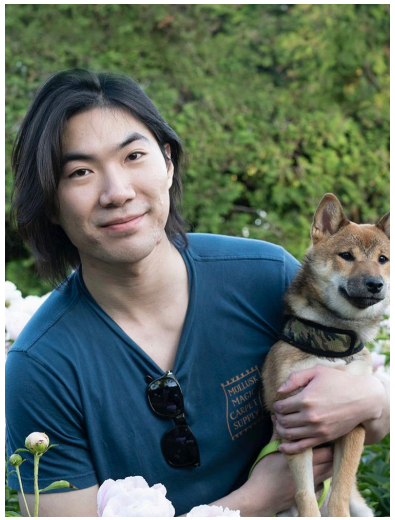
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

NEURAL INFORMATION PROCESSING SYSTEMS

Revisiting Scalarization in Multi-Task Learning: A *Theoretical* Perspective

Speaker: ***Yuzheng Hu***

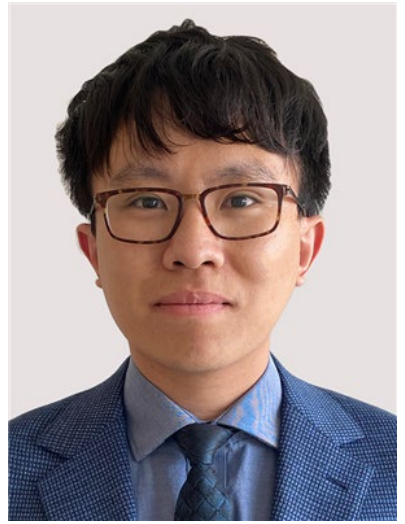


*Ruicheng Xian**

*Qilong Wu**

Qiuling Fan

Lang Yin

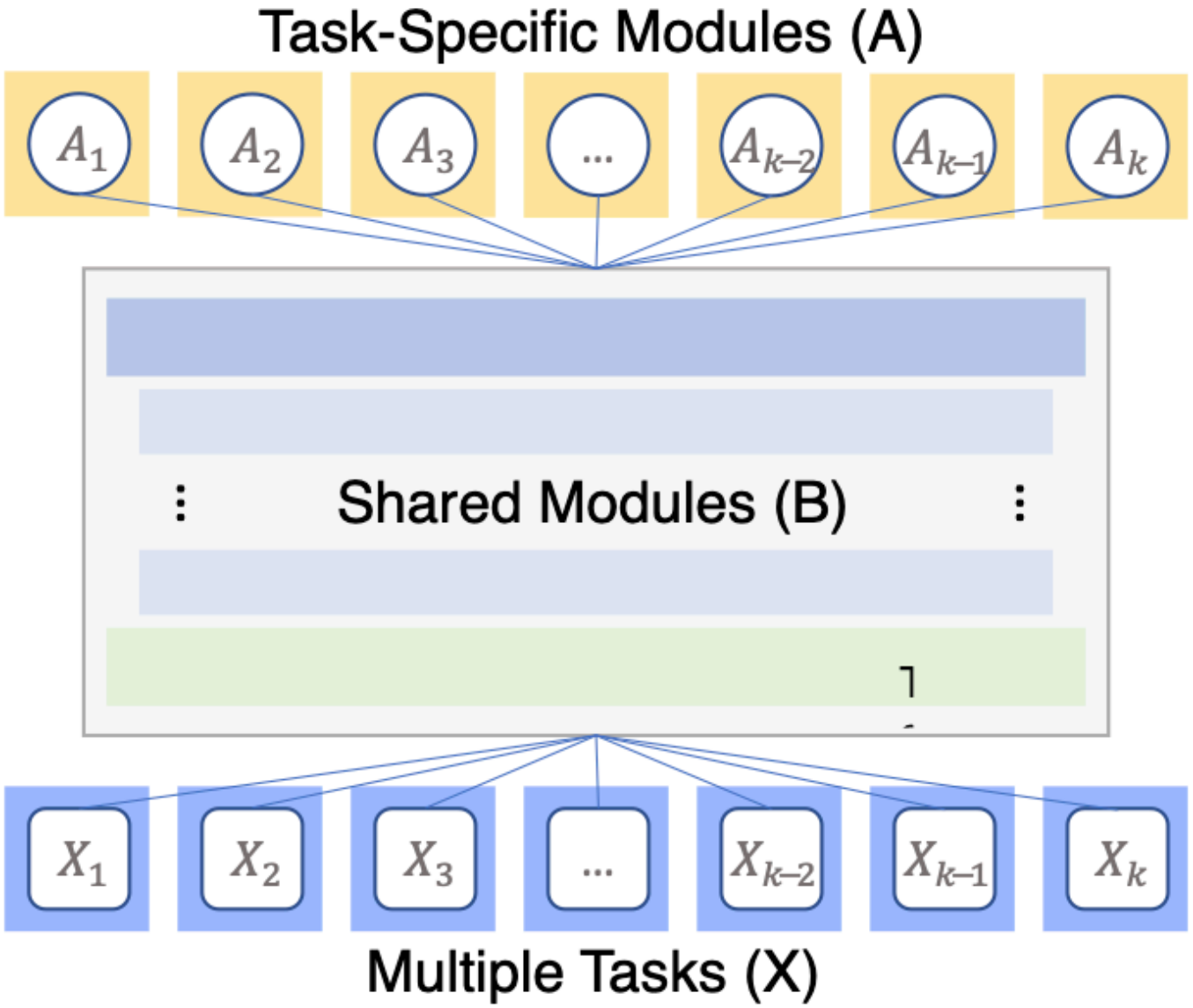


Han Zhao

*equal contribution. **Qilong is applying for PhD!**

https://arxiv.org/pdf/2308.13985.pdf

Multi-task learning



From https://hazyresearch.stanford.edu/blog/2020-03-01-multi_task_transfer_learning

- **Goal**: learn multiple *related* tasks simultaneously
- **Benefit**: improved *generalization*
- **Application**: autonomous driving

Two lines of research

- **Linear scalarization**
  - First choose a *fixed* set of non-negative weights $\{\lambda_i\}_{i\in[k]}$, then solve the scalar optimization problem:

$$\theta^* = \arg\min_{\theta} \sum_{i\in[k]} \lambda_i L_i(\theta)$$

  - Simple and scalable

- **Specialized multi-task optimizers (SMTOs)**
  - *Dynamic* multi-objective optimization
  - Goal: finding ***Pareto-optimal solutions***
  - MGDA[1], Gradient Surgery[2]…

[1] Désidéri, Jean-Antoine. "Multiple-gradient descent algorithm (MGDA) for multiobjective optimization." *Comptes Rendus Mathematique* 350.5-6 (2012): 313-318.
[2] Yu, Tianhe, et al. "Gradient surgery for multi-task learning." *Advances in Neural Information Processing Systems* 33 (2020): 5824-5836.

Heated debate

In Defense of the Unitary Scalarization for Deep Multi-Task Lea

Vitaly Kurin*
University of Oxford
vitaly.kurin@cs.ox.ac.uk

Alessa
Unive
adepalma@

Ilya Kostrikov
University of California, Berkeley
New York University

Shimon Whiteson
University of Oxford

Kurin et al., NeurIPS 2022

Do Current Multi-Task Optimization Methods in Deep Learning Even Help?

Derrick Xin*
Google Research
Mountain View, CA
dxin@google.com

Behrooz Ghorbani*
Google Research
Mountain View, CA
ghorbani@google.com

Ankush Garg
Google Research
Mountain View, CA
ankugarg@google.com

Orhan Firat
Google Research
Mountain View, CA
orhanf@google.com

Justin Gilmer
Google Research
Mountain View, CA
gilmer@google.com

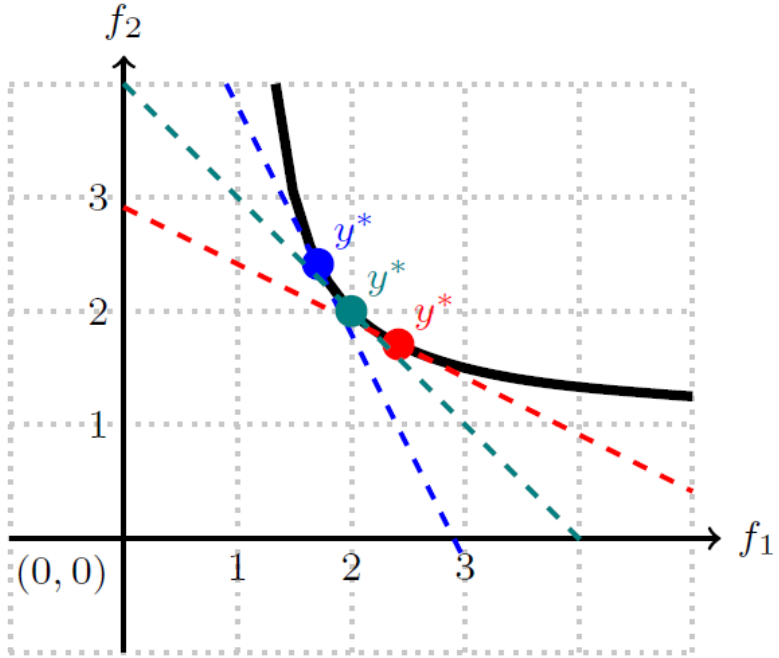
Xin et al., NeurIPS 2022

With proper choices of hyperparameters and regularization techniques, scalarization matches or even surpasses SMTOs.

Motivation

- Understand linear scalarization on the *representation* level
- *Full-exploration* problem:

  For every Pareto optimum v, does there exist a set of weights, such that the optimal solution of the linearly scalarized objective corresponds to v?



From Emmerich and Deutz, 2018

Theorem [BV04] *When the loss functions are convex, linear scalarization with proper weights can reach every Pareto optimum.*

What if the loss functions are non-convex?

[BV04] Boyd, Stephen P., and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Setting

- Two-layer multi-task linear network for **regression**: for task $i \in [k]$, the prediction is given by

$$f_i(x, W, a_i) = x^\top W a_i$$

$x \in \mathbb{R}^p$	input
$W \in \mathbb{R}^{p \times q}$	shared layer
$a_i \in \mathbb{R}^q$	task-specific head

- Shared input $X \in \mathbb{R}^{n \times p}$, target vector $y_i \in \mathbb{R}^n$, training loss for task i :

$$L_i(W, a_i) = \|XWa_i - y_i\|^2$$

Setting (cont.)

- **Over-parametrized regime** ($q \geq k$)
  - [WZR20] (linear case): The network has sufficient capacity to fit all tasks perfectly; the Pareto front reduces to a singleton $\{\vec{0}\}$ and can be achieved by linear scalarization with any choices of convex coefficients
  - True for general non-linear models (our work)

- **Under-parametrized regime** ($q < k$, our focus)
  - $q = 1$ --- extremely under-parametrized
  - $q = k - 1$ --- mildly under-parametrized

[WZR20] Wu, Sen, Hongyang R. Zhang, and Christopher Ré. "Understanding and Improving Information Transfer in Multi-Task Learning." *International Conference on Learning Representations*. 2019.

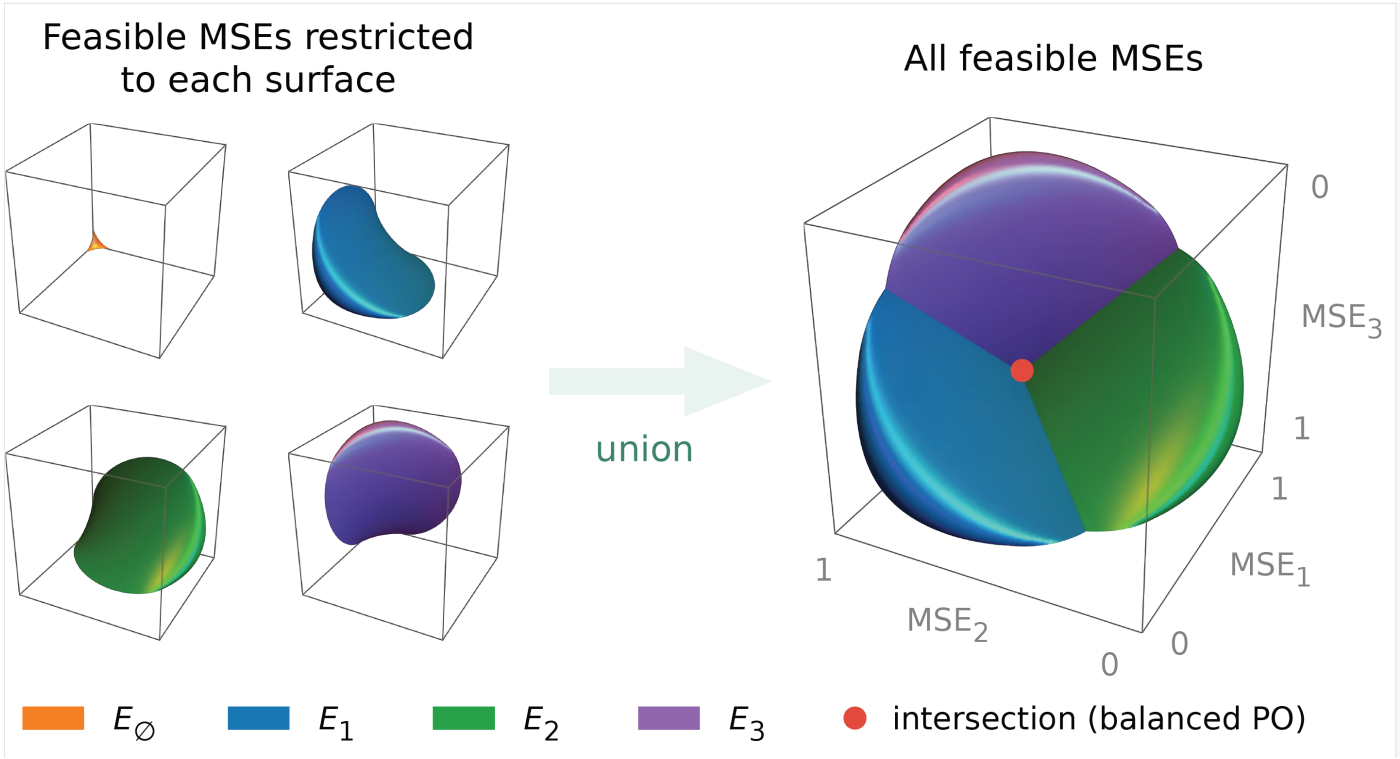
Main results

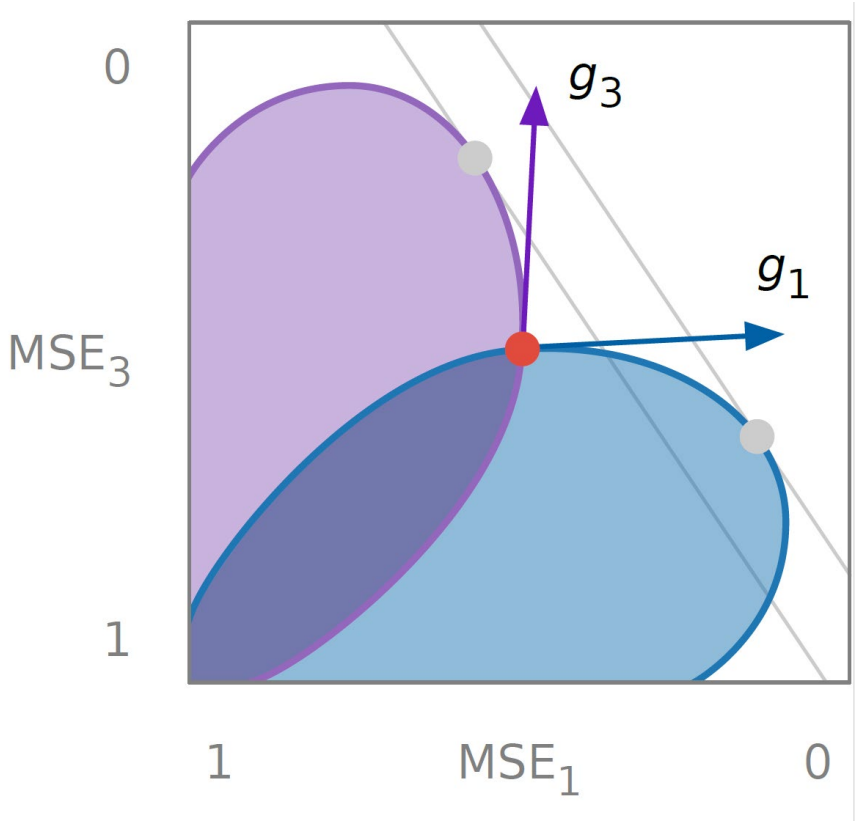
- Denote $\hat{y}_i = X(X^\top X)^\dagger X^\top y_i$ as the optimal linear predictor for task i
- Let $\hat{Y} = [\hat{y}_1, \cdots, \hat{y}_k] \in \mathbb{R}^{n \times k}$
- We develop *sufficient and necessary* conditions for full exploration

Theorem ($q = 1$): Linear scalarization is capable of fully exploring the Pareto front, if and only if $G := \hat{Y}^\top \hat{Y}$ is doubly non-negative, i.e., the inner products for all pairs of $\hat{y}_i$ and $\hat{y}_j$ are non-negative, up to negating the direction of some $\hat{y}_i$'s.

Theorem ($q = k - 1$): Linear scalarization is capable of fully exploring the Pareto front, if and only if $Q = G^{-1}$ is doubly non-negative, up to negating the direction of some $\hat{y}_i$'s.

Key observations

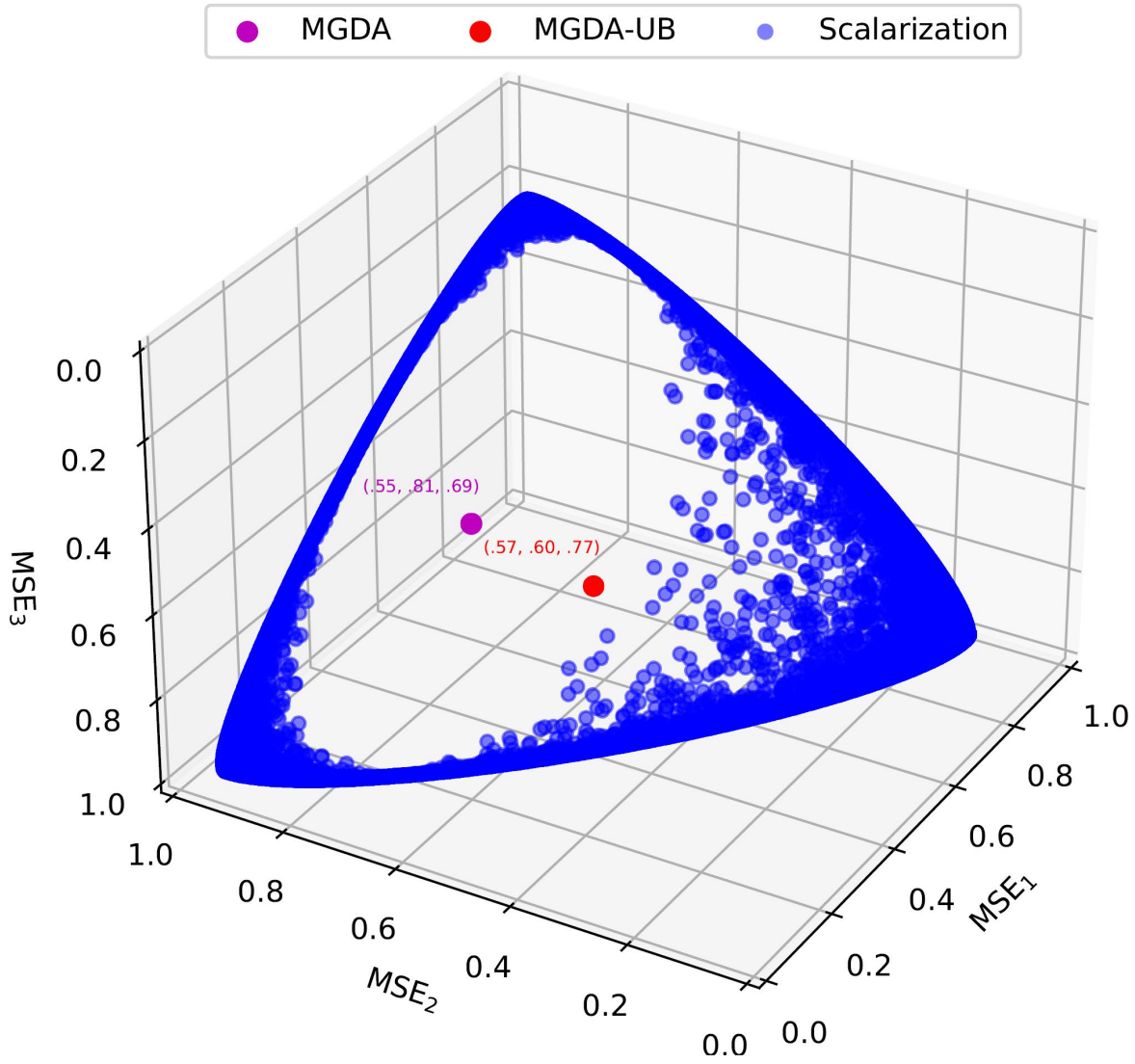




Multi-surface structure

Gradient disagreement

Experiment



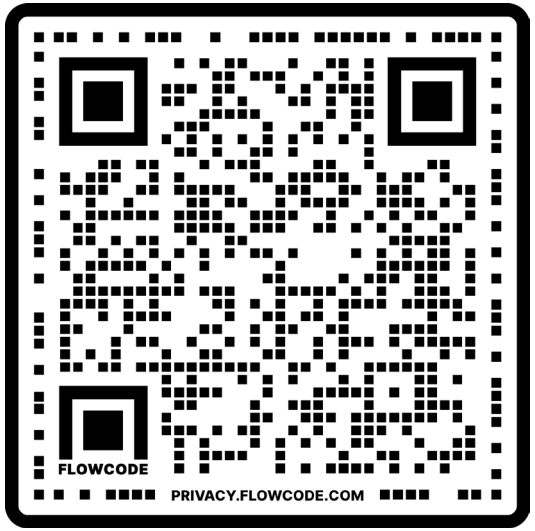
SMTOs are capable of finding *balanced* solutions, which are not achievable by linear scalarization

Takeaway

- We demonstrate a **representation limitation** of linear scalarization: it is generally not capable of full exploration for linear MTL
- On the empirical side, we reveal the potential of SMTOs in finding balanced solutions

We hope our work could:

- Foster a balanced development among linear scalarization and SMTOs
- Motivate the research community to develop a better theory explaining the empirical success of linear scalarization



Paper

Thank you!

Meet us at **Great Hall & Hall B1+B2 #1004**,
Dec 12th (Tuesday) 5:15 - 7:15 pm!