Pareto Multi-Task Learning

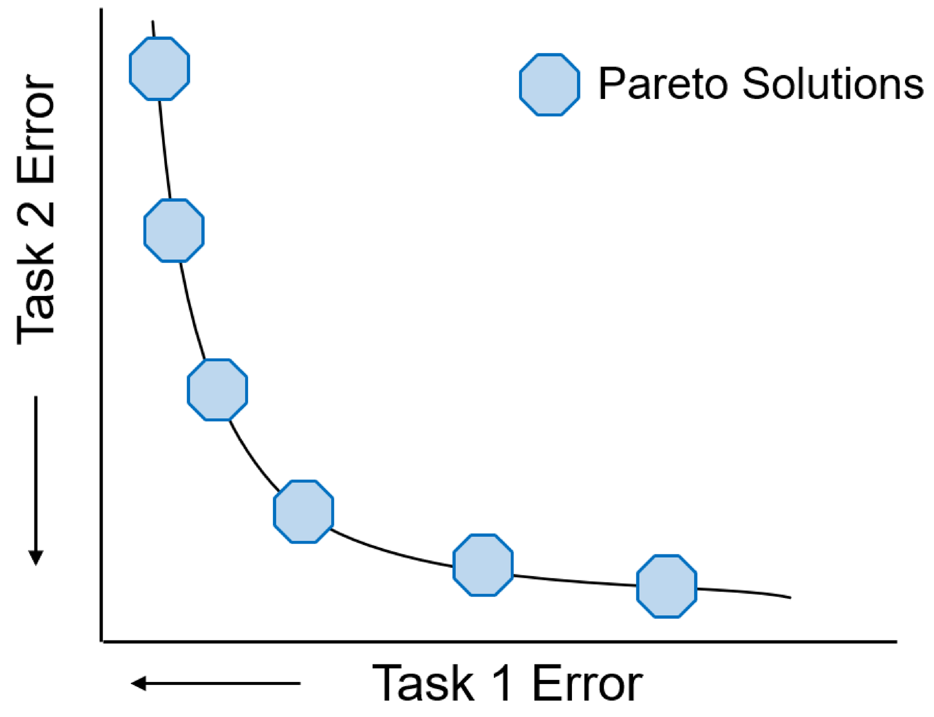
Xi Lin[1], Hui-Ling Zhen[1], Zhenhua Li[2], Qingfu Zhang[1], Sam Kwong[1]

[1]City University of Hong Kong, [2]Nanjing University of Aeronautics and Astronautics

xi.lin@my.cityu.edu.hk

Problem Definition and Contribution

Goal: To generate widely distributed Pareto solutions with different trade-offs for multi-task learning (MTL).



Then MTL practitioners can easily select their preferred solution(s) with different optimal trade-offs.

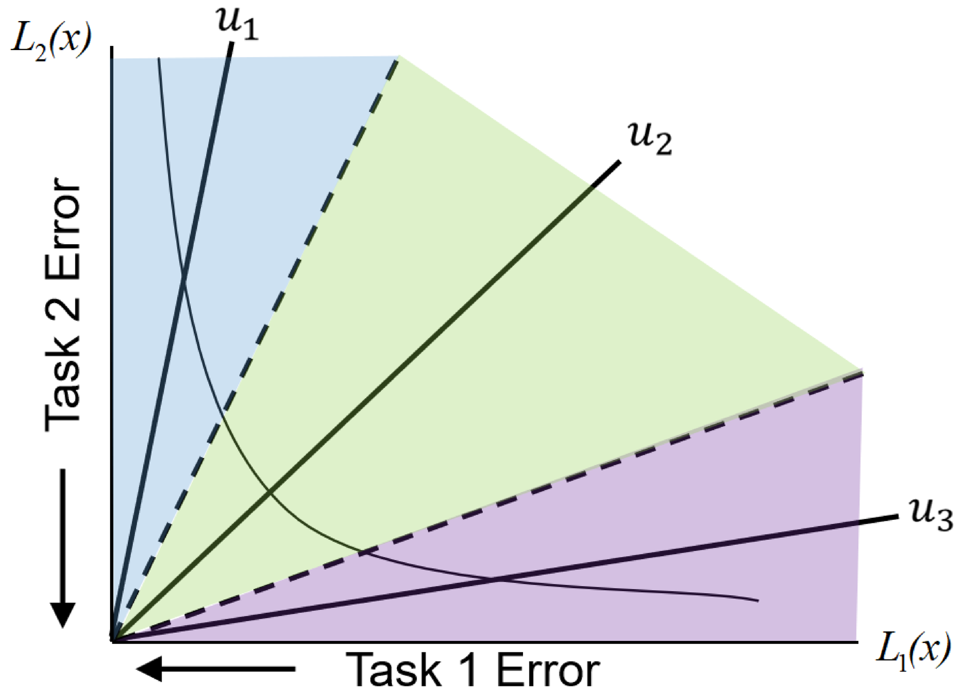
Motivations:

- Current linear scalarization methods need exhaustive weights search which could be inefficient.
- Existing adaptive weight methods can not make different trade-offs among tasks.

