FERERO: A Flexible Framework for Preference-Guided Multi-Objective Learning

Lisha Chen
Rensselaer Polytechnic Institute
Troy, NY, United States
chenl21@rpi.edu

AFM Saif
Rensselaer Polytechnic Institute
Troy, NY, United States
saifa@rpi.edu

Yanning Shen
University of California, Irvine
Irvine, CA, United States
yannings@uci.edu

Tianyi Chen
Rensselaer Polytechnic Institute
Troy, NY, United States
chentianyi19@gmail.com

Abstract

Finding specific preference-guided Pareto solutions that represent different trade-offs among multiple objectives is critical yet challenging in multi-objective problems. Existing methods are restrictive in preference definitions and/or their theoretical guarantees. In this work, we introduce a Flexible framEwork for pREfeRence-guided multi-Objective learning (**FERERO**) by casting it as a constrained vector optimization problem. Specifically, two types of preferences are incorporated into this formulation – the *relative preference* defined by the partial ordering induced by a polyhedral cone, and the *absolute preference* defined by constraints that are linear functions of the objectives. To solve this problem, convergent algorithms are developed with both single-loop and stochastic variants. Notably, this is the *first single-loop primal algorithm* for constrained vector optimization to our knowledge. The proposed algorithms adaptively adjust to both constraint and objective values, eliminating the need to solve different subproblems at different stages of constraint satisfaction. Experiments on multiple benchmarks demonstrate the proposed method is very competitive in finding preference-guided optimal solutions. Code is available at https://github.com/lisha-chen/FERERO/.

1 Introduction

Many machine learning tasks inherently involve multiple objectives, which can be different performance metrics such as accuracy, fairness, and privacy; or, the same metrics defined on different data [40, 31]. To tackle such multi-objective problems, it is common to learn a shared model that simultaneously performs well on all the objectives. Compared to learning one model for each objective, learning a shared model has the benefit of reducing both the model size and the inference time. This can be achieved through multi-objective optimization [40, 45, 25, 6], which is to learn a model that minimizes the vector-valued objective. In practical applications, it is of interest to learn solutions with controlled trade-offs or preferences. To further illustrate, we give two examples below.

In fairness-aware machine learning, a trade-off exists between the fairness $f_{\text{fair}}(\theta)$ and accuracy $f_{\text{acc}}(\theta)$, see also Figure 1a. With θ denoting the model parameter, and C denoting the partial order

The work of L. Chen, AFM Saif, and T. Chen was supported by the National Science Foundation (NSF) projects 2401297, 2412486, the RPI-IBM Artificial Intelligence Research Collaboration (AIRC), the Cisco Research Award, and the IEEE Signal Processing Society scholarship. The work of Y. Shen was supported by NSF ECCS-2412484.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Comparison to existing methods. "Flexibility" represents preference modeling, such as by using weights, preference vectors (rays), or constraints. "Exactness" represents the ability to align with a preference vector exactly. "Deter.", "Stoch." represent deterministic and stochastic, respectively. "✗" means not provided in the corresponding work, and "-" means not relevant.

Method	Preference Flexibility	Exactness	Controlled ascent	Single loop	Convergence Deter.	Convergence Stoch.
Linear Scalarization	weight	-	✗	✓	T^{-1}	$T^{-\frac{1}{2}}$
(Smooth) Tchebycheff [23]	weight	-	✗	✓	non-asymptotic	✗
PMTL [24]	inequalities (absolute)	✗	✗	✗	asymptotic	✗
EPO [30]	r^{-1} ray (ratio, absolute)	✓	✓	✗	asymptotic	✗
(X)WC-MGDA [33]	shifted ray (absolute)	✓	✗	✗	✗	✗
FERERO (ours)	relative & absolute	✓	✓	✓	T^{-1}	$T^{-\frac{1}{2}}$

cone, to find the optimal models that consider different trade-offs, one can solve the following problem with different thresholds ϵ [8]

$$\text{maximize}_C \ \ (f_{\text{acc}}(\theta), f_{\text{fair}}(\theta))^\top \ \text{ s.t. } \ f_{\text{fair}}(\theta) \geq \epsilon. \tag{1.1}$$

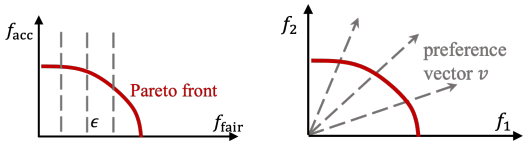
Another example is in drug or molecule design, where the goal is to design drugs or molecules with multiple desired properties $f_1(\theta), f_2(\theta), \ldots, f_M(\theta)$. Aiming to align the values of the properties $F(\theta)$ with a predefined preference vector v as in Figure 1b, one can solve the following problem [29, 1, 46]

$$\text{maximize}_C \ \ F(\theta) \coloneqq \big(f_1(\theta), \ldots, f_M(\theta)\big)^\top \ \text{ s.t. } \ BF(\theta) = Bv, \ Bv = 0 \tag{1.2}$$

where $B \in \mathbb{R}^{(M-1)\times M}$ is full row rank.

Then a natural question arises:

Can we develop a principled framework to capture flexible preferences and admit provably convergent deterministic and stochastic algorithms?



(a) (b)

Figure 1: Illustration of preferences in different examples. The solid red curves represent the Pareto front, dashed lines represent preference constraints.

Our answer to this question is affirmative. Recognizing that all the aforementioned applications can be addressed within a unified framework, we formulate preference-guided multi-objective learning (PMOL) as a constrained vector optimization problem. Specifically, given a model $\theta \in \mathbb{R}^q$, and the objectives $f_m : \mathbb{R}^q \to \mathbb{R}$, $m = 1, \ldots, M$, we define the constrained vector optimization problem as

$$\min_{\theta\in\mathbb{R}^q} F(\theta) \coloneqq \big(f_1(\theta), \ldots, f_M(\theta)\big)^\top, \quad \text{s.t.} \quad G(\theta) \leq 0, \ H(\theta) = 0 \tag{PMOL}$$

where $G(\theta)$ and $H(\theta)$ are the vector-valued preference constraints such as the examples in (1.1) and (1.2). Here "$\leq$" and "$=$" are element-wise relations on the vectors, with each row representing one constraint. In these examples, the preferences are directly defined in the objective space, as intersections of half-spaces defined by the hyperplanes; see Figure 1. Thus, $G(\cdot)$ and $H(\cdot)$ in (PMOL) can be expressed as linear functions of $F(\theta)$, given by

$$G(\theta) = B_g F(\theta) + b_g, \ \ H(\theta) = B_h F(\theta) + b_h \tag{1.3}$$

where $B_g \in \mathbb{R}^{M_g \times M}, B_h \in \mathbb{R}^{M_h \times M}$, and $b_g \in \mathbb{R}^{M_g}, b_h \in \mathbb{R}^{M_h}$. Different B_g, B_h, b_g, b_h correspond to different preferences, and thus different trade-offs among the objectives.

A comparison of our methods to existing methods is summarized in Table 1. Specifically, our contributions are listed as follows:

- **C1)** We cast the PMOL problem as a constrained vector optimization problem, and develop the FERERO framework to capture flexible preferences.
- **C2)** We develop a method with an adaptive subprogram that efficiently finds update directions to meet flexible preferences, eliminating the need for multiple subprograms under different active constraints. This approach ensures iterative improvement in a general partial ordering while allowing controlled ascent of objectives to satisfy preferences.

C3) We propose a practical single-loop algorithm and establish non-asymptotic convergence guarantees for deterministic and stochastic variants of the proposed algorithm. To our best knowledge, this is the *first single-loop primal algorithm* in constrained vector optimization with convergence guarantees.

C4) We apply the proposed algorithms to various synthetic and real-world image and speech datasets to demonstrate its ability to find flexible preference-guided optimal models.

In our theoretical analysis, we address the following technical challenges.

T1) The commonly used constraint qualification assumptions do not generally hold for the PMOL problem. We overcome this challenge by leveraging the specific structure that the constraints are linear functions of F to prove the calmness condition holds for PMOL. See more details in Lemma 2.

T2) The commonly used merit or Lyapunov functions for constrained optimization are usually non-smooth, making it difficult to derive a descent lemma on the functions, and thus difficult to derive the non-asymptotic convergence guarantees. We overcome this challenge by exploiting the optimality conditions and proper step size choices. See Lemma 8.

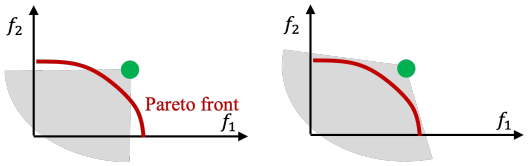
T3) The convergence of our single-loop algorithm is slower with the commonly-used merit functions. We provide a sharper analysis by introducing a different merit function and exploiting the algorithm properties; see Theorem 3.

2 Problem Setup and A Meta Algorithm

To characterize the optimality conditions of PMOL, we introduce the generalized notion of dominance and the related concept of optimality. We then present a meta-algorithm to solve PMOL.

2.1 Problem setup and preliminaries

We first introduce optimality definitions for PMOL that go beyond the standard definitions of Pareto optimality [12, 10, 26]. Given two vectors v and w, we use $v < w$ and $v \leq w$ to denote $v_i < w_i$ for all i, and $v_i \leq w_i$ for all i, respectively. We use $v \lneq w$ to denote $v \leq w$ and $v \neq w$, and define $>, \geq, \gneq$ analogously.



(a) (b)

Figure 2: Illustration of C_A-dominance. The solid red curves are the Pareto fronts, green dots are the reference points, gray shaded regions are the set of objectives dominating the reference points, under different C_A in (a) and (b).

Definition 1 (C_A-dominance [11, 19])**.** *Given $v, w \in \mathbb{R}^M$, $A \in \mathbb{R}^{M\times M}$, and $C_A := \{y \in \mathbb{R}^M \mid Ay \geq 0\} \neq \emptyset$, we say v strictly dominates w based on C_A if and only if $A(v - w) < 0$.*

The generalized dominance defines a partial order on $\mathbb{R}^M$, i.e., the relation between two vectors. Illustrations of different partial orders are given in Figure 2. Figure 2a shows the dominance relation under the widely used non-negative orthant cone with $C_A = \mathbb{R}^M_+$, corresponding to Pareto optimality. However, as illustrated by the figure, given the initial green reference point, a descent method such as MGDA [12] cannot find points on the Pareto front but outside of the gray shaded region. This poses a critical challenge for applications where specific preference-guided solutions on the Pareto front are needed. Nevertheless, this issue can be addressed by substituting $\mathbb{R}^M_+$ with a more general definition of C_A as displayed in Figure 2b. Under this partial order, a general descent method is able to reach any points on the Pareto front starting from the green reference point.

Based on the partial order, one can then find the minimum or optimal elements in the vector-valued objective space, whose formal definition is provided below.

Definition 2 (C_A-optimal)**.** *A point $\theta \in \mathbb{R}^q$ is C_A-optimal if there is no $\theta' \neq \theta$ such that, $AF(\theta') \lneq AF(\theta)$. A point θ is weakly C_A-optimal if there is no $\theta' \neq \theta$ such that, $AF(\theta') < AF(\theta)$.*

Note that, C_A is a polyhedral cone, or the intersection of half-spaces defined by the rows of the inequality $Ay \geq 0$. When $A = I_M$, an $M \times M$ identity matrix, $C_A = \mathbb{R}^M_+ := \{y \in \mathbb{R}^M \mid y_m \geq 0 \ \forall m \in [M]\}$, then Definition 1 reduces to the commonly used notion of dominance associated with Pareto optimality. The cone C_A can be interpreted as a *relative preference* that defines the objectives'

improvement directions, which generalizes the relative preference defined by $\mathbb{R}_+^M$. In contrast, the preference defined by constraints in (1.3) can be interpreted as an *absolute preference* that defines the feasible or preferred set of objective function values. In practice, C_A can be chosen based on the requirements of specific applications. For example, when the controlled ascent of objectives is needed [30], we can choose C_A such that the controlled ascent direction belongs to $-C_A$. We defer the detailed implementation to Section 3.2. The C_A-optimal set, denoted as $\mathcal{P}_A$, contains all the C_A-optimal models. When $A = I_M$, $\mathcal{P}_A$ is the Pareto optimal set $\mathcal{P}$. The Pareto front is the set of function values evaluated at Pareto optimal models, i.e., $\mathcal{F} = \{F(\theta) \mid \theta \in \mathcal{P}\}$.

We make the following standard assumptions throughout the paper [12, 17, 6].

Assumption 1. *1. (Non-negative objectives) $AF(\theta) \geq 0$, and $\mathbf{1}^\top AF(\theta) \geq c_{AF} > 0$ for all $\theta \in \mathbb{R}^q$.*
2. (Differentiable objectives) F is twice continuously differentiable.
3. (Ordering cone with non-empty interior) C_A has a non-empty interior.

2.2 Find the preference-guided direction

In this section, we proceed to discuss an adaptive method to solve (PMOL). At iteration t, the algorithm finds an update direction d_t and performs the iterative update $\theta_{t+1} = \theta_t + \alpha_t d_t$ with a step size α_t. Ideally, the update direction d_t is chosen to improve the objective $F(\theta)$ and to satisfy the preference constraints. It is desirable that when the constraints are not satisfied, d_t decreases the violation of constraints and improves the objectives in the general partial ordering sense; when the constraints are satisfied, d_t improves the objectives and ensures the constraints are satisfied. To achieve this, we find a direction $d^*(\theta)$ that solves following subprogram

$$\psi(\theta) \coloneqq \min_{(d,c)\in\mathbb{R}^q\times\mathbb{R}} c + \frac{1}{2}\|d\|^2 \quad \text{s.t. } A\nabla F(\theta)^\top d \leq \frac{c}{\mathbf{1}^\top AF(\theta)} AF(\theta) \tag{2.1}$$
$$\nabla G(\theta)^\top d + c_g G(\theta) \leq 0,\ \nabla H(\theta)^\top d + c_h H(\theta) = 0$$

where $\|\cdot\|$ denotes the ℓ_2-norm, c_g and c_h are pre-defined positive constants. Larger c_g and c_h put more emphasis on constraint satisfaction than objective improvement. We call this subprogram *adaptive* since it deals with constraints in an adaptive way, which does not require the initial model to be feasible, nor θ_t to be feasible at each iteration. But rather, it finds an update direction that decreases the constraint violation. Because of this, it neither requires solving different subprograms at different stages nor requires different treatment of the active set of inequalities as in existing works [24, 30, 33].

We then show in Lemma 1 that the desired properties can be satisfied.

Lemma 1. *For the subprogram (2.1), the following holds:*
If θ is a local optimal solution with $AF(\theta) > 0$, then $d^*(\theta) = 0$, $\psi(\theta) = 0$. Otherwise, if θ is not a local optimal solution, then $d^*(\theta) \neq 0$, $\psi(\theta) < 0$, and when θ is feasible,

$$2\psi(\theta) \leq -\|d^*(\theta)\|^2 < 0. \tag{2.2}$$

Let θ be a weak C_A-optimal solution, with $(AF(\theta))_m = 0$ for some $m \in [M]$. If there exists feasible and non-strictly improving directions at θ with $A\nabla F(\theta)^\top d \lneq 0$, then $d^*(\theta) \neq 0$, $\psi(\theta) < 0$. Otherwise, $d^*(\theta) = 0$, $\psi(\theta) = 0$.

Algorithm 1 A meta FERERO algorithm

```
Initialize t = 0, θ_0, step size {α_t}; define A.
while ψ(θ_t) ≠ 0 do
    Compute gradient ∇F(θ_t);
    Compute λ_t by (approximately) solving (2.3);
    Compute the update direction
    d_t = −∇F(θ_t) A_ag^⊤ λ_t;
    Update θ_t by θ_{t+1} = θ_t + α_t d_t;
    Set t = t + 1;
end while
```

By Lemma 1, $\|d^*(\theta)\| = 0$ is a stationary condition for PMOL. Recall the feasibility condition requires $[G(\theta)]_+ = 0$ and $|H(\theta)|_{\text{ab}} = 0$, where $[\cdot]_+$ and $|\cdot|_{\text{ab}}$ are entry-wise ReLU and absolute functions, respectively. And the complementary slackness condition requires ${\lambda_g^*}^\top [-G(\theta)]_+ = 0$. Thus $\|d^*(\theta)\|^2 + {\lambda_g^*}^\top [-G(\theta)]_+ + \|[G(\theta)]_+\|_1 + \|H(\theta)\|_1$ achieves zero if and only if the model θ satisfies the first-order KKT condition. Besides the properties in Lemma 1, it has an additional scale-invariant property that is deferred to Lemma 4 due to space limit.

By the Lagrangian of (2.1), the optimal update direction can be expressed in a simple form as a weighted combination of the gradients, i.e. $d^*(\theta) = -\nabla F(\theta) A_{ag}^\top \lambda^*$, with $A_{ag} \coloneqq [A; B_g; B_h]$, and

$$\lambda^* \in \arg\min_{\lambda\in\Omega_\lambda(\theta)} \varphi(\lambda;\theta) \coloneqq \frac{1}{2}\|\nabla F(\theta) A_{ag}^\top \lambda\|^2 - c_g \lambda_g^\top G(\theta) - c_h \lambda_h^\top H(\theta) \tag{2.3}$$

where $\lambda = [\lambda_f; \lambda_g; \lambda_h]$, $\Omega_\lambda(\theta)$ is the domain of the Lagrangian multipliers, given by [1]

$$\Omega_\lambda(\theta) := \Omega_{\lambda_f}(\theta) \times \mathbb{R}_+^{M_g} \times \mathbb{R}^{M_h}, \text{ with } \Omega_{\lambda_f}(\theta) := \{\lambda \in \mathbb{R}_+^M \mid \lambda^\top AF(\theta) = \mathbf{1}_M^\top AF(\theta)\}. \quad (2.4)$$

Our goal is to design an algorithm that converges to a KKT solution based on (2.1). However, the KKT condition is not necessary unless certain constraint qualifications (CQs) hold. Prior works [14, 24] assume certain CQs hold, e.g., the Linear Independence Constraint Qualification (LICQ). However, the LICQ assumption (c.f., [14, Section 3.1, (A2)]) does not generally hold at a local optimal solution for problem (PMOL), c.f., Example 1 in Appendix D.3.2. Though some commonly used CQs do not hold generally, in our case, leveraging the specific structure that the constraints are linear functions of F, we can justify the calmness CQ in Definition 9 tailored for our problem in Lemma 2, thus the KKT condition is a necessary optimality condition. The proof is deferred to Appendix D.3.2.

Lemma 2. *Let $\bar{\theta} \in \mathbb{R}^q$ be a global solution to (PMOL). Define $\Sigma(p, q) := \{y \in \mathbb{R}^M \mid B_g y + b_g \leq p, B_h y + b_h = q\}$. If $\Sigma(p, q)$ is a line, the PMOL calmness condition in Definition 9 is satisfied for (PMOL) at $\bar{\theta}$ if $A \in \mathbb{R}^{M \times M}$ is full rank, $H(\theta), G(\theta)$ defined by (1.3) satisfy $[B_h^\top, B_g^\top] \neq 0$, and B_h, B_g are full row rank. Consequently, the KKT condition is a necessary optimality condition.*

Lemma 2 provides a sufficient condition for the KKT condition to be a necessary optimality condition without relying on unjustified assumptions. The requirement that the constraint set is a line in the objective space is common for applications such as alignment to a preference vector.

We then discuss a generic preference-guided multi-objective algorithm based on the subprogram.

2.3 A meta algorithm for preference-guided multi-objective learning

Given the model θ_t at iteration t, one can then solve (2.3) to obtain λ_t. The direction $d_t = -\nabla F(\theta) A_{ag}^\top \lambda_t$ is used to update the model θ_t by $\theta_{t+1} = \theta_t + \alpha_t d_t$ iteratively until convergence. The full procedure of this meta algorithm is summarized in Algorithm 1, where Step 4 is a generic step and can be customized in Section 3.

To establish the non-asymptotic convergence rate, we use the following standard smoothness assumption that has been commonly used in prior works for multi-objective learning [6, 26].

Assumption 2 (Smooth objectives)**.** *For all $m \in [M]$, $\nabla f_m(\theta)$ is $\ell_{f,1}$-Lipschitz continuous.*

We then state the convergence result for Algorithm 1 in Theorem 1.

Theorem 1 (Convergence of the generic FERERO algorithm)**.** *Suppose Assumptions 1, 2 hold. Let $\{\theta_t\}$ be the sequences produced by Algorithm 1, with d_t being an ϵ-optimal solution to the subprogram (2.1). If $\|\lambda^*(\theta_t)\|_1 \leq c$, $\alpha_t \leq \min\{\frac{1}{c\ell_{f,1}\|A_{ag}^\top\|_{\infty,1}}, c_g^{-1}, c_h^{-1}\}$, and $\alpha_t = \Theta(1)$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \underbrace{\|\nabla F(\theta_t) A_{ag}^\top \lambda^*(\theta_t)\|^2}_{\text{stationarity}} + \underbrace{\lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+}_{\text{complementary slackness}} + \underbrace{\|[G(\theta_t)]_+\|_1 + \|H(\theta_t)\|_1}_{\text{feasibility}} = \mathcal{O}\big(T^{-1} + \epsilon\big). \quad (2.5)$$

Theorem 1 guarantees the non-asymptotic convergence for the generic FERERO algorithm. In Algorithm 1, λ_t can be solved through projected gradient descent or Frank Wolfe algorithm iteratively within an inner loop. In practice, we usually do not need to solve the subprogram exactly. Next, we discuss the efficient single-loop approximate algorithm based on Algorithm 1.

3 Efficient Algorithm Development

In this section, we first discuss efficient algorithm development with the approximate update rules and practical choice of preferences. We focus on (PMOL) with *equality constraints only*, i.e., $M_g = 0$. Building upon this, we then discuss the stochastic variants of the algorithms that can be applied to large-scale learning problems.

Algorithm 2 FERERO-SA

1: Initialize $t = 0$, θ_0, λ_0, step sizes $\{\alpha_t, \gamma_t\}$; define A, number of iterations T.
2: **for** $t = 0, \ldots, T-1$ **do**
3: Compute gradient $\nabla F(\theta_t)$;
4: Compute direction $d_t = -\nabla F(\theta_t) A_{ag}^\top \lambda_t$;
5: Update θ_t by $\theta_{t+1} = \theta_t + \alpha_t d_t$;
6: Update λ_t by (3.1);
7: **end for**

[1] Note that, $A\nabla F(\theta)^\top d \leq \frac{c}{\mathbf{1}^\top AF(\theta)} AF(\theta)$ in (2.1) can be replaced by $A\nabla F(\theta)^\top d \leq c \cdot \mathbf{1}$, which leads to a simplified subprogram with $\Omega_{\lambda_f}(\theta) = \Omega_{\lambda_f} = \Delta^M$, and can be covered by our analysis.

3.1 Single-loop approximate algorithm

In practice, if one only requires the converging solutions generated by the algorithm to be feasible, but not all the iterates, then further approximations can be made to the subprogram (2.3). At iteration t, to obtain an approximate direction d_t, we adopt the following update

$$\lambda_{t+1} = \Pi_{\Omega_\lambda(\theta_t)}\big(\lambda_t - \gamma_t \nabla_\lambda \varphi(\lambda_t; \theta_t)\big). \tag{3.1}$$

The single-loop algorithm with the approximate solution is summarized in Algorithm 2. We name it FERERO with Single-loop Approximate update (FERERO-SA) algorithm.

We make the following additional assumption of Lipschitz objectives to prove the convergence of Algorithm 2, which is standard in optimization literature.

Assumption 3 (Lipschitz objectives)**.** *For all $m \in [M]$, $f_m(\theta)$ is ℓ_f-Lipschitz continuous.*

To prove the convergence of Algorithm 2, we can use the same merit function, which leads to a convergence rate of $\mathcal{O}\big(T^{-\frac{1}{6}}\big)$. See Theorem 2 below and its proof in Appendix F.2.

Theorem 2 (Convergence of the FERERO-SA algorithm)**.** *Suppose Assumptions 1, 2, 3 hold, and $M_g = 0$. Let $\{\theta_t\}, \{\lambda_t\}$ be the sequences produced by the simplified Algorithm 2 with $\Omega_{\lambda_f}(\theta) = \Delta^M$. Assume $\lambda^*(\theta_t), \lambda_\rho^*(\theta_t), \lambda_t$ are bounded. With properly chosen step sizes $\alpha = \Theta(T^{-\frac{5}{6}})$, $\gamma = \Theta(T^{-\frac{1}{6}})$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2 + \|H(\theta_t)\|_1 = \mathcal{O}\Big(T^{-\frac{1}{6}}\Big). \tag{3.2}$$

To obtain a sharper convergence rate, we consider a different merit function $\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2 + \|H(\theta_t)\|^2$, which achieves zero if θ_t satisfies the KKT condition. We use this merit function because it provides better properties for sharper analysis. The detailed proof is deferred to Appendix F.3.

Theorem 3 (Sharper convergence of the FERERO-SA algorithm)**.** *Suppose Assumptions 1, 2, 3 hold, and $M_g = 0$. Let $\{\theta_t\}, \{\lambda_t\}$ be the sequences produced by Algorithm 2. With properly chosen step sizes $\alpha_t = \Theta(1)$, $\gamma_t = \Theta(T^{-1})$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2 + \|H(\theta_t)\|^2 = \mathcal{O}\Big(T^{-1}\Big). \tag{3.3}$$

Theorem 3 states that $\{\theta_t\}$ produced by Algorithm 2 converges to a KKT solution of the PMOL problem in the general nonconvex case. Moreover, both $\|d_t\|^2$ and $\|H(\theta_t)\|^2$ converge to zero at a rate of $\mathcal{O}(T^{-1})$, implying the convergence of both the objective values and the preference constraints. Note that, the convergence in terms of $\|H(\theta_t)\|^2$ at a rate of $\mathcal{O}(T^{-1})$ is weaker compared to the one with $\|H(\theta_t)\|_1$ at the same rate for Algorithm 1. This is reasonable since Algorithm 2 only uses a one-step approximate update of λ_t instead of exactly solving the subprogram.

The stochastic variant. We employ a stochastic variant of Algorithm 2 based on the double sampling techniques developed in the recent work [6]. The update is given by

$$\theta_{t+1} = \theta_t + \nabla F_{\xi_{t,1}}(\theta_t) A_{ag}^\top \lambda_t \tag{3.4a}$$

$$\lambda_{t+1} = \Pi_{\Omega_\lambda(\theta_t)}\big(\lambda_t - \gamma_t \tilde{\nabla}_\lambda \varphi(\lambda_t; \theta_t)\big) \tag{3.4b}$$

$$\tilde{\nabla}_\lambda \varphi(\lambda_t; \theta_t) = A_{ag} \nabla F_{\xi_{t,1}}(\theta_t)^\top \nabla F_{\xi_{t,2}}(\theta_t) A_{ag}^\top \lambda_t - [0^\top, c_h H_{\xi_{t,1}}(\theta_t)^\top]^\top \tag{3.4c}$$

where $\tilde{\nabla}$ is the unbiased stochastic estimate of the gradient, and $\xi_{t,1}$ and $\xi_{t,2}$ are two independent stochastic samples obtained at iteration t.

The full description of the stochastic algorithms and their convergence guarantees are deferred to Appendix G. Note that, compared to [6], we adopt a more efficient implementation that reduces the per-iteration computational complexity.

Theorem 4 (Convergence of the stochastic FERERO algorithm)**.** *Suppose Assumptions 1, 2, 3 hold, and $M_g = 0$. Let $\{\theta_t\}, \{\lambda_t\}$ be the sequences produced by Algorithm 3. Suppose the variance of $\nabla F_\xi(\theta_t), \tilde{\nabla}_\lambda \varphi(\lambda_t; \theta_t)$ are bounded. With properly chosen step sizes $\alpha_t = \alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma_t = \gamma = \Theta(T^{-\frac{3}{2}})$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2 + \|H(\theta_t)\|^2\Big] = \mathcal{O}\Big(T^{-\frac{1}{2}}\Big). \tag{3.5}$$

Theorem 4 generalizes Theorem 3 to its stochastic variants, with a matching convergence rate to the unconstrained stochastic MOO algorithms and stochastic gradient descent. This allows us to apply the algorithm to large-scale machine learning problems, which we detail in Section 5.

3.2 Choice of relative preferences

As briefly discussed in Section 2.1, the ordering cone and the corresponding matrix A can be specified according to practical needs. We first discuss how to obtain matrix A for the relative preference given the set of improvement directions. Then we discuss how to choose the relative preference to allow controlled ascent update, which is useful for touring the Pareto front [30].

Ordering cone generation. In practice, to obtain the polyhedral cone that defines the partial order, one can usually first define the extreme rays of the polyhedral cone. We then show how to convert the extreme ray description of the cone to the half-space description given by matrix A, i.e., $C_A = \{y \in \mathbb{R}^M \mid Ay \geq 0\}$, by showing how to compute A from the extreme rays.

Let $Y = [y_1 \cdots y_M] \in \mathbb{R}^{M \times M}$ be a matrix that contains all the extreme rays of C_A as its column vectors, then $C_A = \{Y\lambda \mid \lambda \geq 0\}$. Let $a_m^\top \in \mathbb{R}^{1 \times M}$ denote the row vectors of A for all $m \in [M]$. Then all a_m can be found by a that solves the following linear feasibility program

$$\operatorname*{find}_{a \neq 0, \lambda \geq 0} \quad \text{s.t. } Y\lambda = c, \;\; c^\top a = 0, \;\; Y^\top a \geq 0. \tag{3.6}$$

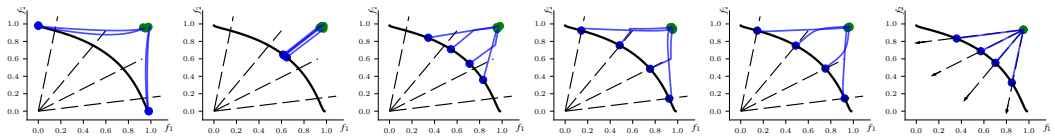
Choice of C_A for controlled ascent. If C_A is not pre-specified, and the decision maker wants to choose C_A to allow controlled ascent, it can be achieved with the following procedure. Let $F_0 = F(\theta_0)$ be the objective of the initial iterate of the algorithm, and F_{go} be the target function value along the controlled ascent direction. To ensure $F_{go} - F_0 \in -C_A$ for controlled ascent, we include $(F_0 - F_{go})/\|F_0 - F_{go}\|$ in the set of extreme rays, then take the extreme rays of the convex hull of the new set to form the columns of Y. Finally, we obtain C_A by solving (3.6).

4 Related Works

To put our work in context, we review the most relevant literature in (preference-guided) multi-objective optimization, constrained optimization, with a focus on gradient-based approaches.

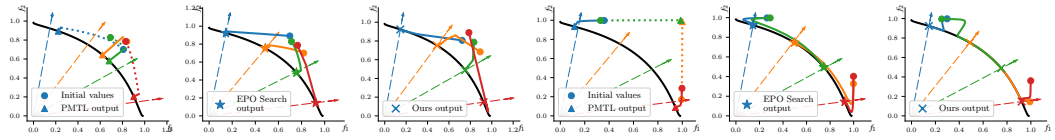
Multi-objective optimization (MOO). A straightforward approach of MOO is to use scalarization to transform MOO into a single-objective optimization problem [32]. Another popular approach focuses on finding update directions which avoid conflicts with the gradients of the objectives [40, 45, 25]. A foundational algorithm in this domain is the Multiple Gradient Descent Algorithm (MGDA) [10, 26, 6], which dynamically weights gradients to find a steepest common descent direction for all objectives. However, a single solution usually cannot capture different trade-offs on the Pareto front. This motivates the development of *preference-guided multi-objective optimization* methods.

Preferences can be modeled through weights or thresholds assigned to different objectives [32]. For example, scalarization-based methods use the ℓ_p-norm of the weighted vector-valued objective to convert the vector-valued objective into a scalar-valued objective, e.g., Linear scalarization (LS), Tchebycheff scalarization. Then the problem can be solved by single-objective optimization on the scalar objective. The ϵ-constraint methods enforce threshold constraints on different objectives, then solve the problem by constrained optimization. More recently, preferences have been modeled by preference vectors defined in the objective space. Then the problem can be formulated as finding Pareto optimal solutions satisfying the constraints defined by the preference vectors [24], or optimizing the distance to the preference vectors [30, 33]. The key difference between FERERO and these works is that FERERO can capture more flexible preferences based on a general partial order. Moreover, we provide non-asymptotic convergence guarantees for the proposed algorithms.



(a) LS (b) MGDA (c) PMTL (d) EPO (e) FERERO (f) FERERO

Figure 3: Converging solutions (blue dots) and optimization trajectories (blue lines) on the objective space of different methods on synthetic objectives given in (5.1). Dashed arrows represent pre-specified preference vectors. The green dots represent initial objective values.



(a) PMTL (b) EPO (c) FERERO (d) PMTL (e) EPO (f) FERERO

Figure 4: Outputs (colored markers) and optimization trajectories (colored lines) of different methods when initial objectives are near the Pareto front. Different colors represent different preferences.

Constrained optimization. Constrained optimization methods include primal methods, penalty and barrier methods, and primal-dual methods [4, 28]. Our proposed method is related to the primal method that finds a feasible update direction to ensure the models are feasible and improving along the optimization trajectory. To address the limitation that it usually requires a stage-one procedure to ensure the initialization is feasible, we use an adaptive approach to ensure the constraint violation is decreasing and converging to zero. This idea can also be found in sequential quadratic programming (SQP) [13]. However, compared to SQP, we use an identity matrix to approximate the Hessian of each objective, and we use an adaptive approach that automatically adjust the descent amount of objectives. Furthermore, existing SQP algorithms typically require an inner loop to solve the optimal Lagrangian multiplier, resulting in double-loop algorithms. In contrast, we develop a single-loop algorithm which is more efficient to implement.

Vector optimization. Vector optimization [11, 19] generalizes multi-objective optimization by substituting the commonly used component-wise partial order with a more general partial order, such as a general convex-cone induced partial order used in this paper. In the unconstrained setting, the MGDA method is extended to a steepest cone descent method in the vector optimization setting in [17]. Besides gradient-based vector optimization, another line of works focus on black-box vector optimization with discrete design space [3, 2]. In the constrained setting, the first-order optimality conditions are studied in [16, 43]. Algorithms based on projected gradient [9] or conditional gradient [7] are developed to solve vector optimization with parameters constrained in a set, to name a few. To our best knowledge, we are the first to design single-loop (stochastic) primal algorithms for constrained vector optimization with convergence rate guarantees.

5 Experiments

In this section, we conduct experiments to verify our theory and show the applicability of the algorithms to preference-guided multi-task learning, and multi-objective finetuning of large multi-lingual speech recognition models. We use Linear scalarization (LS), MGDA [40], PMTL [24], EPO [30], XWC-MGDA [33] as baselines for comparison.

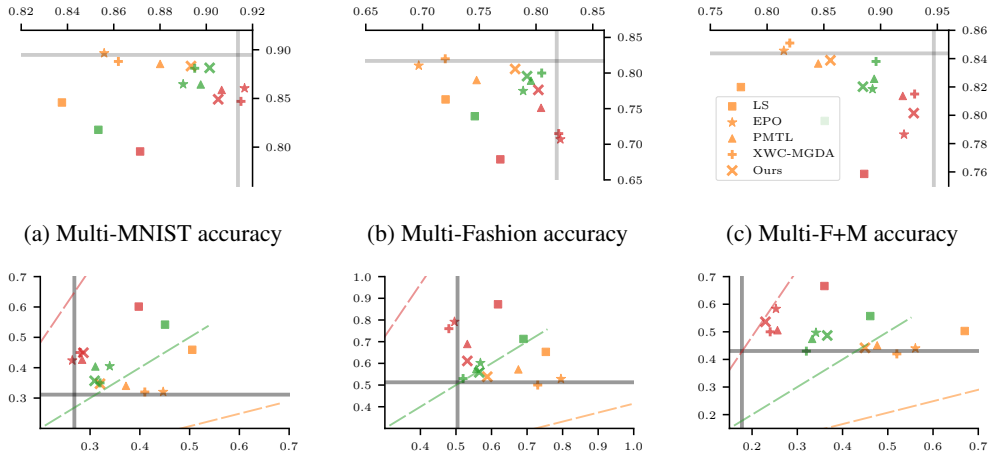
Metrics. *Objective loss and accuracy.* We report the objective losses and accuracies in classification. *Relative loss profile.* We use the element-wise product of the preference vector and the objective values as a measure of the relative loss profile. *Hypervolume.* Let $F' \in \mathbb{R}^M$ denote a reference point, and $\mathcal{S}$ denote a set of objective function values of the obtained models. Hypervolume measures the size of the dominated space of $\mathcal{S}$ relative to F', which can be computed by $H(\mathcal{S}) = \Lambda(\{q \in \mathbb{R}^M \mid \exists F \in \mathcal{S} : F \leq q \leq F'\})$, where $\Lambda(\cdot)$ denotes the Lebesgue measure. For a fair comparison, we use the Nadir point, i.e., the worst performance on single-task baselines, as the reference point F'.

Additional details. The implementation and additional experiments can be found in Appendix H.

5.1 Synthetic data

Following [24, 30, 33], the first objective we consider is

$$F(\theta) = \left(1 - e^{-\|\theta - \frac{1}{\sqrt{q}}\mathbf{1}\|_2^2},\ 1 - e^{-\|\theta + \frac{1}{\sqrt{q}}\mathbf{1}\|_2^2}\right). \tag{5.1}$$



(a) Multi-MNIST accuracy (b) Multi-Fashion accuracy (c) Multi-F+M accuracy

(d) Multi-MNIST loss (e) Multi-Fashion loss (f) Multi-F+M loss

Figure 5: Training losses and accuracies of various methods with different preferences across three image datasets. The horizontal and vertical axes represent results for objective 1 and objective 2, respectively. Different colored dashed arrows indicate various preference vectors. Different markers denote the solutions obtained by different methods, with marker colors matching the preferences.

The objective has a nonconvex Pareto front (PF). See the results of different methods in Figure 3. With uniformly generated weights from a simplex, LS only finds extreme points on the PF with one objective minimized. MGDA can only find points close to the center of the PF. PMTL can find points in the subregions but not aligned well with the exact preference vectors. Similar to EPO, in Figure 3e, our method finds points that align well with the exact preferences; and in Figure 3f, our method can handle different definitions of preferences.

We conduct another experiment in a more difficult setting where the initial objectives are close to the PF. In Figures 4a-4c, we consider a relatively easier case where the initial model is not too close to the Pareto optimal. For our method, by solving (3.6), $a_1 = [\frac{1}{\sqrt{5}}; \frac{2}{\sqrt{5}}], a_2 = [\frac{2}{\sqrt{5}}; \frac{1}{\sqrt{5}}]$. The corresponding matrix A is given by $A = [a_1, a_2]^\top$. In this setting, all methods converge to the PF, and our method takes the least number of iterations (PMTL takes 100, EPO Search takes 60, and our method takes only 10 iterations). PMTL does not align exactly with the preference vectors, while EPO and our method do. In Figures 4d-4f, PMTL and our method take 200 iterations, EPO Search takes 80 iterations. Results show that for the green and yellow preferences, PMTL moves further away from the PF in the first stage, and does not perform any update in the second stage. It converges to the PF only in 2 out of 4 cases. In contrast, with controlled ascent updates, EPO and our method can converge to the PF and trace the PF until the objectives align exactly with the preferences.

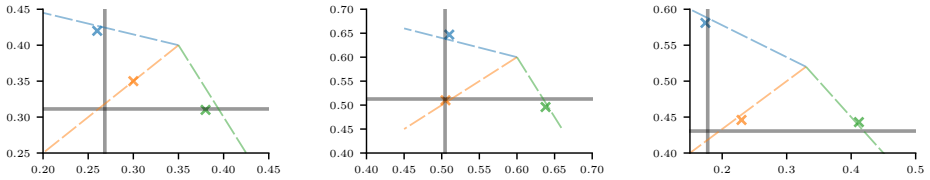
5.2 Real data

Multi-patch image classification. Following [24, 30, 33], we consider three datasets for image classification, including Multi-MNIST, Multi-Fashion, and Multi-Fashion+MNIST. The two tasks or objectives in all three datasets are to classify the top-left and the bottom-right images, respectively. For a fair comparison, we use LeNet as the backbone neural network. The training losses and accuracies of different methods given different preference vectors are plotted in Figure 5. Experiments for our method are repeated 5 times. Hypervolumes with means and standard deviations are reported in Table 2. The results for other methods in Table 2 are referenced from [33].

One limitation of EPO is that the preference is defined as a ray from the origin in the objective space, whose corresponding objectives can be unattainable, e.g., the yellow preferences in Figure 5. As a result, the losses of all methods are far away from the preference vectors. In this case, a more flexible choice of preferences is helpful to ensure preference satisfaction. To demonstrate this, we conduct experiments with more flexible preferences; see the results in Figure 6, where the obtained solutions align better with the preference lines compared to those in Figure 5. Moreover, it can perform controlled ascent updates during optimization, which cannot be achieved by PMTL or XWC-MGDA.

