Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness

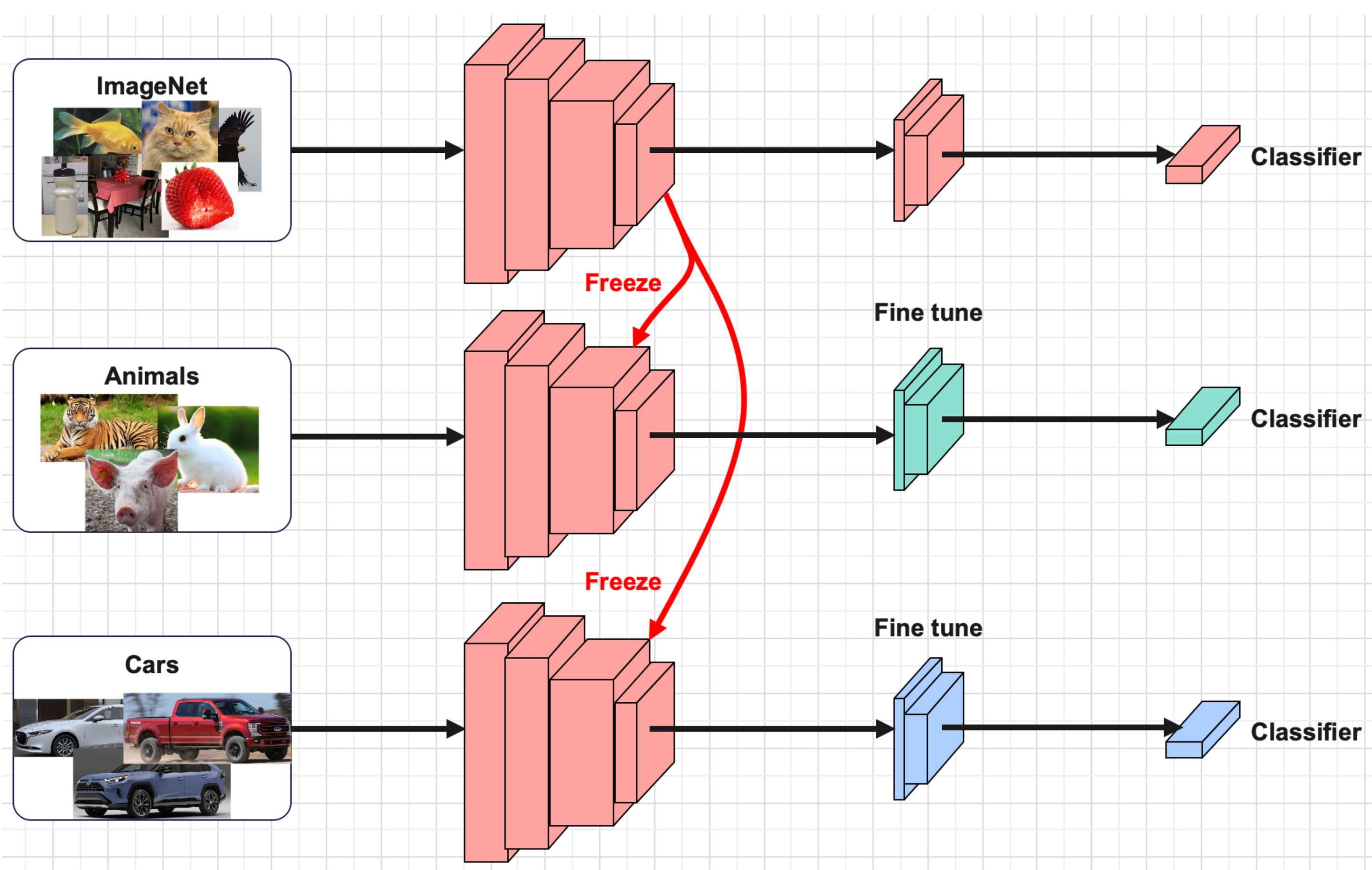
Ye Tian†, Yuqi Gu†, and Yang Feng‡

†Department of Statistics, Columbia University
‡Department of Biostatistics, School of Global Public Health, New York University

Columbia University

Introduction

Representation multi-task (MTL) and transfer learning (TL)



Lack of theoretical understanding

- The representations may NOT be the same! But most of theoretical studies impose such an assumption.
- As the number of tasks grow, there can be outlier tasks or adversarial attacks.

