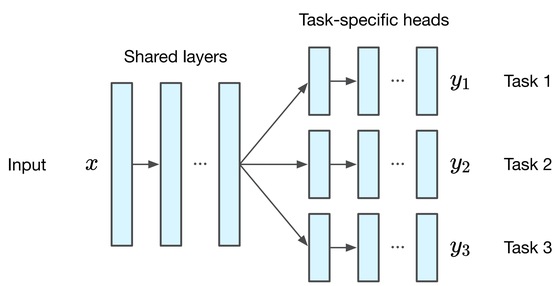
Revisiting Scalarization in Multi-Task Learning: A Theoretical Perspective

Yuzheng Hu Ruicheng Xian* Qilong Wu* Qiuling Fan Lang Yin Han Zhao
University of Illinois Urbana-Champaign

Background of Multi-Task Learning

Multi-task learning (MTL) aims to learn a joint model for multiple tasks for efficiency and generalization, and a popular approach is learning shared feature representations for all tasks.



The ideal goal is to find a model that minimizes the losses of all tasks, $L_1, \cdots, L_k$, simultaneously. Two typical approaches to train and optimize the joint model are:

- **Linear Scalarization.** Combine the losses with fixed weights, $\lambda_1, \cdots, \lambda_k \geq 0$, and solve the scalar optimization problem of $\min_\theta \sum_{i=1}^k \lambda_i L_i(\theta)$. This approach is simple and scalable.
- **Specialized Multi-Task Optimizers (SMTOs).** Cast as a multi-objective optimization problem [1, 5], and find a solution that is *Pareto optimal* among tasks (i.e., no task can be further improved without hurting some other task).

Motivation, and the Full-Exploration Problem

There are heated debates on whether SMTOs have an inherent advantage for MTL over scalarization, as recent empirical studies show that scalarization achieves better performance [2, 4, 3].

We study this problem theoretically by analyzing scalarization at the representation level—whether it is capable of *full exploration*:

> *For every Pareto optimum υ, does there exist a set of weights, $\lambda_1, \cdots, \lambda_k$, such that the optimal solution of the linearly scalarized objective corresponds to υ?*

Key Takeaways

In linear MTL, when model is under-parameterized:

- Scalarization has a representation limitation in that it generally cannot fully explore the Pareto frontier.
- Empirically, Scalarization cannot achieve balanced solutions found by SMTOs.

This limitation can be overcome with over-parametrization or randomization.

Conditions for Full-Exploration on Linear MTL

Let the k-task MTL regression problem be given by n sample pairs of input $x \in \mathbb{R}^p$ and target $y \in \mathbb{R}^k$, where y_i is the target of task i.

We train two-layer linear multi-task networks (linear MTL) under squared loss. The predicted target for task i on x is $x^\top W a_i$, where $W \in \mathbb{R}^{p \times q}$ is shared layer, and $a_i \in \mathbb{R}^q$ is task-specific head.

Results

We provide sufficient and necessary conditions for linear scalarization to fully explore the Pareto frontier. When model is:

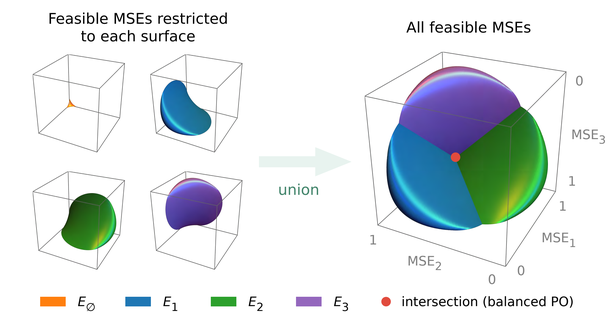
- **Over-parameterized ($q \geq k$).** Always capable, since the Pareto front reduces to a singleton in this case.
- **Mildly under-parameterized ($q = k - 1$).** Iff G^{-1} is doubly non-negative, up to negating some $\hat{y}_i$'s.
- **Extremely under-parameterized ($q = 1$).** Iff G is doubly non-negative (i.e., $\hat{y}_i^\top \hat{y}_j \geq 0$, $\forall i, j$), up to negating some $\hat{y}_i$'s.

Where, we defined:

- $X \in \mathbb{R}^{n \times p}$ the matrix of stacked inputs, $Y_i \in \mathbb{R}^n$ the concatenation of task i targets.
- $\widehat{Y}_i = X(X^\top X)^\dagger X^\top Y_i$ is optimal linear predictor for task i (OLS).
- $\widehat{Y} = [\widehat{Y}_1, \cdots, \widehat{Y}_k] \in \mathbb{R}^{n \times k}$, and $G = \widehat{Y}^\top \widehat{Y}$.

[1] Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. 2012.
[2] Kurin et al. In defense of the unitary scalarization for deep multi-task learning. 2022.
[3] Lin et al. Reasonable Effectiveness of Random Weighting: A Litmus Test for Multi-Task Learning. 2022.
[4] Xin et al. Do Current Multi-Task Optimization Methods in Deep Learning Even Help? 2022.
[5] Yu et al. Gradient surgery for multi-task learning. 2020.

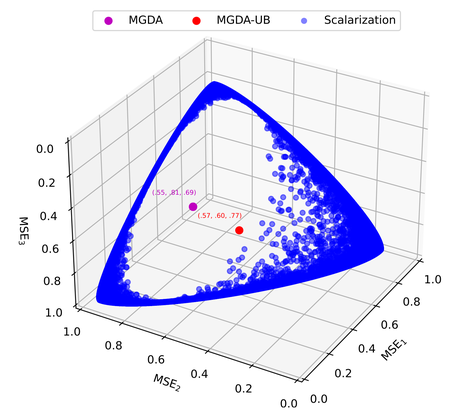
Visualizing the Feasible Region of Linear MTL



For linear MTL, in the under-parameterized settings,

- the feasible region has a *multi-surface structure*, which could be non-convex;
- when non-convex (i.e., conditions are not met), scalarization cannot reach the intersections of two or more surfaces,
- at those points, the gradients w.r.t. the surfaces disagree.

Scalarization vs. SMTOs for Linear MTL on SARCOS Dataset



SMTOs achieves red and purple points with balanced performance on all tasks. Scalarization only achieves blue points that are skewed towards a subset of tasks (under different weights).