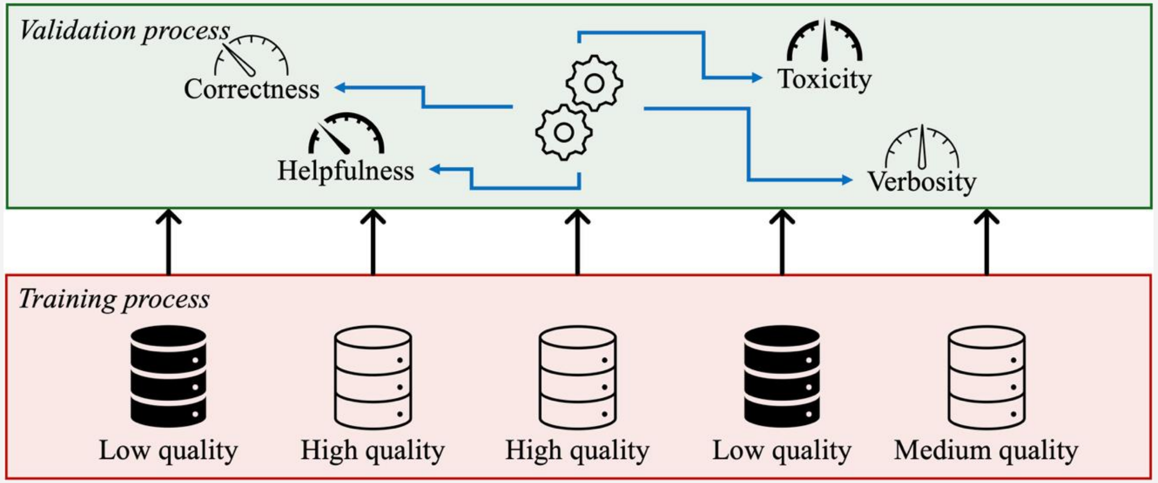
A Tale of Two Problems: Multi-Task Bilevel Learning Meets Equality Constrained Multi-Objective Optimization

Zhiyao Zhang[1], Myeung Suk Oh[1], Zhen Qin[1], Jiaxiang Li[2], Xin Zhang[2], Jia Liu[1]

[1]Department of Electrical and Computer Engineering, The Ohio State University
[2]Meta Platforms, Inc.

Multi-Task Bilevel Learning (MTBL)

In modern machine learning paradigms, many problems involve not only ***multiple*** optimization objectives, but also variables in ***nested*** structure. For instance, we can consider a data curation task in large language model (LLM) fine-tuning as follows:



To this end, we formulate the following multi-task bilevel learning (MTBL) problem:

$$\min_{x,y} F(x,y) = \left[f_1(x,y), \ldots, f_S(x,y)\right]^\top$$
$$\text{s.t. } y \in \mathcal{M}(x) := \arg\min_{y} g(x,y),$$

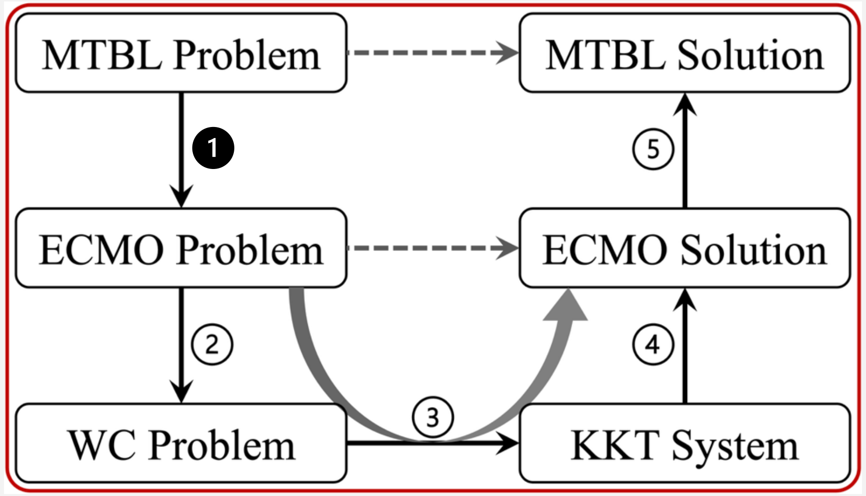
where F is the upper-level (UL) function, and g is the lower-level (LL) function, and S denotes the number of objectives.