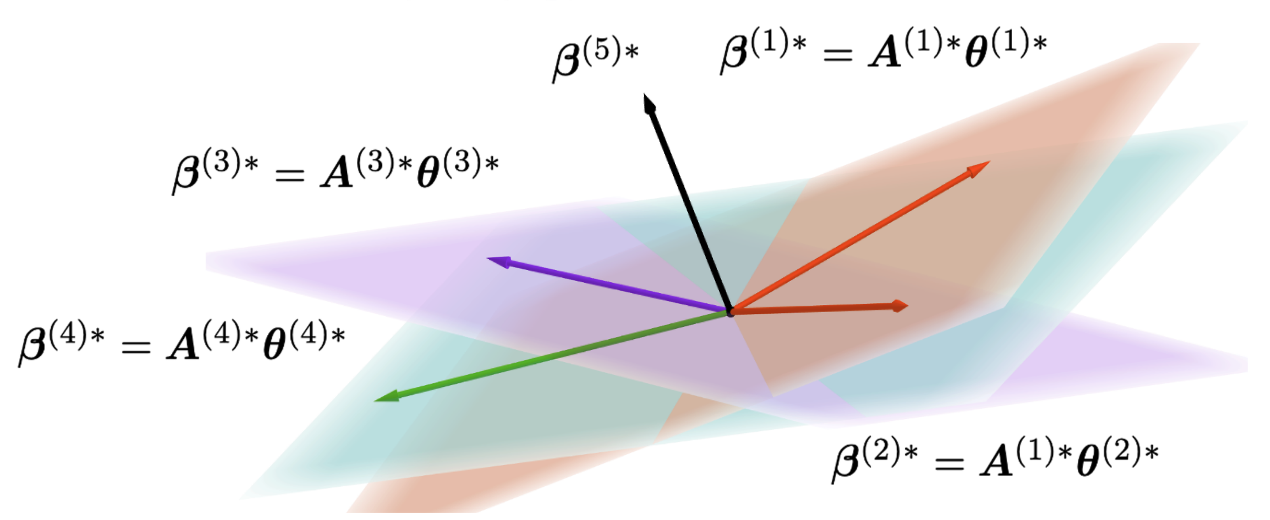
Multi-task Learning

Problem setting

- T tasks, data $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, with $\boldsymbol{x}_i^{(t)} \in \mathbb{R}^p$, $y_i^{(t)} \in \mathbb{R}$;
- There exists an unknown subset $S \subseteq [T]$, such that for all $t \in S$,

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

  where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathcal{O}^{p\times r} = \{\boldsymbol{A} \in \mathbb{R}^{p\times r} : \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_r\}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$, $r \leq p$, and $\{\epsilon_i^{(t)}\}_{i=1}^n$ are i.i.d. zero-mean sub-Gaussian $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$;
- $\min_{\overline{\boldsymbol{A}} \in \mathcal{O}^{p\times r}} \max_{t\in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \leq h$;
- Data $\{\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n\}_{t\notin S}$ from outlier tasks $\sim$ an arbitrary distribution $\mathbb{Q}_{S^c}$;
- **Goal:** jointly learning $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$



Two-step algorithm: (r known) Set $\lambda \asymp \sqrt{r(p+\log T)}$ and $\gamma \asymp \sqrt{p+\log T}$

- $\{\widehat{\boldsymbol{A}}^{(t)}\}_{t=1}^T, \{\widehat{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T, \widehat{\overline{\boldsymbol{A}}} \in \underset{\{\boldsymbol{A}^{(t)}\}_{t=1}^T, \{\boldsymbol{\theta}^{(t)}\}_{t=1}^T, \overline{\boldsymbol{A}}}{\arg\min} \Big\{ \sum_{t=1}^T \frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}_i^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \Big\}$;
- $\widehat{\boldsymbol{\beta}}^{(t)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \Big\{\frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}_i^{(t)})^T\boldsymbol{\beta}]^2 + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2\Big\}$ for $t \in [T]$

Assumptions

Notations: Coefficient matrix $\boldsymbol{B}_S^* \in \mathbb{R}^{p\times|S|}$, each column of which is a coefficient vector in $\{\boldsymbol{\beta}^{(t)*}\}_{t\in S}$. $\boldsymbol{\Sigma}^{(t)} := \mathbb{E}[\boldsymbol{x}^{(t)}(\boldsymbol{x}^{(t)})^T]$.