Main Contributions:

- A novel method to decompose a MTL problem into multiple subproblems with different trade-offs.
- Show the proposed Pareto MTL can be reformulated as a linear scalarization approach to solve MTL with dynamically adaptive weights.
- A scalable optimization algorithm to solve all constrained subproblems with different preferences.

Key Idea:



Pareto MTL decomposes a given MTL problem into several subproblems with a set of preference vectors. Each MTL subproblem aims at finding one Pareto solution in its restricted preference region.

Problem Formulation

Problem: Consider a MTL problem with m correlated tasks:

$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), \cdots, \mathcal{L}_m(\theta))^{\mathrm{T}}. \quad (1)$$

Traditional linear scalarization:

$$\min_\theta \mathcal{L}(\theta) = \sum_{i=1}^{m} \boldsymbol{w}_i \mathcal{L}_i(\theta). \quad (2)$$

where $\boldsymbol{w}_i$ is hard to be set.

Pareto MTL: Decompose a MTL problem with a set of unit preference vectors $\{\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_K\}$ in R_+^m:

$$\min_\theta \mathcal{L}(\theta) = (\mathcal{L}_1(\theta), \cdots, \mathcal{L}_m(\theta))^{\mathrm{T}}, s.t. \mathcal{L}(\theta) \in \Omega_k, \quad (3)$$

where $\Omega_k (k = 1, ..., K)$ is a subregion:

$$\Omega_k = \{\boldsymbol{v} \in R_+^m | \boldsymbol{u}_j^T \boldsymbol{v} \leq \boldsymbol{u}_k^T \boldsymbol{v}, \forall j = 1, ..., K\}. \quad (4)$$

The constraints can be reformulated as:

$$\mathcal{G}_j(\theta_t) = (\boldsymbol{u}_j - \boldsymbol{u}_k)^T \mathcal{L}(\theta_t) \leq 0, \forall j = 1, ..., K. \quad (5)$$

Solving the Subproblems: Find a valid direction to minimize all loss functions and activated constraints.

$$\begin{aligned}(d_t, \alpha_t) = &\arg \min_{d \in R^n, \alpha \in R} \alpha + \frac{1}{2}||d||^2 \\ s.t. \ &\nabla \mathcal{L}_i(\theta_t)^T d \leq \alpha, i = 1, ..., m. \\ &\nabla \mathcal{G}_j(\theta_t)^T d \leq \alpha, j \in I_\epsilon(\theta_t),\end{aligned} \quad (6)$$

where $I_\epsilon(\theta)$ is the index set of activated constraints.

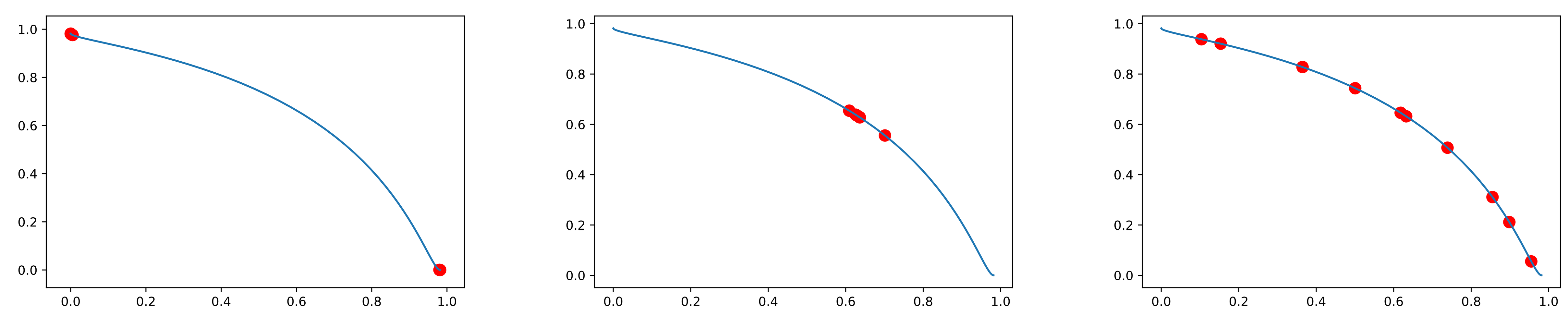
Pareto MTL as Adaptive Linear Scalarization:

$$\begin{aligned}&\mathcal{L}(\theta_t) = \sum_{i=1}^{m} \alpha_i \mathcal{L}_i(\theta_t), \\ &\text{where } \alpha_i = \lambda_i + \sum_{j \in I_{\epsilon(\theta)}} \beta_j (\boldsymbol{u}_{ji} - \boldsymbol{u}_{ki}),\end{aligned} \quad (7)$$

where λ_i and β_j are obtained by solving the dual form of problem (6) with assigned reference vector $\boldsymbol{u}_k$.

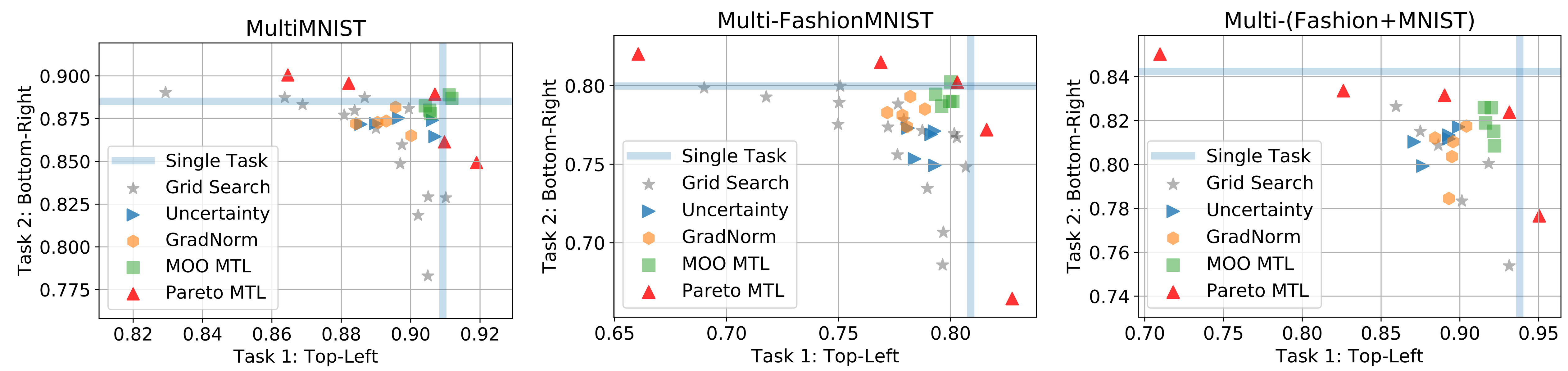
Experiments & Results

Synthetic Example with Concave Pareto Front:



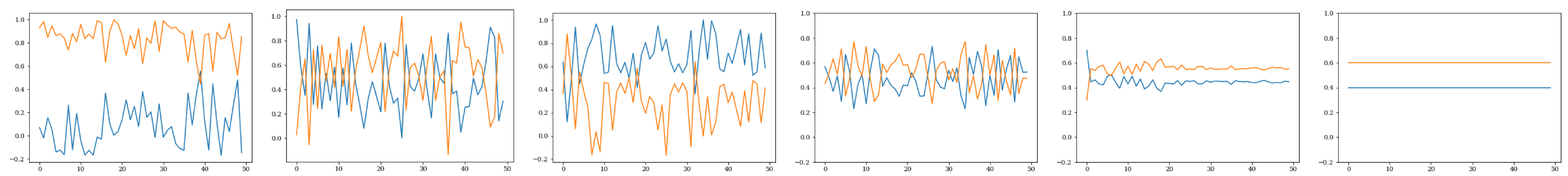
Random Linear Scalarization — MOO MTL — Pareto MTL (Ours)

MultiMNIST, MultiFashionMNIST and Multi-(Fashion+MNIST):



MultiMNIST — MultiFashionMNIST — Multi-(Fashion+MNIST)

Adaptive Weights for Different Algorithms:



ParetoMTL (1,0) — ParetoMTL ($\sqrt{2}/2$, $\sqrt{2}/2$) — ParetoMTL (0,1) — MOO-MTL — GradNorm — Fixed Weight

Pareto MTL with Randomly Generated Preference Vectors: