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

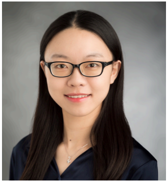
Ye Tian

Department of Statistics, Columbia University

Columbia Statistical Machine Learning Symposium 2023

April 7, 2023

# Joint work with



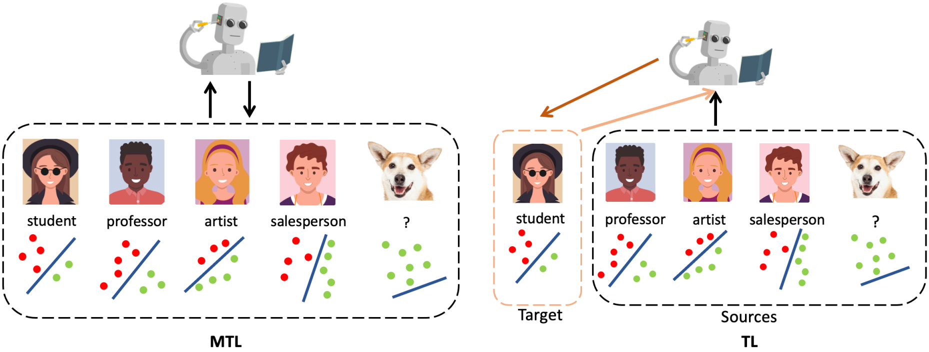
Yuqi Gu (Columbia stats)

Yang Feng (NYU biostats)

Greatest thanks to Yuqi and Yang!

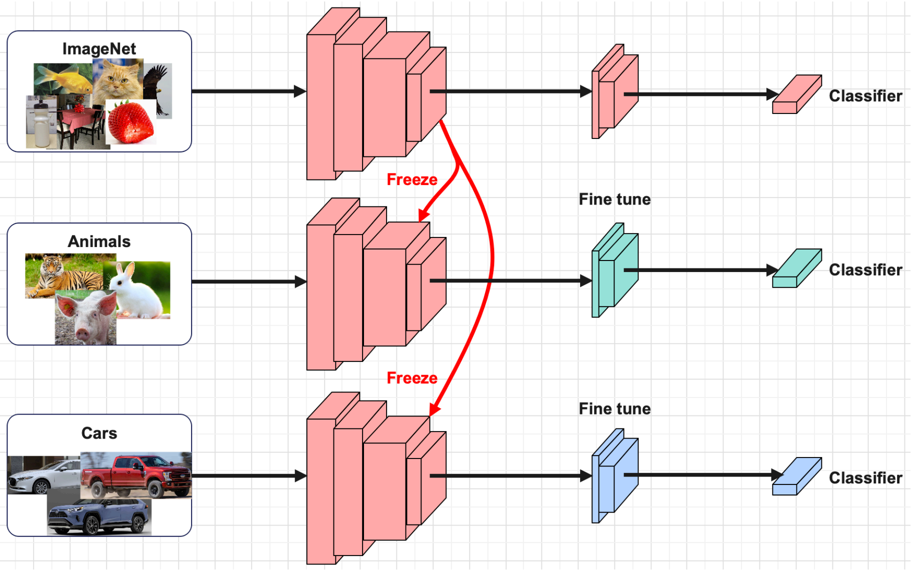
# Representation MTL and TL

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




- Representation MTL and TL
  - Learn a representation jointly and learn a low-dim parameter locally

# An example


ImageNet
Classifier
Freeze
Animals
Fine tune
Classifier
Freeze
Cars
Fine tune
Classifier

# A theoretical formulation

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

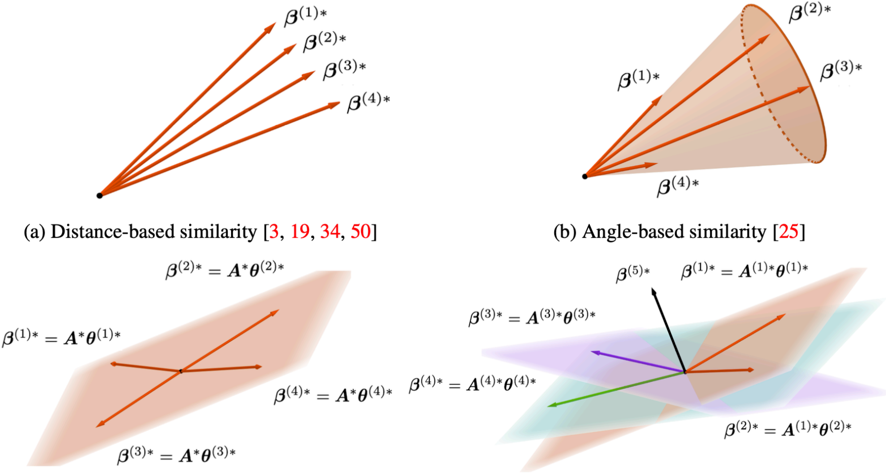
$$y_i^{(t)} = (\boldsymbol{x}_i^{(t)})^T \boldsymbol{\beta}^{(t)*} + \epsilon_i^{(t)}, \quad i = 1:n,$$