A.1 $\{\boldsymbol{x}_i^{(t)}\}_{i=1}^n$ are i.i.d. sub-Gaussian, and $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(t)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(t)}) \leq C$.
A.2 (Task diversity) $\max_{t\in S}\|\boldsymbol{\theta}^{(t)*}\|_2 \leq C < \infty$, and $\sigma_r(\boldsymbol{B}_S^*/\sqrt{T}) \geq \frac{c}{\sqrt{r}}$.
A.3 (Not many outlier tasks) $\frac{|S^c|}{T}r^{3/2} \leq$ a small constant c.

Upper bounds: When $n \gtrsim p + \log T$, $\forall S \subseteq [T]$ and an arbitrary $\mathbb{Q}_{S^c}$, w.h.p.,

$$\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left(r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \sqrt{\frac{p}{n}}\cdot\frac{|S^c|}{T}r^{3/2}\right) \wedge \sqrt{\frac{p+\log T}{n}}.$$

If outlier tasks in S^c also follow linear model (1), w.h.p.,

$$\max_{t\in[T]}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \sqrt{\frac{p+\log T}{n}}.$$

Lower bounds: $\forall\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T$, $\exists S \subseteq [T]$, $\{\boldsymbol{\beta}^{(t)}\}_{t\in S}$, $\mathbb{Q}_{S^c}$, w.p. $\geq 1/10$,

$$\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \sqrt{\frac{pr}{nT}} + h \wedge \sqrt{\frac{p+\log T}{n}} + \sqrt{\frac{r+\log T}{n}} + \frac{\epsilon r}{\sqrt{n}}.$$

If outlier tasks in S^c also follow linear model (1), $\forall\{\widehat{\boldsymbol{\beta}}^{(t)}\}_{t=1}^T$, $\exists\{\boldsymbol{\beta}^{(t)}\}_{t=1}^T$, w.p. $\geq 1/10$,

$$\max_{t\in[T]}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \gtrsim \sqrt{\frac{p+\log T}{n}}.$$

Transferring to New Tasks

Problem setting:

- Data $\{(\boldsymbol{x}_i^{(0)}, y_i^{(0)})\}_{i=1}^{n_0}$ from a new task, generated from model (1);
- $\max_{t\in S}\|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \boldsymbol{A}^{(0)*}(\boldsymbol{A}^{(0)*})^T\|_2 \leq h$;
- **Goal:** learning $\boldsymbol{\beta}^{(0)*}$

Two-step algorithm: (r known) Take $\widehat{\overline{\boldsymbol{A}}}$ from MTL algorithm, $\gamma \asymp \sqrt{p+\log T}$

- $\widehat{\boldsymbol{\theta}}^{(0)} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^r}{\arg\min}\Big\{\frac{1}{n_0}\sum_{i=1}^{n_0}[y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^T\widehat{\overline{\boldsymbol{A}}}\boldsymbol{\theta}]^2\Big\}$;
- $\widehat{\boldsymbol{\beta}}^{(0)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{n_0}\sum_{i=1}^{n_0}[y_i^{(0)} - (\boldsymbol{x}_i^{(0)})^T\boldsymbol{\beta}]^2 + \frac{\gamma}{\sqrt{n_0}}\|\boldsymbol{\beta} - \widehat{\overline{\boldsymbol{A}}}\widehat{\boldsymbol{\theta}}^{(0)}\|_2\Big\}$ for $t \in [T]$

Assumptions

A.4 $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^{n_0}$ are i.i.d. sub-Gaussian, and $0 < c \leq \lambda_{\min}(\boldsymbol{\Sigma}^{(0)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}^{(0)}) \leq C$.
A.5 $\|\boldsymbol{\theta}^{(0)*}\|_2 \leq C$.

