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

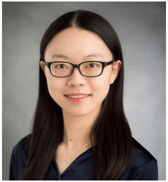
Ye Tian

Department of Statistics, Columbia University

2023 Berkeley-Columbia Meeting in Engineering and Statistics

April 20, 2023

Joint work with

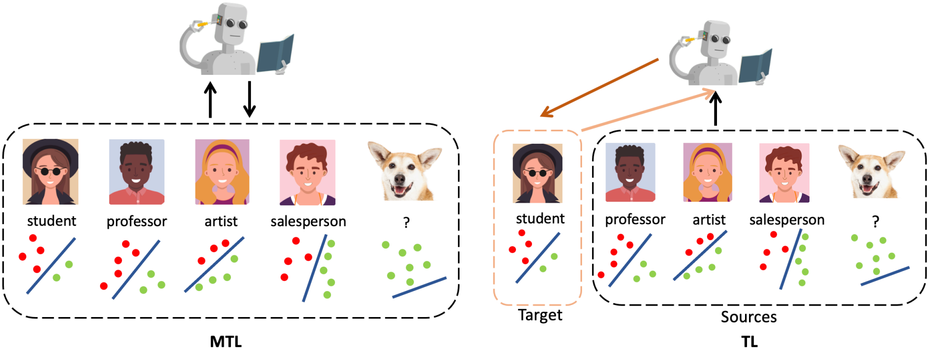


Yuqi Gu (Columbia stats) Yang Feng (NYU biostats)

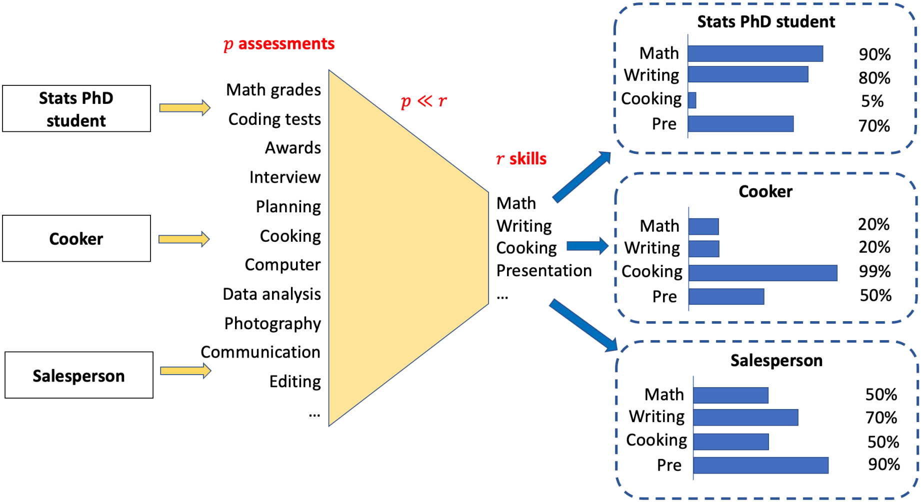
Greatest thanks to Yuqi and Yang!

Multi-task learning (MTL) and transfer learning (TL)

