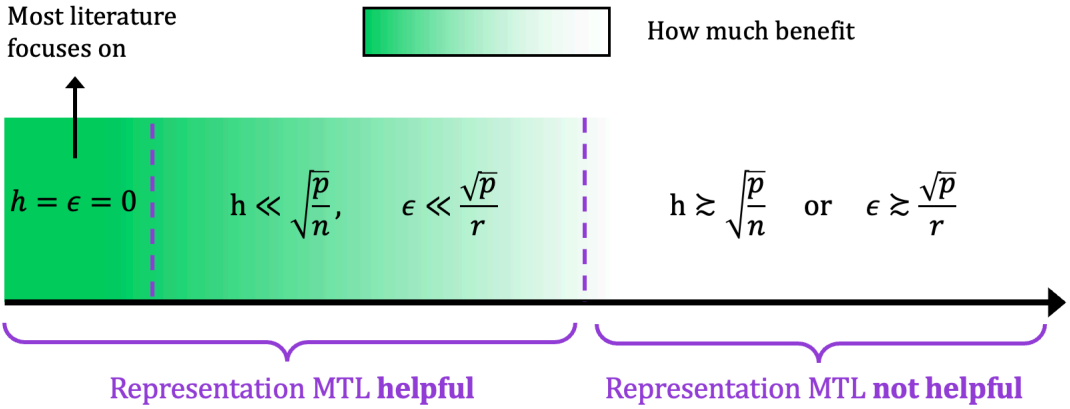


(d) Our setting

Related works: other parametric knowledge transfer



Different Regimes



Algorithm 1: penalized ERM

Algorithm 1 (Penalized ERM): Given $\lambda, \gamma > 0$

- **Step 1: (Pooling)** $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \leftarrow \min_{\boldsymbol{A}^{(t)}, \overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}, \boldsymbol{\theta}^{(t)} \in \mathbb{R}^r} \sum_{t=1}^{T} \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i^{(t)} - (\boldsymbol{x}^{(t)})^\top \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}} \| \boldsymbol{A}^{(t)} (\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}} (\overline{\boldsymbol{A}})^\top \|_2 \right\}$
- **Step 2: (Regularized Fine-Tuning)** $\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i^{(t)} - (\boldsymbol{x}^{(t)})^\top \boldsymbol{\beta}]^2 + \frac{\gamma}{\sqrt{n}} \| \boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(t)} \widehat{\boldsymbol{\theta}}^{(t)} \|_2 \right\}$

- Step 1 can be solved in $\mathbb{R}^{p\times r}$ by replacing $\boldsymbol{A}\boldsymbol{A}^\top$ with projection matrix onto $\mathrm{Col}(\boldsymbol{A})$

Algorithm 1: penalized ERM

Algorithm 1 (Penalized ERM): Given $\lambda, \gamma > 0$

- **Step 1: (Pooling)** $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \leftarrow \min_{\boldsymbol{A}^{(t)}, \overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}, \boldsymbol{\theta}^{(t)} \in \mathbb{R}^r}$

  $$\sum_{t=1}^{T} \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i^{(t)} - (\boldsymbol{x}^{(t)})^\top \boldsymbol{A}^{(t)} \boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}} \| \boldsymbol{A}^{(t)} (\boldsymbol{A}^{(t)})^\top - \overline{\boldsymbol{A}} (\overline{\boldsymbol{A}})^\top \|_2 \right\}$$

- **Step 2: (Regularized Fine-Tuning)**

  $$\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_i^{(t)} - (\boldsymbol{x}^{(t)})^\top \boldsymbol{\beta}]^2 + \frac{\gamma}{\sqrt{n}} \| \boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(t)} \widehat{\boldsymbol{\theta}}^{(t)} \|_2 \right\}$$

- Step 1 can be solved in $\mathbb{R}^{p\times r}$ by replacing $\boldsymbol{A}\boldsymbol{A}^\top$ with projection matrix onto $\mathrm{Col}(\boldsymbol{A})$

Theorem 1 (Estimation error of pERM)

(i) When $\lambda \gtrsim \sqrt{r(p+\log T)}$, $\gamma \asymp \sqrt{p+\log T}$*: up to log-terms, w.h.p.*

$$\max_{t\in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \Bigg(\underbrace{r\sqrt{\frac{p}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{\sqrt{r}h}_{\boldsymbol{A}^{(t)*} \text{not equal}} + \underbrace{r\sqrt{\frac{1}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\frac{\lambda\epsilon r}{\sqrt{n}}}_{\text{outlier tasks}} \Bigg) \wedge \underbrace{\sqrt{\frac{p}{n}}}_{\text{single-task rate}}$$

(ii) If tasks in S^c *also follow linear model: up to log-terms, w.h.p.*

$$\max_{t\in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\tfrac{p}{n}}.$$

- No contamination ($\epsilon = 0$)
  - Taking $\lambda = +\infty$ reduces to the ERM in Du et al. (2020); Tripuraneni et al. (2021)
  - No "bad" local minimizers
- With contamination ($\epsilon > 0$): let $\lambda \asymp \sqrt{r(p+\log T)}$
  - No "bad" local minimizers with good initializers
  - Adaptivity to h, robustness against contamination

Theorem 1 (Estimation error of pERM)

(i) When $\lambda \gtrsim \sqrt{r(p+\log T)}$, $\gamma \asymp \sqrt{p+\log T}$*: up to log-terms, w.h.p.*

$$\max_{t\in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \Bigg(\underbrace{r\sqrt{\frac{p}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{\sqrt{r}h}_{\boldsymbol{A}^{(t)*} \text{not equal}} + \underbrace{r\sqrt{\frac{1}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\frac{\lambda\epsilon r}{\sqrt{n}}}_{\text{outlier tasks}} \Bigg) \wedge \underbrace{\sqrt{\frac{p}{n}}}_{\text{single-task rate}}$$

(ii) If tasks in S^c *also follow linear model: up to log-terms, w.h.p.*

$$\max_{t\in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\tfrac{p}{n}}.$$

- No contamination ($\epsilon = 0$)
  - Taking $\lambda = +\infty$ reduces to the ERM in Du et al. (2020); Tripuraneni et al. (2021)
  - No "bad" local minimizers
- With contamination ($\epsilon > 0$): let $\lambda \asymp \sqrt{r(p+\log T)}$
  - No "bad" local minimizers with good initializers
  - Adaptivity to h, robustness against contamination

Theorem 1 (Estimation error of pERM)

(i) When $\lambda \gtrsim \sqrt{r(p+\log T)}$, $\gamma \asymp \sqrt{p+\log T}$*: up to log-terms, w.h.p.*

$$\max_{t\in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \Bigg(\underbrace{r\sqrt{\frac{p}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{\sqrt{r}h}_{\boldsymbol{A}^{(t)*} \text{not equal}} + \underbrace{r\sqrt{\frac{1}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\frac{\lambda\epsilon r}{\sqrt{n}}}_{\text{outlier tasks}} \Bigg) \wedge \underbrace{\sqrt{\frac{p}{n}}}_{\text{single-task rate}}$$

(ii) If tasks in S^c *also follow linear model: up to log-terms, w.h.p.*

$$\max_{t\in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\tfrac{p}{n}}.$$

- No contamination ($\epsilon = 0$)
  - Taking $\lambda = +\infty$ reduces to the ERM in Du et al. (2020); Tripuraneni et al. (2021)
  - No "bad" local minimizers
- With contamination ($\epsilon > 0$): let $\lambda \asymp \sqrt{r(p+\log T)}$
  - No "bad" local minimizers with good initializers
  - Adaptivity to h, robustness against contamination

Algorithm 2: spectral method

- When $h = 0$: $\boldsymbol{B}^* = \{\boldsymbol{\beta}^{(t)*}\}_{t \in S} \in \text{Col}(\overline{\boldsymbol{A}})$
  - SVD should work
  - Need to estimate $\boldsymbol{B}^*$ in a robust way

Algorithm 2 (Spectral method): Given $\gamma, R > 0$, an upper bound $\bar{\epsilon}$ (for ϵ)

- **Step 1: (Single-task regression)** $\widetilde{\boldsymbol{\beta}}^{(t)} = \text{OLS}$ on task t
- **Step 2: (Projection and concatenation)** $\widehat{\boldsymbol{B}} = (\prod_R(\widetilde{\boldsymbol{\beta}}^{(t)}) \ \ldots \ \prod_R(\widetilde{\boldsymbol{\beta}}^{(T)}))$, where $R = \texttt{quantile}(\{\|\widetilde{\boldsymbol{\beta}}^{(t)}\|_2\}_{t=1}^T, 1 - \bar{\epsilon})$.
- **Step 3: (SVD)** $\widehat{\widehat{\boldsymbol{A}}} =$ left sing. matrix of $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{\theta}}^{(t)} = \text{OLS}$ in $\text{Col}(\widehat{\widehat{\boldsymbol{A}}})$ on task t
- **Step 4: (Biased regularization)**

$$\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T \boldsymbol{\beta}]^2 + \frac{\gamma}{\sqrt{n}} \|\boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(t)} \widehat{\boldsymbol{\theta}}^{(t)}\|_2 \right\}$$

Algorithm 2: spectral method

Theorem 2 (Estimation error of the spectral method)

(i) With a proper $R > 0$, when $\gamma \asymp \sqrt{p + \log T}$: up to log-terms, w.h.p.

$$\max_{t \in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \Bigg(\underbrace{\sqrt{\frac{pr}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{h}_{\boldsymbol{A}^{(t)*} \text{ not equal}} + \underbrace{\sqrt{\frac{r}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\sqrt{\epsilon r}}_{\text{outlier tasks}} \Bigg) \wedge \underbrace{\sqrt{\frac{p}{n}}}_{\text{single-task rate}}$$

(ii) If tasks in S^c also follow linear model: up to log-terms, w.h.p.

$$\max_{t \in S^c} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p}{n}}.$$

- Polynomial-time algorithm
- Better than the bound of pERM when $\epsilon = 0$
- Minimax optimal when $\epsilon = 0$

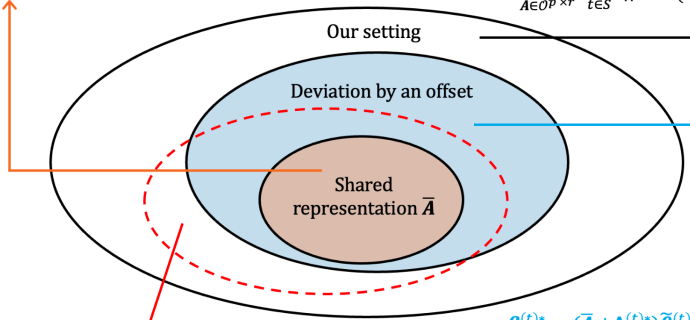
Comparing with exisiting methods

$\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}\boldsymbol{\theta}^{(t)*}, t \in S = [T]$ (Du et al., 2020; Thekumparampil et al., 2021; Tripuraneni et al., 2021)

$\max_{t\in[T]} ||\boldsymbol{\theta}^{(t)*}||_2 \lesssim 1$

$\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}, \mathrm{t} \in S \subseteq [T]$

$\min_{\overline{\boldsymbol{A}}\in\mathcal{O}^{p\times r}} \max_{t\in S} ||\boldsymbol{A}^{(t)*}\left(\boldsymbol{A}^{(t)*}\right)^{\top} - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^{\top}||_2 \leq h$



$\boldsymbol{\beta}^{(t)*} = (\overline{\boldsymbol{A}} + \boldsymbol{\Delta}^{(t)*})\widetilde{\boldsymbol{\theta}}^{(t)*}, t \in S = [T]$

$\max_{t\in[T]} ||\widetilde{\boldsymbol{\theta}}^{(t)*}||_2 \lesssim 1$ (Chua et al., 2021)

$\boldsymbol{\beta}^{(t)*} = \overline{\boldsymbol{A}}\overline{\boldsymbol{\theta}}^{(t)*} + \boldsymbol{\delta}^{(t)*}, \mathrm{t} \in S = [T]$

$\max_{t\in[T]} ||\overline{\boldsymbol{\theta}}^{(t)*}||_2 \lesssim 1$ (Duan and Wang, 2023)

The regime for rate comparison in Table 1, our setting with $\max_{t\in S} ||\boldsymbol{\theta}^{(t)*}||_2 \lesssim 1, S \subseteq [T]$ and $\sqrt{|S|^{-1}\sum_{t\in S}||\boldsymbol{\theta}^{(t)*}||_2^2} \gtrsim 1$

Review: Comparison of estimation error

Regime	Algorithm	$\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)}-\boldsymbol{\beta}^{(t)*}\|_2$	Optimal ($\epsilon=0$)?	Poly-time?
$h=\epsilon=0$	ERM (Du et al., 2020; Tripuraneni et al., 2021)	$r\sqrt{\frac{p}{nT}}+r\sqrt{\frac{1}{n}}$	No	No
	MoM (Tripuraneni et al., 2021)	$r\sqrt{\frac{p}{nT}}+\sqrt{\frac{r}{n}}$	No	Yes
	AltMinGD (Thekumparampil et al., 2021)	$r\sqrt{\frac{p}{nT}}+\sqrt{\frac{r}{n}}$	No	Yes
$h\neq 0$	AdaptRep (Chua et al., 2021)	$\left[r\sqrt{\frac{p}{nT}}+r\sqrt{\frac{1}{n}}+\sqrt{rh}(\frac{p}{n})^{1/4}\right]\wedge\sqrt{\frac{p}{n}}$	No	No
$\epsilon=0$	ARMUL (Duan and Wang, 2023)	$\left(r\sqrt{\frac{p}{nT}}+r\sqrt{\frac{1}{n}}+rh\right)\wedge\left(r\sqrt{\frac{p}{n}}\right)$	No	No
$h,\epsilon\neq 0$	pERM (Algorithm 1)	$\left(r\sqrt{\frac{p}{nT}}+r\sqrt{\frac{1}{n}}+\sqrt{r}h+\epsilon\frac{r^{3/2}\sqrt{p}}{\sqrt{n}}\right)\wedge\sqrt{\frac{p}{n}}$	No	No
	Spectral (Algorithm 2)	$\left(\sqrt{\frac{pr}{nT}}+\sqrt{\frac{r}{n}}+h+\sqrt{\epsilon r}\right)\wedge\sqrt{\frac{p}{n}}$	Yes	Yes
	Single-task	$\sqrt{\frac{p}{n}}$	No	Yes
	Lower-bound	$\left(\sqrt{\frac{pr}{nT}}+\sqrt{\frac{r}{n}}+h+\epsilon\frac{r}{\sqrt{n}}\right)\wedge\sqrt{\frac{p}{n}}$	–	–

§3 Simulation Studies

Simulation Settings

- Generate sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, $t = 1 : T$, and

$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^\top \boldsymbol{\beta}^{(t)*} + \epsilon_i^{(t)}, \quad i = 1 : n,$$

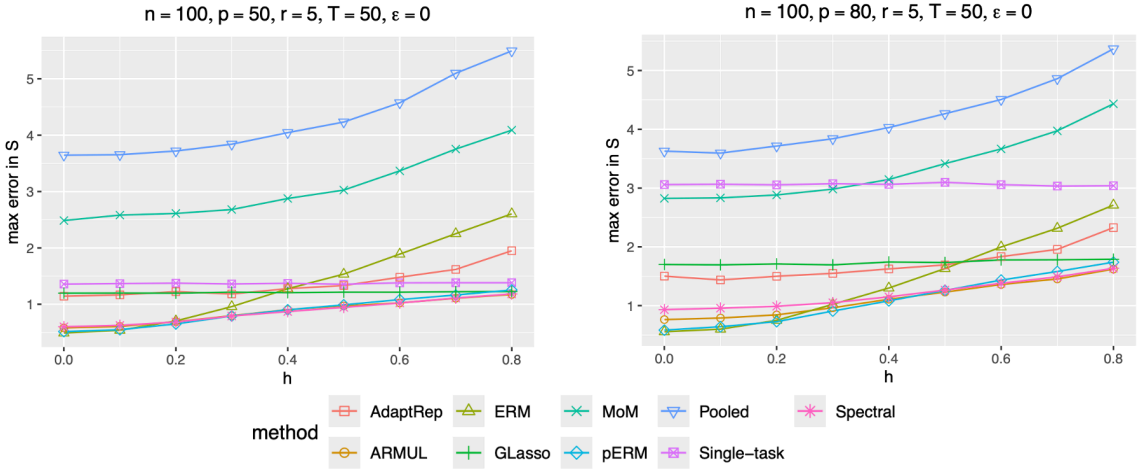
  where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p\times r}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$.

- We generate
  - $\boldsymbol{x}_i^{(t)} \overset{i.i.d.}{\sim} N(\boldsymbol{0}_p, \boldsymbol{I}_p)$
  - $\epsilon_i^{(t)} \overset{i.i.d.}{\sim} N(0, 1)$
  - $p \times r$ matrix $\boldsymbol{C}$ with i.i.d. standard normal entries.
- We defined $\overline{\boldsymbol{A}}$ as the first r columns of the left singular matrix of $\boldsymbol{C}$, $\widetilde{\boldsymbol{A}}^{(t)} = \overline{\boldsymbol{A}} + a^{(t)}(\boldsymbol{I}_{r\times r}, \boldsymbol{0}_{r\times(p-r)})^\top$, and $\boldsymbol{A}^{(t)*} = \widetilde{\boldsymbol{A}}^{(t)}[(\widetilde{\boldsymbol{A}}^{(t)})^\top \widetilde{\boldsymbol{A}}^{(t)}]^{-1}(\widetilde{\boldsymbol{A}}^{(t)})^\top$ for $t \in [T]$, where $a^{(t)}$'s are i.i.d. sampled from Unif$([-h, h])$.
- We generated each coordinate of $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$ from Unif$([-2, 2])$ independently.
- 100 replicates

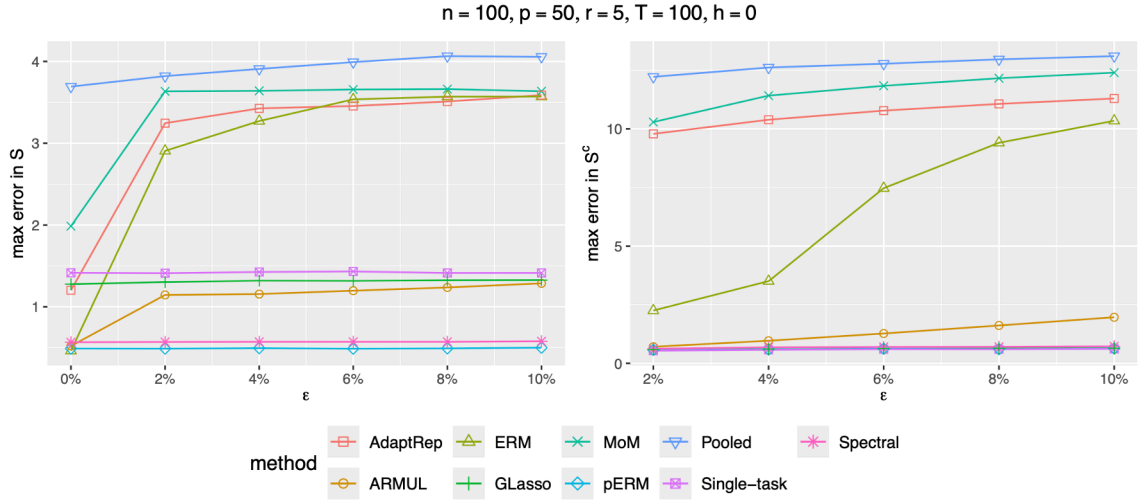
Implementation and benchmarks

- We used the Adam solver (Kingma and Ba, 2015) in PyTorch to solve the optimization problems
- Our methods:
  - Penalized ERM
  - Spectral method
- Benchmarks:
  - Single-task regression
  - Pooled regression (Crammer et al., 2008)
  - ERM (Du et al., 2020; Tripuraneni et al., 2021)
  - Method-of-moments (MoM) (Tripuraneni et al., 2021)
  - ARMUL (Duan and Wang, 2023)
  - AdaptRep (Chua et al., 2021)
  - Group Lasso (GLasso) (Lounici et al., 2011)

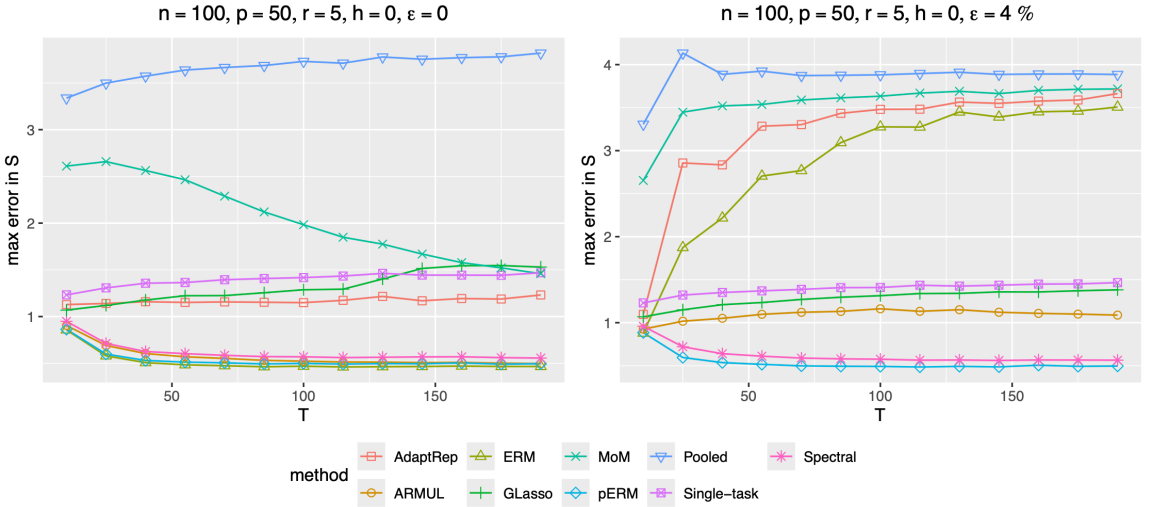
Different heterogeneity parameter h



Different contamination proportion ϵ



Different number of tasks T



§4 Human Activity Recognition Dataset Analysis

Human Activity Recognition Dataset

- The data come from **30 volunteers** performing six activities (walking, walking upstairs, walking downstairs, sitting, standing, and lying) while carrying a smartphone (Anguita et al., 2013).[2]
- Each observation contains $p = 561$ time- and frequency-domain features extracted from smartphone sensor signals.
- We treat each volunteer as a **separate task**, with sample sizes per task ranging from **281 to 409**.
- We focus on a **binary classification** problem: distinguishing *active postures* (walking, walking upstairs, walking downstairs, standing) from *inactive postures* (sitting, lying).
- For each task, in each of the **100 replicates**, 50% of the samples are used for training and 50% for testing.
- We report the **average test misclassification error rates** across all $T = 30$ **tasks** under varying values of r.

[2] Available at the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`.

Human Activity Recognition Dataset

- The data come from **30 volunteers** performing six activities (walking, walking upstairs, walking downstairs, sitting, standing, and lying) while carrying a smartphone (Anguita et al., 2013).[2]
- Each observation contains $p = 561$ time- and frequency-domain features extracted from smartphone sensor signals.
- We treat each volunteer as a **separate task**, with sample sizes per task ranging from **281 to 409**.
- We focus on a **binary classification** problem: distinguishing *active postures* (walking, walking upstairs, walking downstairs, standing) from *inactive postures* (sitting, lying).
- For each task, in each of the **100 replicates**, 50% of the samples are used for training and 50% for testing.
- We report the **average test misclassification error rates** across all $T = 30$ **tasks** under varying values of r.

[2] Available at the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`.

Mis-classification test error rates

r/Method	Single-task	Pooled	ERM	ARMUL	pERM	Spectral	GLasso
$r = 5$	1.66 (0.20)	1.79 (0.21)	1.62 (1.42)	2.12 (0.27)	**1.33 (0.23)**	1.85 (0.27)	1.44 (0.25)
$r = 10$	1.66 (0.20)	1.79 (0.21)	1.42 (0.23)	1.77 (0.23)	**1.25 (0.20)**	1.47 (0.18)	1.44 (0.25)
$r = 15$	1.66 (0.20)	1.79 (0.21)	1.36 (0.23)	1.68 (0.21)	**1.36 (1.07)**	1.50 (0.19)	1.44 (0.25)

Table: The average mis-classification test error rates (standard deviations) of different methods over $T = 30$ tasks with different r values. All values are in percentages.

§5 Take-away

Take-away

- **Robust and adaptive representation learning** is essential for integrating heterogeneous data across domains.
- **Theoretical guarantees** provide rigor and reliability even under contamination and unknown similarity.
- **Future directions:** nonlinear situation, scaling to multimodal and dynamic settings such as EHR, genomics, and imaging.

- *Tian, Y., Gu, Y., & Feng, Y. (2025). Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness. JMLR.*

My other works in transfer learning/multi-task learning

- **Transfer Learning in High-Dimensional GLMs (JASA, 2023)**
  Y. Tian & Y. Feng
- **Theory of Unsupervised Federated Learning (ICML, 2024)**
  Y. Tian, H. Weng, & Y. Feng
- **Unsupervised Multi-Task and Transfer Learning on GMMs (arXiv:2209.15224, 2024)**
  Y. Tian, H. Weng, L. Xia, & Y. Feng
- **Federated Transfer Learning with Differential Privacy (arXiv:2403.11343, 2024)**
  M. Li, Y. Tian, Y. Feng, & Y. Yu
- **GeoERM: Geometry-Aware Multi-Task Learning (arXiv:2505.02972, 2025)**
  A. Chen & Y. Feng

SHORT COURSE on TL

An Invitation to Contribute to *JASA Reviews*

Write a Review Article for *JASA*

- We welcome high-quality review articles on emerging and foundational areas of statistics and data science.
- A well-written review can define a field, highlight key advances, and inspire future research.
- As the Reviews Editor (2026-2028 term), I would be delighted to discuss potential ideas with you.

Have You Written a New Book?

- We regularly feature reviews of newly published research monographs and textbooks.
- If you have a recent book, we would be happy to invite a qualified reviewer.
- Please reach out to share your title or publisher details.

Let's work together to highlight and synthesize impactful contributions to our field.


tusind tak
ngiyabonga
谢谢
dakujem vám
dziękuję
ありがとう
merci
suksema
danke
thank
baie dankie
धन्यवाद
molte grazie
you
gracias
tak
obrigada
dank u
mahalo
obrigado
gràcies
teşekkür ederim
شكرا
tänan
təşəkkür edirə
tack så mycket

References I

D Anguita, A Ghio, L Oneto, X Parra, JL Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 437–442. CIACO, 2013.

Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

References II

Yaqi Duan and Kaizheng Wang. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039, 2023.

Peter J Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR: international conference on learning representations*, pages 1–15. ICLR US., 2015.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International conference on machine learning*, pages 3488–3498. PMLR, 2019.

Nikola Konstantinov, Elias Frantar, Dan Alistarh, and Christoph Lampert. On the sample complexity of adversarial multi-source pac learning. In *International Conference on Machine Learning*, pages 5416–5425. PMLR, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

References III

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations (ICLR)*. International Conference on Learning Representations, 2020.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4): 2164–2204, 2011.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

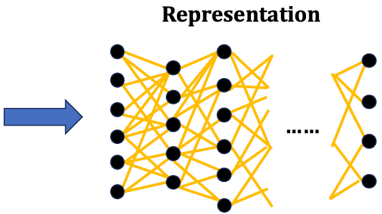
Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

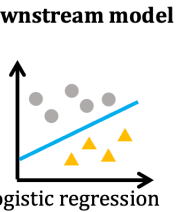
Representation transfer/multi-task learning

ImageNet
~1.3M training images

Representation transfer/multi-task learning

ImageNet
~1.3M training images

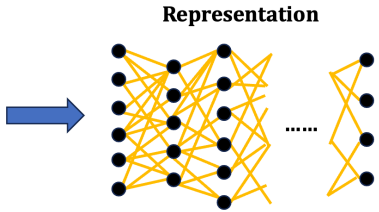




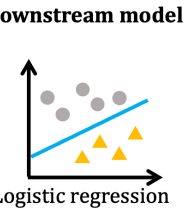
Representation transfer/multi-task learning

ImageNet
~1.3M training images

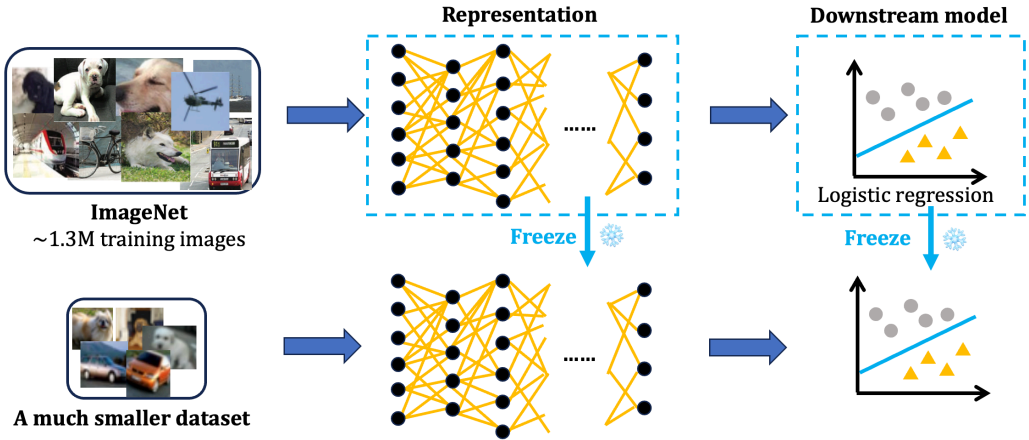
A much smaller dataset



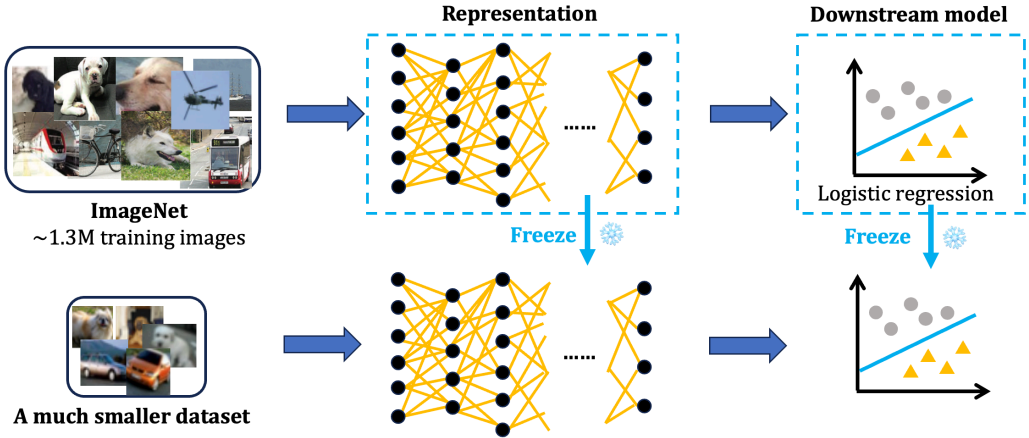
?



Representation transfer/multi-task learning

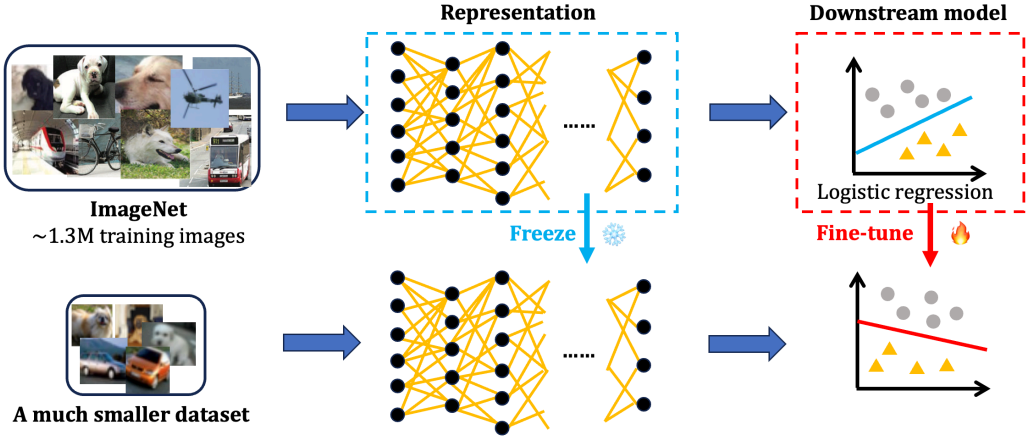


Representation transfer/multi-task learning

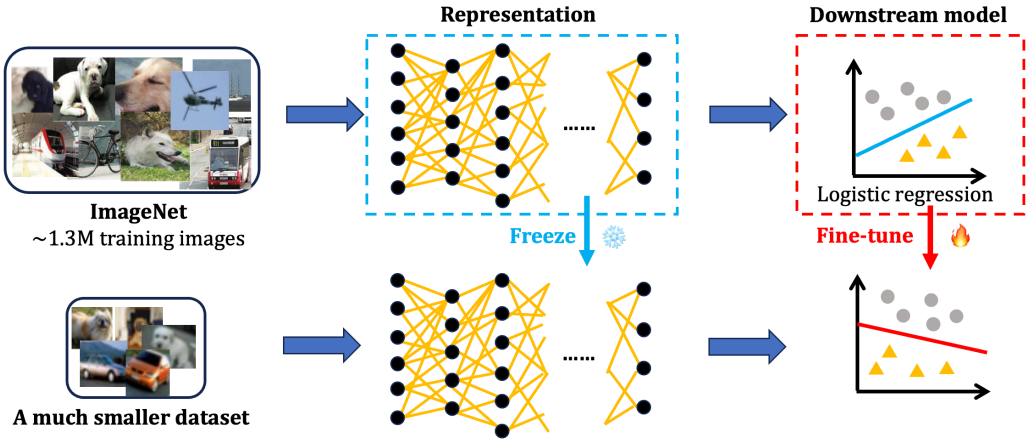


❄ representation + ❄ downstream models: not a good idea

Representation transfer/multi-task learning

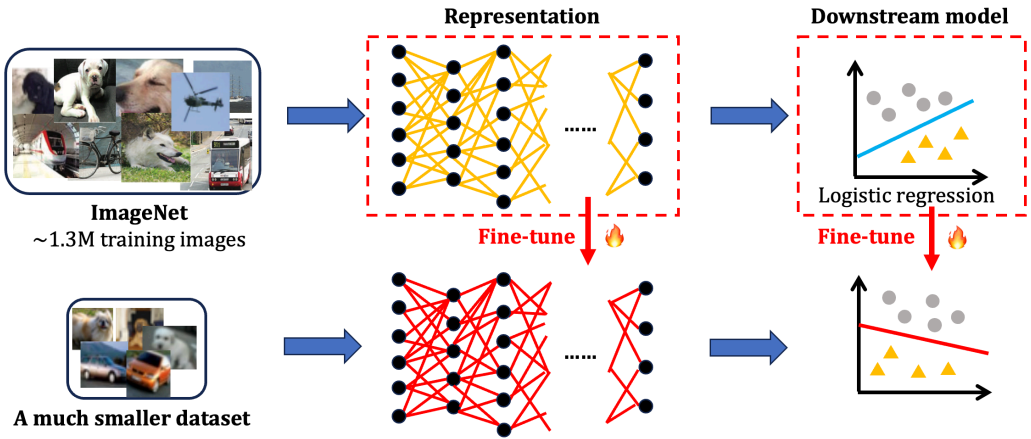


Representation transfer/multi-task learning

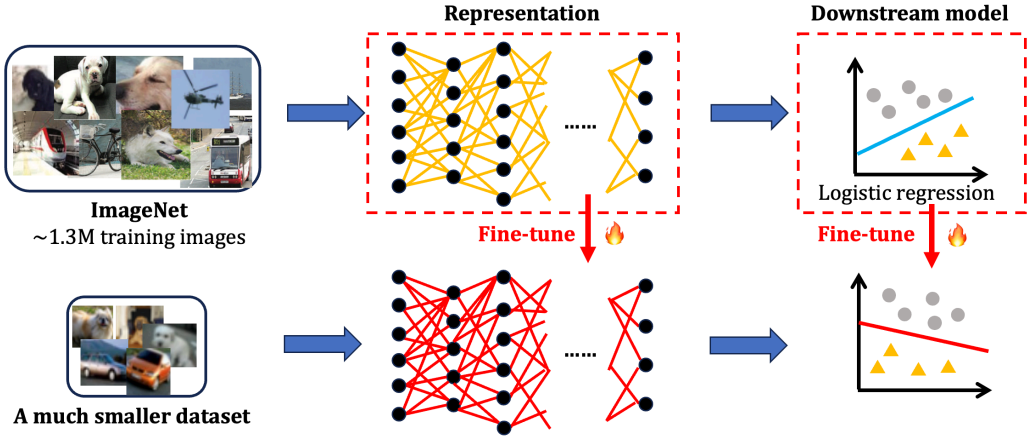


❄ representation + 🔥 downstream models (Donahue et al., 2014): can lead to negative transfer (Kornblith et al., 2019).

Representation transfer/multi-task learning



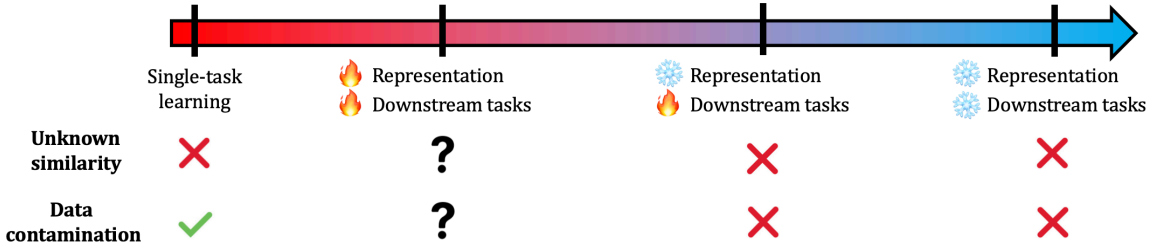
Representation transfer/multi-task learning



🔥 representation + 🔥 downstream models (Kornblith et al., 2019): sometimes unstable, "catastrophic forgetting" (Lee et al., 2020)

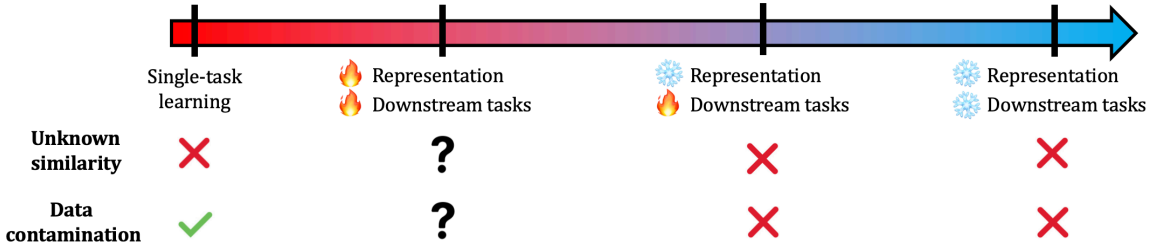
A quick summary

- Different approaches and two challenges:



A quick summary

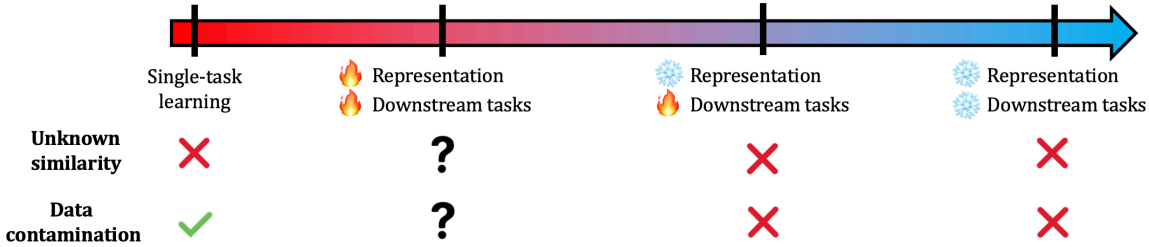
- Different approaches and two challenges:



- **Goal 1:** Show how **regularized fine-tuning** can replace both **?** with ✓

A quick summary

- Different approaches and two challenges:



- **Goal 1:** Show how **regularized fine-tuning** can replace both **?** with ✓
- **Goal 2:** Quantify the impact of two challenges on representation transfer/multi-task learning.