Upper bound: When $n, n_0 \gtrsim p + \log T$, $\forall S \subseteq [T]$ and arbitrary $\mathbb{Q}_{S^c}$, w.h.p.,

$$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \lesssim \left(r\sqrt{\frac{p}{nT}} + \sqrt{r}h + \sqrt{r}\sqrt{\frac{r+\log T}{n}} + \sqrt{\frac{p}{n}}\cdot\frac{|S^c|}{T}r^{3/2}\right) \wedge \sqrt{\frac{p}{n_0}} + \sqrt{\frac{r}{n_0}}.$$

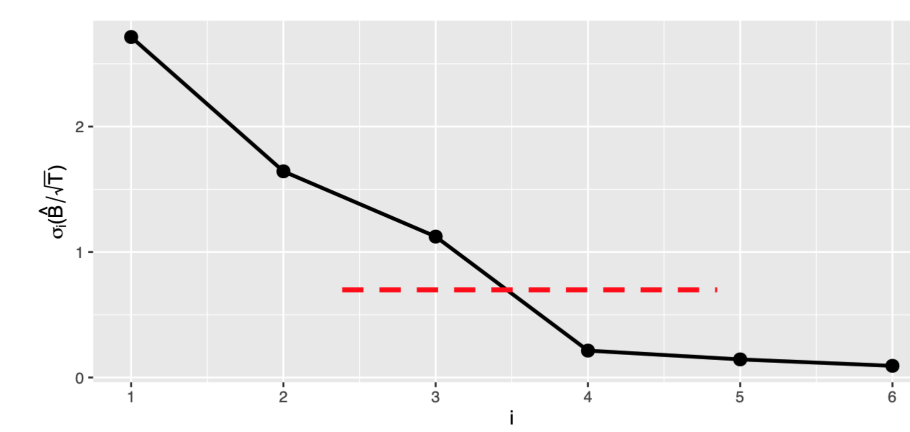
Lower bound: $\forall\widehat{\boldsymbol{\beta}}^{(0)}$, $\exists S \subseteq [T]$, $\{\boldsymbol{\beta}^{(t)}\}_{t\in\{0\}\cup S}$, $\mathbb{Q}_{S^c}$, w.p. $\geq 1/10$,

$$\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^{(0)*}\|_2 \gtrsim \left(\sqrt{\frac{pr}{nT}} + h\right) \wedge \sqrt{\frac{p}{n_0}} + \sqrt{\frac{r}{n_0}} + \frac{\epsilon r}{\sqrt{n}} \wedge \sqrt{\frac{1}{n_0}}.$$

Adaptation to Unknown Intrinsic Dimension

Intuition:

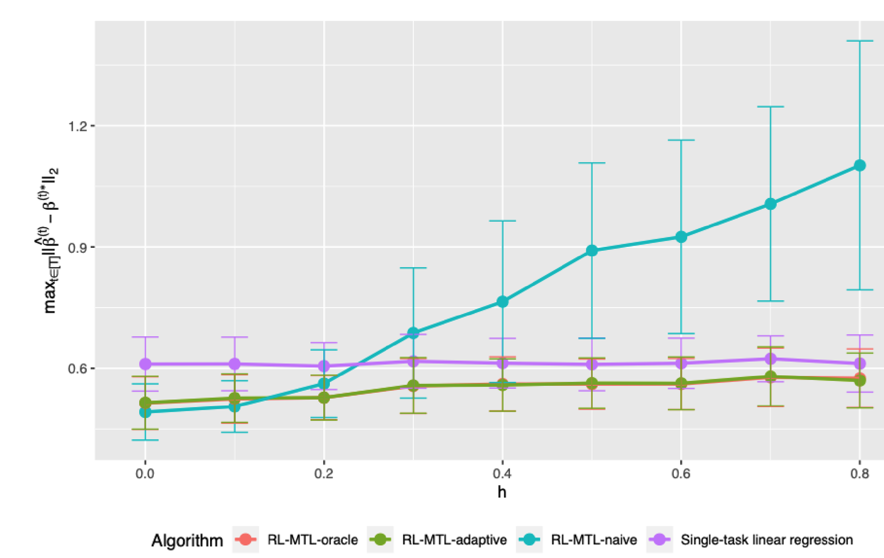
- It suffices to estimate r well when h and the proportion of outlier task $|S^c|/T$ are small;
- $\sigma_r(\boldsymbol{B}_S^*/\sqrt{T}) \gtrsim 1/\sqrt{r}$, $\sigma_{r+1}(\boldsymbol{B}_S^*/\sqrt{T}) \lesssim h \lesssim \sqrt{\frac{p+\log T}{nr}} \lesssim 1/\sqrt{r}$
- Do a thresholding to estimate r



Consistency: When $h \lesssim \sqrt{\frac{p+\log T}{rn}}$ (MTL) or $h \lesssim \sqrt{\frac{p}{rn_0}}$ (TL), under A.3, $\widehat{r} = r$ w.h.p.

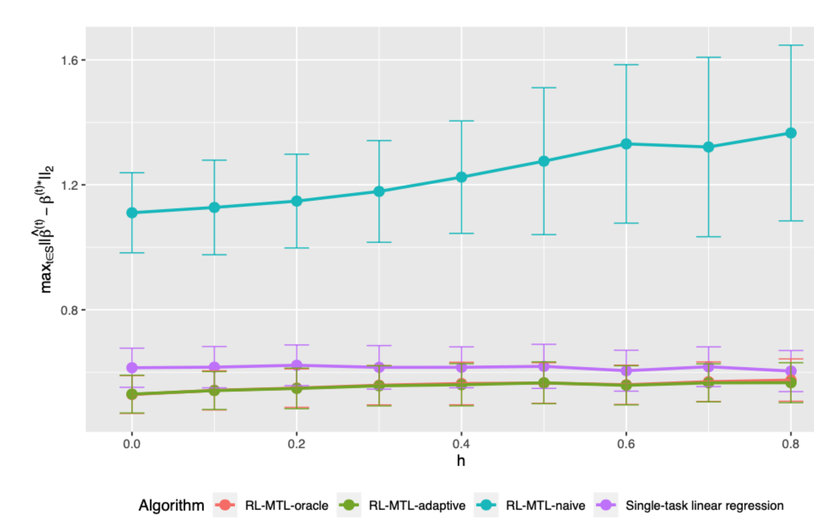
MTL Simulations

No outlier tasks: $n = 100$, $T = 6$, $p = 20$, $r = 3$

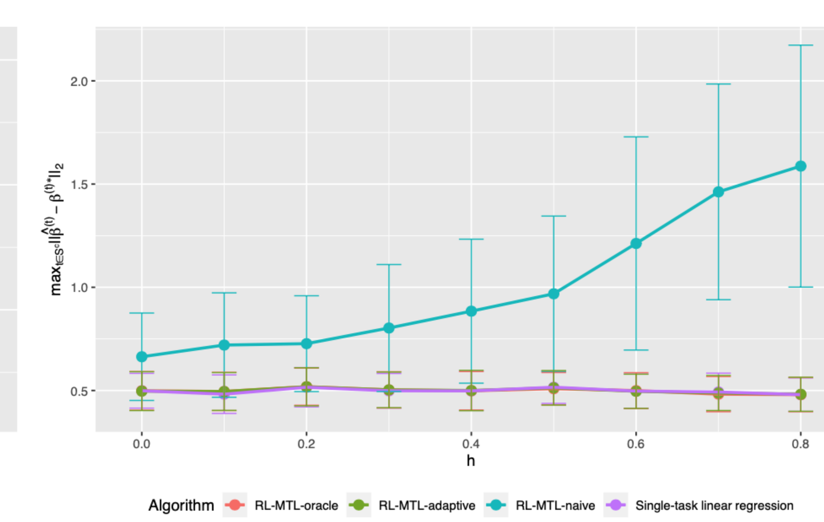


Estimation error $\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$

With outlier tasks: $n = 100$, $T = 7$, $|S^c| = 1$, $p = 20$, $r = 3$



(a) Estimation error $\max_{t\in S}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$



(b) Estimation error $\max_{t\in S^c}\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$