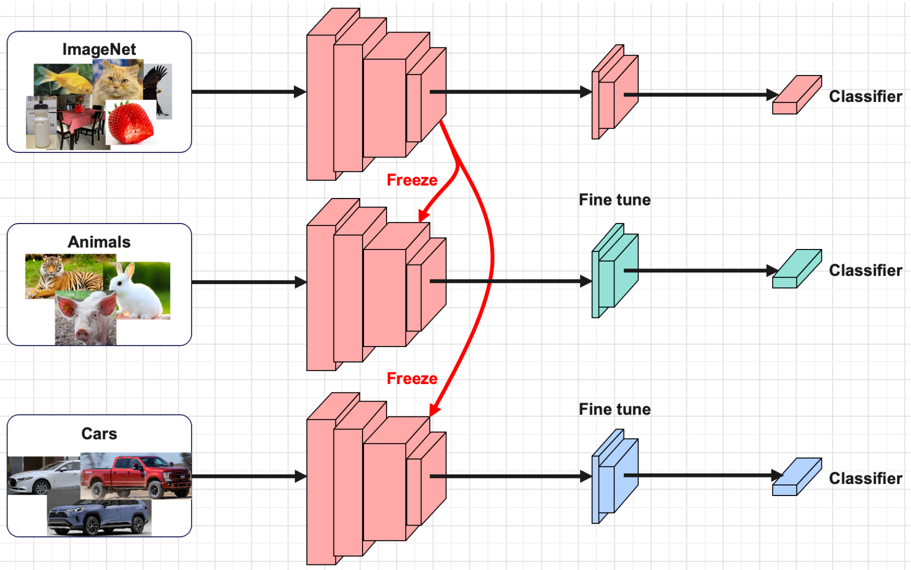
- Multi-task learning (MTL): Perform well on **all (or most)** tasks
- Transfer learning (TL): Perform well on the **target** task



Representation MTL and TL



In neural nets: freezing + fine tuning



A theoretical formulation

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

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

  where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^* \boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^* \in \mathbb{R}^{p \times r}$ with $(\boldsymbol{A}^*)^T \boldsymbol{A}^* = \boldsymbol{I}_{r \times r}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$.
- Theory was studied in Du et al. (2020); Tripuraneni et al. (2021)
- **Questions:**
  - What if the representations are **NOT** the same?
  - Outlier tasks?
- We suppose $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*} \boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}}} \max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \leq h.$$

  Sample $\{x_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T] \backslash S$ can be **arbitrarily** distributed. $\Longrightarrow$ **Outlier tasks**

A theoretical formulation

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

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

  where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^* \boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^* \in \mathbb{R}^{p \times r}$ with $(\boldsymbol{A}^*)^T \boldsymbol{A}^* = \boldsymbol{I}_{r \times r}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$.
- Theory was studied in Du et al. (2020); Tripuraneni et al. (2021)
- **Questions:**
  - What if the representations are **NOT** the same?
  - Outlier tasks?
- We suppose $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*} \boldsymbol{\theta}^{(t)*}$ with

$$\min_{\overline{\boldsymbol{A}}} \max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \leq h.$$

  Sample $\{x_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T] \backslash S$ can be **arbitrarily** distributed. $\Longrightarrow$ **Outlier tasks**

A theoretical formulation

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

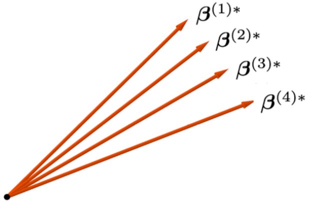
$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^T \boldsymbol{\beta}^{(t)*} + \epsilon_i^{(t)}, \quad i = 1:n,$$

  where $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^* \boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^* \in \mathbb{R}^{p\times r}$ with $(\boldsymbol{A}^*)^T \boldsymbol{A}^* = \boldsymbol{I}_{r\times r}$, $\boldsymbol{\theta}^{(t)*} \in \mathbb{R}^r$.
- Theory was studied in Du et al. (2020); Tripuraneni et al. (2021)
- **Questions:**
  - What if the representations are **NOT** the same?
  - Outlier tasks?
- We suppose $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*} \boldsymbol{\theta}^{(t)*}$ with

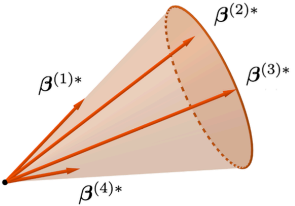
$$\min_{\overline{\boldsymbol{A}}} \max_{t\in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \leq h.$$

  Sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T]\backslash S$ can be **arbitrarily** distributed. $\Longrightarrow$ **Outlier tasks**

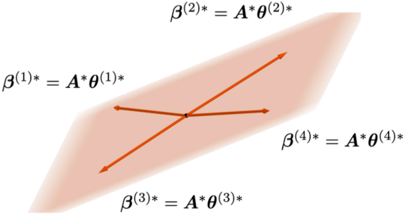
Different paradigms of MTL and TL



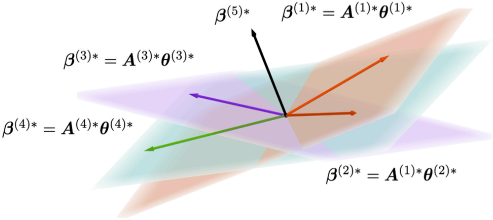
(a) Distance-based similarity [3, 19, 34, 50]



(b) Angle-based similarity [25]



(c) The same representation [18, 52]



(d) Similar representations with outliers (ours)

Problem review + algorithm

- **Problem:** Collected sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, $t = 1 : T$.
  - $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathbb{R}^{p\times r}$ with $\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T = \boldsymbol{I}_{r\times r}$:

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

  - Sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T]\backslash S$ can be **arbitrarily** distributed.
- **Two-step algorithm:** $\lambda \asymp \sqrt{r(p + \log T)}$, $\gamma \asymp \sqrt{p + \log T}$
  - $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \leftarrow$ Minimize $\sum_{t=1}^T \frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2$
  - $\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow$ Minimize $\frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{\beta}^{(t)}]^2 + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2$

Problem review + algorithm

- **Problem:** Collected sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, $t = 1 : T$.
  - $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathbb{R}^{p\times r}$ with $\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T = \boldsymbol{I}_{r\times r}$:

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

  - Sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T]\backslash S$ can be **arbitrarily** distributed.
- **Two-step algorithm:** $\lambda \asymp \sqrt{r(p+\log T)}$, $\gamma \asymp \sqrt{p+\log T}$
  - $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \leftarrow$ Minimize $\sum_{t=1}^T \frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2$
  - $\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow$ Minimize $\frac{1}{n}\sum_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{\beta}^{(t)}]^2 + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2$

Problem review + algorithm

- **Problem:** Collected sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from the t-th task, $t = 1:T$.
  - $\exists S \subseteq [T]$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$, $\boldsymbol{A}^{(t)*} \in \mathbb{R}^{p\times r}$ with $\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T = \boldsymbol{I}_{r\times r}$:
$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^T\boldsymbol{\beta}^{(t)*} + \epsilon_i^{(t)}, \quad i = 1:n, \quad t \in S.$$
  - Sample $\{\boldsymbol{x}_i^{(t)}, y_i^{(t)}\}_{i=1}^n$ from $t \in S^c = [T]\backslash S$ can be **arbitrarily** distributed.
- **Two-step algorithm:** $\lambda \asymp \sqrt{r(p+\log T)}$, $\gamma \asymp \sqrt{p+\log T}$
  - $\widehat{\boldsymbol{A}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\overline{\boldsymbol{A}}} \leftarrow$ Minimize $\sum\limits_{t=1}^T \frac{1}{n}\sum\limits_{i=1}^n [y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{A}^{(t)}\boldsymbol{\theta}^{(t)}]^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{A}^{(t)}(\boldsymbol{A}^{(t)})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2$
  - $\widehat{\boldsymbol{\beta}}^{(t)} \leftarrow$ Minimize $\frac{1}{n}\sum_{i=1}^n[y_i^{(t)} - (\boldsymbol{x}^{(t)})^T\boldsymbol{\beta}^{(t)}]^2 + \frac{\gamma}{\sqrt{n}}\|\boldsymbol{\beta}^{(t)} - \widehat{\boldsymbol{A}}^{(t)}\widehat{\boldsymbol{\theta}}^{(t)}\|_2$

Upper bounds

Assumptions:

- $\boldsymbol{x}_i^{(t)}$, $\epsilon_i^{(t)}$ sub-Gaussian
- $\max_{t \in S} \|\boldsymbol{\theta}^{(t)*}\|_2 \leq C < \infty$
- (Task diversity) Denote $\boldsymbol{B}_S^* = (\boldsymbol{\beta}^{(t)*})_{p \times |S|}$. Require $\sigma_r(\boldsymbol{B}_S^*) \gtrsim 1/\sqrt{r}$.
- (Not too many outlier tasks) $\epsilon := \frac{|S^c|}{T} \lesssim r^{-3/2}$

Upper bounds: Let $n \gtrsim \sqrt{p + \log T}$.

- $\forall t \in S$, w.p. $1 - o(1)$,

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left(\underbrace{r\sqrt{\frac{p}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{\sqrt{r}h}_{\boldsymbol{A}^{(t)*} \text{not equal}} + \underbrace{\sqrt{r}\sqrt{\frac{r + \log T}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\sqrt{\frac{p}{n}} \cdot \epsilon r^{3/2}}_{\text{outlier tasks}} \right) \wedge \underbrace{\sqrt{\frac{p + \log T}{n}}}_{\text{single-task rate}}$$

- If tasks in S^c also follow linear model: $\forall t \in S^c$, w.p. $1 - o(1)$,

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

Upper bounds

Assumptions:

- $\boldsymbol{x}_i^{(t)}$, $\epsilon_i^{(t)}$ sub-Gaussian
- $\max_{t\in S}\|\boldsymbol{\theta}^{(t)*}\|_2 \leq C < \infty$
- (Task diversity) Denote $\boldsymbol{B}_S^* = (\boldsymbol{\beta}^{(t)*})_{p\times|S|}$. Require $\sigma_r(\boldsymbol{B}_S^*) \gtrsim 1/\sqrt{r}$.
- (Not too many outlier tasks) $\epsilon := \frac{|S^c|}{T} \lesssim r^{-3/2}$

Upper bounds: Let $n \gtrsim \sqrt{p+\log T}$.

- $\forall t \in S$, w.p. $1-o(1)$,

$$\|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2 \lesssim \left(\underbrace{r\sqrt{\frac{p}{nT}}}_{\text{learn } \boldsymbol{A}^{(t)*}} + \underbrace{\sqrt{r}h}_{\boldsymbol{A}^{(t)*}\text{ not equal}} + \underbrace{\sqrt{r}\sqrt{\frac{r+\log T}{n}}}_{\text{learn } \boldsymbol{\theta}^{(t)*}} + \underbrace{\sqrt{\frac{p}{n}}\cdot \epsilon r^{3/2}}_{\text{outlier tasks}}\right) \wedge \underbrace{\sqrt{\frac{p+\log T}{n}}}_{\text{single-task rate}}$$

- If tasks in S^c also follow linear model: $\forall t \in S^c$, w.p. $1-o(1)$,

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

Lower bounds

Upper bounds: Let $n \gtrsim \sqrt{p + \log T}$.

- w.p. $1 - o(1)$,

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

- If tasks in S^c also follow the linear model: w.p. $1 - o(1)$,

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

Lower bounds:

- w.p. $\geq 1/10$,

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

- If tasks in S^c also follow the linear model: w.p. $\geq 1/10$,

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

Adaptation to unknown intrinsic dimension r

- Our algorithm requires r to be known in priori
  - $\widehat{\boldsymbol{A}}^{(t)} \in \mathbb{R}^{p \times r}$, $\widehat{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^{r}$
  - $\lambda \asymp \sqrt{r(p + \log T)}$
- Most prior works assume r is known: Ando et al. (2005); Chua et al. (2021); Collins et al. (2021); Du et al. (2020); Duan and Wang (2022); Duchi et al. (2022); Maurer et al. (2016); Tripuraneni et al. (2021)...

- **Fact:** When $\boldsymbol{A}^{(t)*} \equiv \boldsymbol{A}^*$, we have $\boldsymbol{B}^*_{p \times T} := (\boldsymbol{\beta}^{(t)*})_{t \in [T]} = \boldsymbol{A}^*_{p \times r} \boldsymbol{\Theta}_{r \times T}$ Hence $\sigma_i(\boldsymbol{B}^*) = 0$ for $i \geq r + 1$!
- Thresholding should work when h and outlier proportion ϵ are small
- What happens if h or ϵ is large? $\rightarrow$ No need to estimate r well

Adaptation to unknown intrinsic dimension r

- Our algorithm requires r to be known in priori
  - $\widehat{\boldsymbol{A}}^{(t)} \in \mathbb{R}^{p\times r}$, $\widehat{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^{r}$
  - $\lambda \asymp \sqrt{r(p+\log T)}$
- Most prior works assume r is known: Ando et al. (2005); Chua et al. (2021); Collins et al. (2021); Du et al. (2020); Duan and Wang (2022); Duchi et al. (2022); Maurer et al. (2016); Tripuraneni et al. (2021)...

- **Fact:** When $\boldsymbol{A}^{(t)*} \equiv \boldsymbol{A}^*$, we have $\boldsymbol{B}^*_{p\times T} := (\boldsymbol{\beta}^{(t)*})_{t\in[T]} = \boldsymbol{A}^*_{p\times r}\boldsymbol{\Theta}_{r\times T}$
  Hence $\sigma_i(\boldsymbol{B}^*) = 0$ for $i \geq r+1$!
- Thresholding should work when h and outlier proportion ϵ are small
- What happens if h or ϵ is large? $\rightarrow$ No need to estimate r well

Adaptation to unknown intrinsic dimension r

- Our algorithm requires r to be known in priori
  - $\widehat{\boldsymbol{A}}^{(t)} \in \mathbb{R}^{p \times r}$, $\widehat{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^r$
  - $\lambda \asymp \sqrt{r(p + \log T)}$
- Most prior works assume r is known: Ando et al. (2005); Chua et al. (2021); Collins et al. (2021); Du et al. (2020); Duan and Wang (2022); Duchi et al. (2022); Maurer et al. (2016); Tripuraneni et al. (2021)...

- **Fact:** When $\boldsymbol{A}^{(t)*} \equiv \boldsymbol{A}^*$, we have $\boldsymbol{B}^*_{p\times T} := (\boldsymbol{\beta}^{(t)*})_{t\in[T]} = \boldsymbol{A}^*_{p\times r}\boldsymbol{\Theta}_{r\times T}$
  Hence $\sigma_i(\boldsymbol{B}^*) = 0$ for $i \geq r+1$!
- Thresholding should work when h and outlier proportion ϵ are small
- What happens if h or ϵ is large? $\rightarrow$ No need to estimate r well

Adaptation to unknown intrinsic dimension r

- Our algorithm requires r to be known in priori
  - $\widehat{\boldsymbol{A}}^{(t)} \in \mathbb{R}^{p\times r}$, $\widehat{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^{r}$
  - $\lambda \asymp \sqrt{r(p+\log T)}$
- Most prior works assume r is known: Ando et al. (2005); Chua et al. (2021); Collins et al. (2021); Du et al. (2020); Duan and Wang (2022); Duchi et al. (2022); Maurer et al. (2016); Tripuraneni et al. (2021)...

- **Fact:** When $\boldsymbol{A}^{(t)*} \equiv \boldsymbol{A}^*$, we have $\boldsymbol{B}^*_{p\times T} := (\boldsymbol{\beta}^{(t)*})_{t\in[T]} = \boldsymbol{A}^*_{p\times r}\boldsymbol{\Theta}_{r\times T}$
  Hence $\sigma_i(\boldsymbol{B}^*) = 0$ for $i \geq r+1$!
- Thresholding should work when h and outlier proportion ϵ are small
- What happens if h or ϵ is large? $\rightarrow$ No need to estimate r well

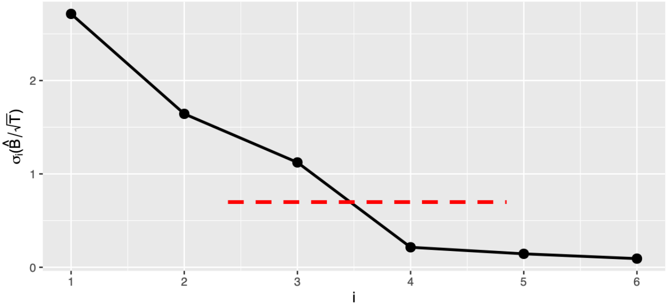
Adaptation to unknown intrinsic dimension r

- Our algorithm requires r to be known in priori
  - $\widehat{\boldsymbol{A}}^{(t)} \in \mathbb{R}^{p\times r}$, $\widehat{\boldsymbol{\theta}}^{(t)} \in \mathbb{R}^{r}$
  - $\lambda \asymp \sqrt{r(p+\log T)}$
- Most prior works assume r is known: Ando et al. (2005); Chua et al. (2021); Collins et al. (2021); Du et al. (2020); Duan and Wang (2022); Duchi et al. (2022); Maurer et al. (2016); Tripuraneni et al. (2021)...

- **Fact:** When $\boldsymbol{A}^{(t)*} \equiv \boldsymbol{A}^*$, we have $\boldsymbol{B}^*_{p\times T} := (\boldsymbol{\beta}^{(t)*})_{t\in[T]} = \boldsymbol{A}^*_{p\times r}\boldsymbol{\Theta}_{r\times T}$
  Hence $\sigma_i(\boldsymbol{B}^*) = 0$ for $i \geq r+1$!
- Thresholding should work when h and outlier proportion ϵ are small
- What happens if h or ϵ is large? $\rightarrow$ No need to estimate r well

Adaptation to unknown intrinsic dimension r

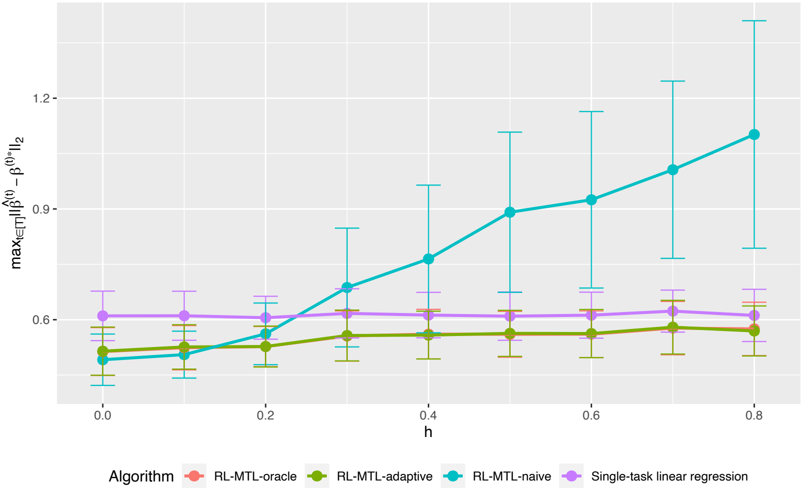
A simulation example: $p = 6$, $r = 3$



- Under almost the same conditions, we can **consistently** estimate r when $h \lesssim \sqrt{\frac{p+\log T}{n}}$ and $\epsilon \lesssim r^{-3/2}$
- Plug the estimated r into the previous algorithm

Simulation 1: No outlier tasks

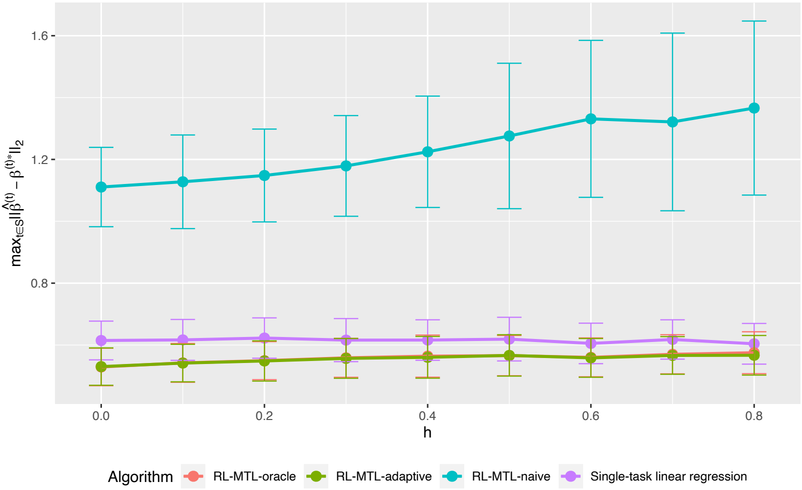
$T = 6$ tasks, $n = 100$, $p = 20$, $r = 3$, no outlier task



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

Simulation 2: With outlier tasks

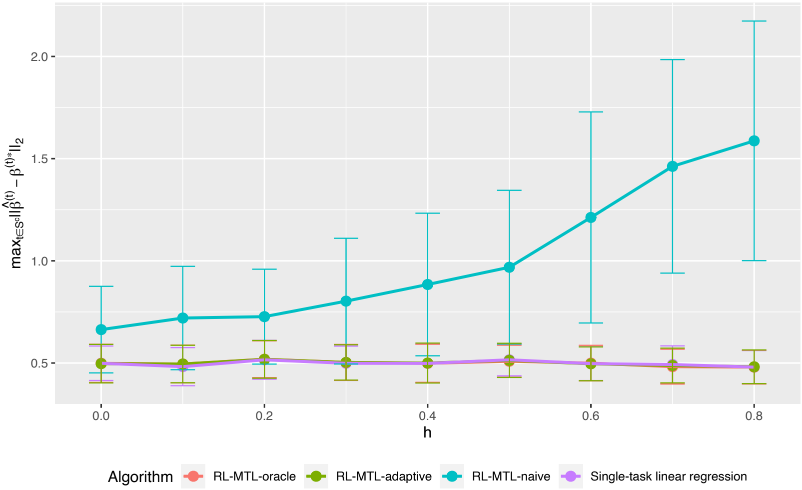
$T = 7$ tasks (1 outlier task), $n = 100$, $p = 20$, $r = 3$



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

Simulation 2: With outlier tasks

$T = 7$ tasks (1 outlier task), $n = 100$, $p = 20$, $r = 3$



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

Take-away

- Always freezing representations across tasks can lead to negative transfer
- We proposed an algorithm to learn from similar linear representations with outlier tasks, which
  - is adaptive to unknown similarity level h and intrinsic dimension r
  - is minimax optimal in a large regime
  - is robust to a small fraction $(\sim r^{-3/2})$ of outlier tasks
- Our paper on arXiv:

  Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness. arXiv preprint arXiv:2303.17765.

Thanks!

References I

Ando, R. K., Zhang, T., and Bartlett, P. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(11).

Chua, K., Lei, Q., and Lee, J. D. (2021). How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871--8884.

Collins, L., Hassani, H., Mokhtari, A., and Shakkottai, S. (2021). Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089--2099. PMLR.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2020). Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*.

Duan, Y. and Wang, K. (2022). Adaptive and robust multi-task learning. *arXiv preprint arXiv:2202.05250*.

References II

Duchi, J., Feldman, V., Hu, L., and Talwar, K. (2022). Subspace recovery from heterogeneous data with non-isotropic noise. *arXiv preprint arXiv:2210.13497*.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1--32.

Tripuraneni, N., Jin, C., and Jordan, M. (2021). Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434--10443. PMLR.