Multi-Task Bilevel Learning (MTBL)

In the literature of MTBL, a ***widely adopted*** yet ***overly restrictive*** assumption is lower-level strong convexity (**LLSC**), i.e., $g(x,\ \cdot)$ is strongly convex w.r.t. y for any given x. LLSC ensures a *singleton* LL solution set, and allows one to leverage well studied methods, such as hyper-gradient-based techniques.

In this paper, we relax LLSC to lower-level general convexity (**LLGC**), i.e., $g(x,\ \cdot)$ is convex (but not necessarily strongly convex) w.r.t. y for any given x. In LLGC-MTBL, not only the hyper-gradients become ill-defined, but the optimality of the UL subproblem also needs to be re-interpreted in the Pareto equilibrium sense.

Our approach is to ***equivalently*** convert the LLGC-MTBL problem into an equality constrained multi-objective (**ECMO**) optimization problem, using the fact that solving the LL subproblem for any given x under convexity condition is equivalent to solving its first-order stationarity condition $\nabla_y g(x,y) = 0$, which is both necessary and sufficient.



Equality Constrained Multi-Objective (ECMO) Optimization

Specifically, the ECMO problem has the following formulation:

$$\min_{z\in\mathbb{R}^k} F(z) = \left[f_1(z),\ldots,f_S(z)\right]^\top$$
$$\text{s.t. } h_i(z) = 0, i = 1,\ldots,q,$$

where $k \coloneqq p+q$, $z := [x^\top,\ y^\top]^\top$, and $h_i(z) := \nabla_{y_i} g(x,\ y) = 0$.

For the ECMO problem, there exist at least the following two technical challenges:

- **Lack of Optimality Characterizations and Appropriate Convergence Metrics**: Unlike the unconstrained multi-objective optimization (MOO) problems, the characterization of Pareto stationarity for ECMO remains unclear, and the literature lacks appropriate convergence metrics for solving the ECMO problem;
- **Algorithm Design and Theoretical Analysis**: Even with the Pareto stationarity characterization and convergence metrics established for ECMO, developing algorithms that can handle the equality constraints in ECMO and enable convergence analysis remains challenging.

Pareto Stationarity for ECMO

Pareto Optimality: In MOO, there does *not* exist a unique minimizer that optimizes all conflicting objectives. Instead, we measure optimality in Pareto sense:

- A solution z dominates another solution z' if and only if $f_s(z) \le f_s(z'), \forall s \in [S]$, and there exists at least one $s \in [S]$ such that the inequality holds strictly. A feasible $\tilde{z}$ is **Pareto optimal** if and only if no other feasible $\hat{z}$ dominates $\tilde{z}$**.**

Intuitively, Pareto optimality means that no objective can be improved without sacrificing at least one other objective.

Pareto Stationarity: However, seeking Pareto optimality is NP hard in general. A practical surrogate is Pareto stationarity:

- For the unconstrained MOO problem, $\tilde{z}$ is a **Pareto stationary** point if and only if there does not exist a direction $d \in \mathbb{R}^k$, such that $\nabla f_s(\tilde{z})^\intercal d < 0, \forall s \in [S]$.

Nevertheless, we construct a counterexample in our paper, showing that this becomes invalid when equality constraints are involved. Therefore, we develop the correct definition of Pareto stationarity in ECMO.

Pareto Stationarity for ECMO:

- For an ECMO problem with its *tangent cone* $\mathcal{D}$, a direction d at a solution z is feasible if $z + \epsilon d \in \mathcal{D}$ for small enough $\epsilon > 0$. In ECMO, a feasible point $\tilde{z}$ is **Pareto stationary** if and only if there does not exist a feasible direction $d \in \mathbb{R}^k$, such that $\nabla f_s(\tilde{z})^\intercal d < 0, \forall s \in [S]$.
- We show that: if $\tilde{z}$ is locally weakly Pareto optimal, then $\tilde{z}$ is also Pareto stationary.

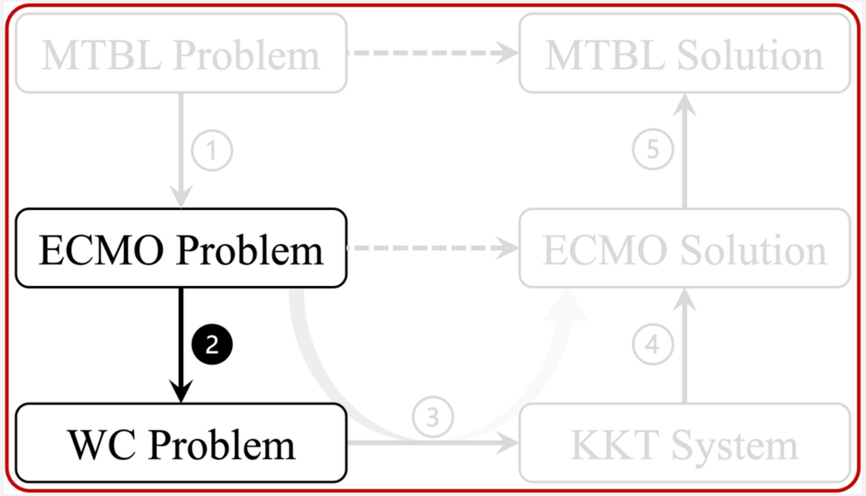
Weighted-Chebyshev (WC) Problem

Weighted-Chebyshev (WC) is a scalarization technique that converts a *vector-valued* optimization problem into a *scalar-valued* problem. It consistently minimizes the "worst-performing" objective at each iteration.

For any preference vector $\lambda \in \Delta_S^+$, where $\Delta_S^+ = \{v \in \mathbb{R}^S : v \geq 0,\ \sum_{s=1}^{S} v_i = 1\}$ denotes the standard simplex, the associated WC problem for ECMO is:

$$\min_{\rho, z} \rho, \quad \text{s.t. } h_i(z) = 0, i \in [q],\ \ \lambda_s f_s(z) \leq \rho, s \in [S].$$

We prove that: there is a **one-to-one** correspondence between *the weakly Pareto optimal front for ECMO* and *the solutions to the corresponding WC problems* (with preference λ varying across Δ_S^+).

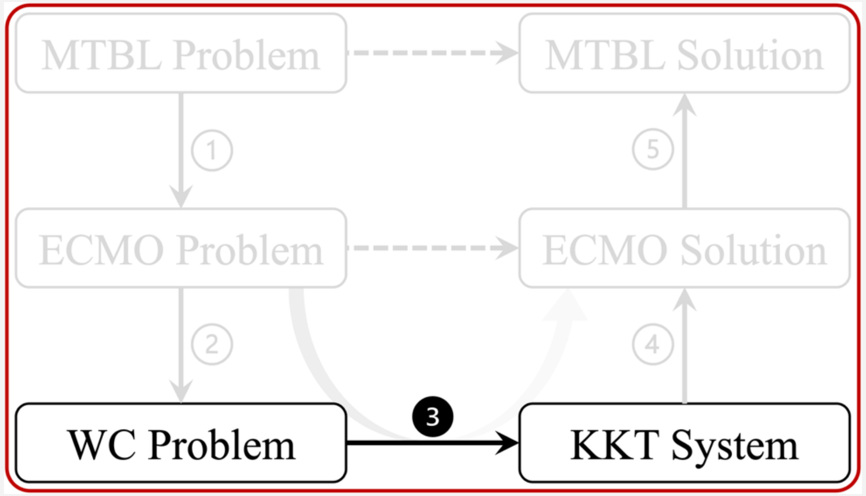


KKT System

Karush–Kuhn–Tucker (KKT) Condition is widely adopted as an optimality condition, and it is known that under some additional condition (such as certain constraint qualification and second-order condition, etc.), the KKT condition is both *necessary* and *sufficient*. Therefore, we use the KKT condition to measure the optimality of WC (and thus ECMO) problem. Considering the special structure of the WC problem, we define the following **KKT system**:

$$\mathcal{K}(\rho, z, \omega, \nu, \lambda) = \begin{pmatrix} \sum_{s=1}^{S} \omega_s - 1 \\ \sum_{s=1}^{S} \omega_s \lambda_s \nabla f_s(z) + \sum_{i=1}^{q} \nu_i \nabla h_i(z) \\ h(z) \\ [\min\{\omega_s, \rho - \lambda_s f_s(z)\}]_{s \in [S]} \end{pmatrix},$$

where the last term considers both *dual feasibility* and *complementary slackness*. This allows us to define a rational metric: for any $\epsilon > 0$, we define a point $\tilde{z}$ to be an **ϵ-Pareto stationary solution** of ECMO if and only if there exist some $\rho \in \mathbb{R}, \omega \in \mathbb{R}^S, \nu \in \mathbb{R}^q, \lambda \in \Delta_S^{++}$, such that $\|\mathcal{K}(\rho, z, \omega, \nu, \lambda)\|_2^2 \leq \epsilon$.



WC-Penalty Algorithm

We propose a penalty-based algorithm, **WC-Penalty algorithm**. To this end, we (i) introduce slackness variables $\delta \in \mathbb{R}_+^S$ to convert the inequality constraints in WC-scalarized problem into equality constraints; and (ii) incorporate all equality constraints as penalty terms in the objective function, which leads to the following formulation:

$$\min_{\theta} P(\theta) = \rho + \frac{u}{2} \sum_{i=1}^{q} h_i(z)^2 + \frac{v}{2} \sum_{s=1}^{S} (\lambda_s f_s(z) + \delta_s - \rho)^2 \quad \text{s.t. } \delta \geq 0,$$

where $\theta := [\rho^\top, z^\top, \delta^\top]^\top$, and $u, v > 0$ are sufficiently large hyper-parameters. By denoting $\mathcal{C} := \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}_+^S$, we can simplify it as $min_{\theta \in \mathcal{C}} P(\theta)$. We then develop our **WC-Penalty algorithm**:

Algorithm 1 The WC-Penalty algorithm.

1: **Input:** Iteration rounds T, initialization $\theta_0 = (\rho_0, z_0, \delta_0) \in \mathcal{C}$, with $\rho_0 \geq 0$, and step-size η.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Compute $\nabla P(\theta_t)$.
4: Update $\theta_{t+1} = \mathcal{P}_{\mathcal{C}}(\theta_t - \eta \nabla P(\theta_t))$.
5: **end for**

WC-Penalty Algorithm

Theoretical Results: Under mild assumptions, for any preference $\lambda \in \Delta_S^{++}$, selecting $\eta = \Theta(T^{-\frac{1}{4}})$ and u=v= $\Theta(T^{\frac{1}{4}})$, WC-Penalty algorithm achieves the following convergence result:

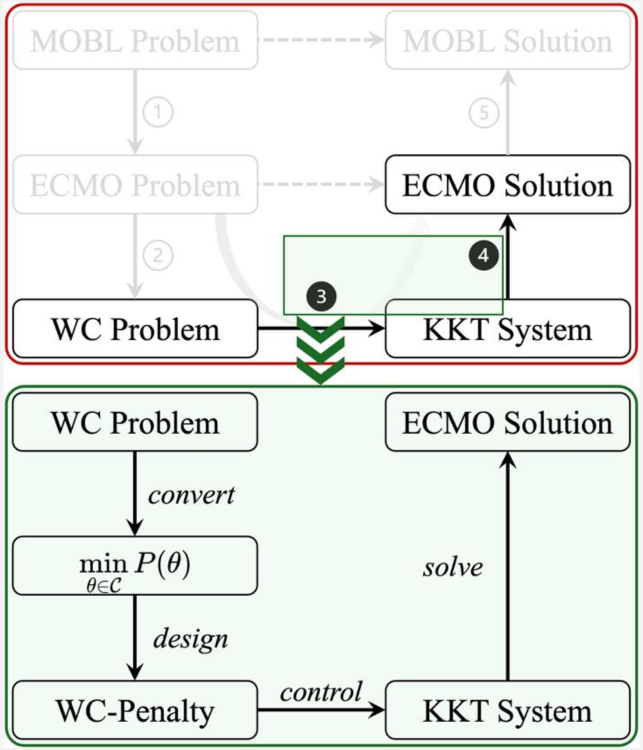
$$\frac{1}{T}\sum_{t=0}^{T-1} \|\mathcal{K}(\rho_t, z_t, \omega_t, \nu_t, \lambda)\|^2 = \mathcal{O}(S/T^{\frac{1}{2}})$$

The theorem is established in two key steps.

- **First**, we consider the dynamics of $P(\theta_t)$ generated by WC-Penalty algorithm, and hence solving the WC-scalarized problem.
- **Second**, we judiciously select the parameters, which, according to the theoretical foundations established before (Pareto stationarity and one-to-one correspondence), allow us to control each term in the KKT system.

Finally, we can easily solve the **MTBL** problem as a special case of the ECMO problem by splitting the variable *z* explicitly into *x, y*:

$$\min_{\rho,x,y,\delta} P(\rho,x,y,\delta) = \rho + \frac{u}{2}\sum_{i=1}^{q}(\nabla_y g(x,y))_i^2 + \frac{v}{2}\sum_{s=1}^{S}(\lambda_s f_s(x,y) + \delta_s - \rho)^2 \quad \text{s.t. } \delta_s \geq 0, s \in [S].$$

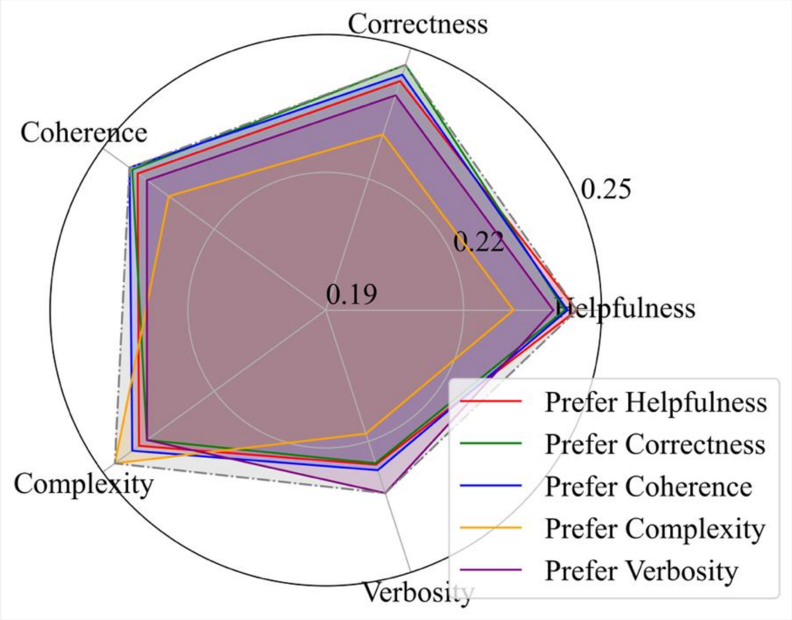


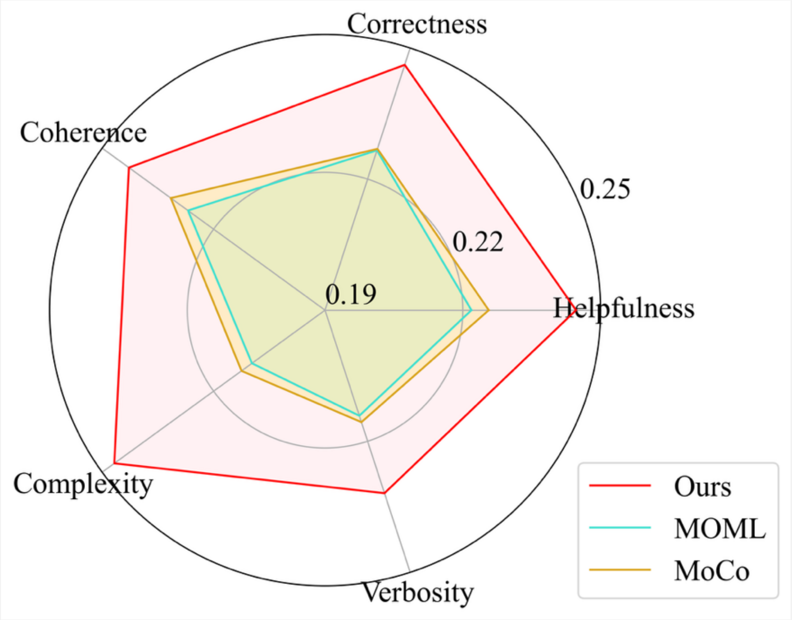
Experiments: LLM-based Data Curation

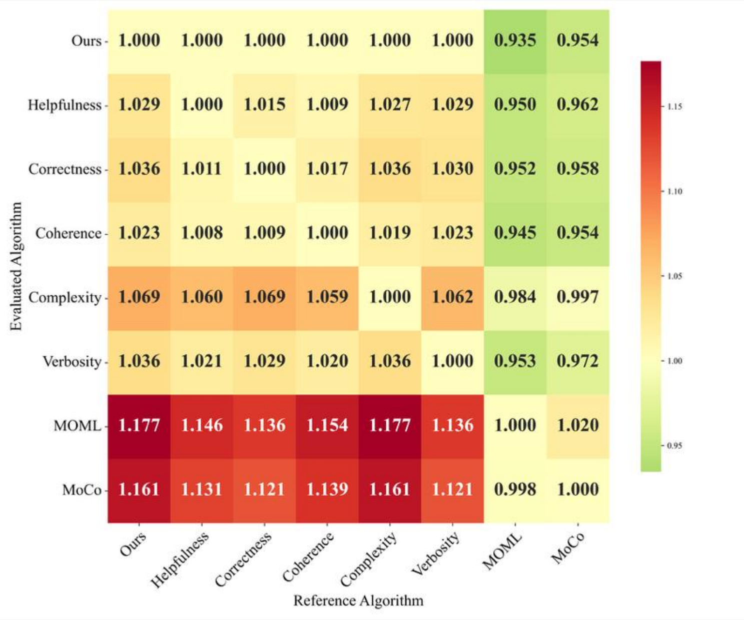
We consider the data curation task on multi-objective LLM alignment, where the goal is to determine the proportion weights of dataset to minimize multiple human-aligned losses (such as helpfulness, correctness, verbosity, etc.) in validation. Here, we denote y as the LLM parameters, x as the weighting vector:

$$\min_{x,y}\left(\sum_{j=1}^{|\mathcal{V}_1|}\mathcal{L}(\tilde{r}_{1,j}(p_{1,j};y(x)),r_{1,j}),\ldots,\sum_{j=1}^{|\mathcal{V}_M|}\mathcal{L}(\tilde{r}_{M,j}(p_{M,j};y(x)),r_{M,j})\right)$$

$$\text{s.t. } y(x)\in\arg\min_{y}\sum_{n=1}^{N}\frac{\exp(x_n)}{\sum_{n'}\exp(x_{n'})}\sum_{i=1}^{|\mathcal{T}_n|}\mathcal{L}(\tilde{r}_{n,i}(p_{n,i};y),r_{n,i}).$$







A Tale of Two Problems: Multi-Task Bilevel Learning Meets Equality Constrained Multi-Objective Optimization

Zhiyao Zhang[1], Myeung Suk Oh[1], Zhen Qin[1], Jiaxiang Li[2], Xin Zhang[2], Jia Liu[1]

[1]Department of Electrical and Computer Engineering, The Ohio State University
[2]Meta Platforms, Inc.

ArXiv Link:

Correspondence to: *zhang.15178@osu.edu*