  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 the linear model




(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)

# Our contributions

Recall: $\min_{\overline{\boldsymbol{A}}} \max_{t \in S} \|\boldsymbol{A}^{(t)*}(\boldsymbol{A}^{(t)*})^T - \overline{\boldsymbol{A}}(\overline{\boldsymbol{A}})^T\|_2 \leq h$, $\boldsymbol{\beta}^{(t)*} = \boldsymbol{A}^{(t)*}\boldsymbol{\theta}^{(t)*}$,

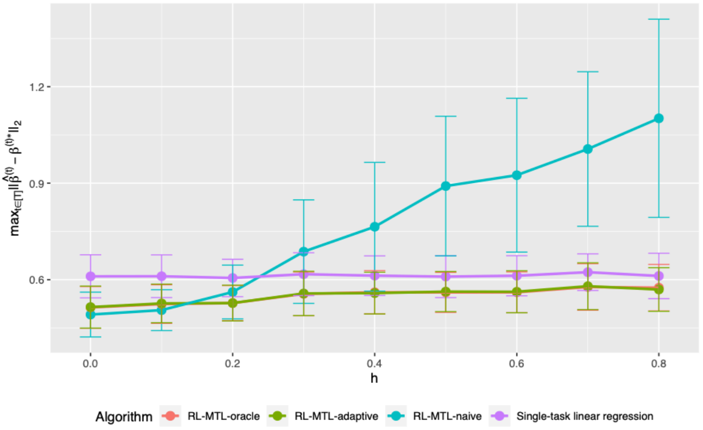
- Proposed algorithms (based on ERM + penalization) to solve this more general problem under MTL and TL setting
- Proved upper bounds of the algorithm
  - **Adaptivity**:

    (i) Never perform worse than single-task learning (No negative transfer)

    (ii) Benefit from similar representations (i.e., small $h$)
  - **Robustness**: robust to a small fraction of outlier tasks
- Proved lower bounds for this problem $\Longrightarrow$ our algorithms are **minimax** optimal in a large regime
- Proposed an algorithm to adapt to unknown intrinsic dimension $r$
- Our paper: Tian, Y., Gu, Y., & Feng, Y. (2023). Learning from Similar Representations: Adaptivity, Minimaxity, and Robustness. *arXiv preprint arXiv:2303.17765.*

# Simulation 1: No outlier tasks

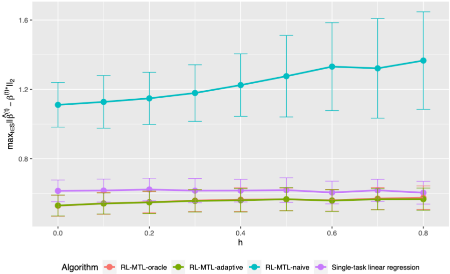
$T = 6$ tasks, $n = 100$, $p = 20$, $r = 3$, no outlier task



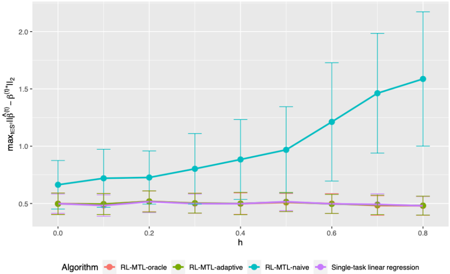

Estimation error $\max_{t\in S} \|\widehat{\boldsymbol{\beta}}^{(t)} - \boldsymbol{\beta}^{(t)*}\|_2$

# Simulation 2: With outlier tasks

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




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




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

Thanks!

# References I

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*.

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