Table 2: Hypervolumes of different methods ($\times 10^{-2}$)

Datasets	LS	PMTL [24]	EPO [30]	XWC-MGDA [33]	FERERO
Multi-MNIST loss	1.68	1.41	1.35	1.42	**1.97**$\pm$**0.21**
Multi-Fashion loss	6.75	5.90	6.02	6.77	**7.76**$\pm$**0.18**
Multi-F+M loss	3.63	3.03	3.76	**3.89**	3.82$\pm$0.21
Multi-MNIST accuracy	0.19	0.15	0.15	0.16	**0.24**$\pm$**0.04**
Multi-Fashion accuracy	0.99	0.87	0.87	0.99	**1.17**$\pm$**0.07**
Multi-F+M accuracy	0.48	0.40	0.50	0.52	**0.53**$\pm$**0.04**
Emotion loss	0.0258	0.0230	**0.0366**	0.0348	0.0357$\pm$0.0006



(a) Multi-MNIST loss (b) Multi-Fashion loss (c) Multi-F+M loss

Figure 6: Losses and preferences of FERERO when the initial objective is close to the Pareto front.

Emotion recognition. We apply our method to predict 6 types of emotions from 593 songs on the Emotions and Music dataset [41]. We follow the experiment settings in [30], with more details summarized in Appendix H.2. The hypervolumes are reported in Table 2.

Multi-lingual speech recognition. We further apply the proposed method to the multi-objective finetuning of pre-trained multi-lingual speech models. We use the Librispeech (100 hours) [36], and AISHELL v1 [5] datasets for multi-lingual speech recognition. A conformer with 8 blocks is used as the model architecture. The total number of parameters is around 64.5M with 58.4M encoder layer parameters and the rest being the classification layer parameters. We consider the objectives associated with the speech recognition Connectionist Temporal Classification (CTC) losses in Chinese and English, denoted as f_t^{ch} and f_t^{en}, respectively. We also use the self-supervised Contrastive Predictive Coding (CPC) loss f_p for representation learning; that is

Table 3: WERs (%) on Librispeech and AISHELL v1.

Method	English	Chinese	Average
Komatsu et al. [20]	7.11	-	-
w/o CPC [39]	11.8	10.2	11.0
Init. (M2ASR) [39]	7.3	6.2	6.7
LS-FT	6.8	5.9	6.4
FERERO-FT	**5.4**	**4.9**	**5.1**

$$\min_{\theta} \quad F(\theta) := \left(f_p(\theta), f_t^{\text{ch}}(\theta), f_t^{\text{en}}(\theta)\right)^\top \quad \text{s.t.} \quad f_p(\theta) \le \epsilon_1,\ f_t^{\text{ch}}(\theta) - f_t^{\text{en}}(\theta) = \epsilon_2 \tag{5.2}$$

where the first constraint ensures to learn a good representation with $\epsilon_1 = 1.2$, and the second constraint avoids one language loss dominates the other with $\epsilon_2 = 0.5$; see more details in Appendix H.1.

Results on the word error rate (WER) are reported in Table 3. The baselines include the state-of-the-art result from Komatsu et al. [20] without an additional large language model, our own implementation of training using only the sum of supervised CTC losses (w/o CPC), the initial pre-trained M2ASR model [39] (init.), linear scalarization of all three objectives for finetuning a pre-trained model with the CPC loss (LS-FT). Results show that considering CPC loss besides the supervised CTC loss improves the average WER by 4.2%, and this can be further improved by 0.3% by finetuning with linear scalarization. However, the LS-FT model has a much better performance in Chinese compared to English. With our proposed approach, the performance gap between different languages is reduced, and the average WER is further improved by 1.3%.

6 Conclusions

In this work, we frame preference-guided multi-objective learning as a constrained vector optimization problem. Specifically, we introduce constraints and partial order to capture the absolute and relative preferences. Under this framework, we develop algorithms to solve the constrained vector optimization problem. Our proposed algorithms use a unified formulation without solving different subprograms at different stages. And they enjoy the benefit of allowing controlled ascent and escaping weak optimal solutions. Theoretical guarantees on the non-asymptotic convergence of the deterministic algorithms and their stochastic variants are provided. Experiments on benchmark datasets demonstrate the broad applicability of the proposed algorithms.

Broader Impacts and Limitations

This paper casts the preference-guided multi-objective learning as a constrained vector optimization problem and proposes an algorithm with stochastic variants and non-asymptotic convergence guarantees to solve the problem. The proposed method is applied to image classification, speech recognition, and emotion classification. The positive impact is that it is a principled method that has broad applications across various domains. There is no negative social impact.

The proposed algorithm is able to model flexible preferences but at a cost of higher per-iteration complexity compared to scalarization methods. The theoretical guarantees make standard assumptions that the objectives are lower bounded, Lipschitz continuous and smooth. These are common assumptions in the optimization literature, and can be satisfied for neural networks with smooth activation functions.

References

[1] Jaqueline S Angelo, Isabella A Guedes, Helio JC Barbosa, and Laurent E Dardenne. Multi-and many-objective optimization: present and future in de novo drug design. *Frontiers in Chemistry*, 11, 2023.

[2] Cagin Ararat and Cem Tekin. Vector optimization with stochastic bandit feedback. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 2165–2190, Valencia, Spain, 2023.

[3] Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 939–947, Cadiz, Spain, 2016.

[4] Dimitri Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series)*. Athena Scientific, 1996.

[5] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment*, pages 1–5, 2017.

[6] Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *Journal of Machine Learning Research*, 2024.

[7] Wang Chen, Xinmin Yang, and Yong Zhao. Conditional gradient method for vector optimization. *Computational Optimization and Applications*, 85(3):857–896, July 2023.

[8] Frank E Curtis, Suyun Liu, and Daniel P Robinson. Fair machine learning through constrained stochastic optimization and an epsilon-constraint method. *Optimization Letters*, pages 1–17, 2023.

[9] L. M. Graña Drummond and A.N. Iusem. A projected gradient method for vector optimization problems. *Computational Optimization and Applications*, 28:5–29, April 2004.

[10] Jean-Antoine Désidéri. Multiple-gradient Descent Algorithm (MGDA) for Multi-objective Optimization. *Comptes Rendus Mathematique*, 350(5-6), 2012.

[11] Matthias Ehrgott. *Multicriteria optimization*. Springer, Berlin; New York, 2nd ed edition, 2005.

[12] Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research*, 51:479–494, 2000.

[13] Jörg Fliege and A. Ismael F. Vaz. A method for constrained multiobjective optimization based on sqp techniques. *SIAM Journal on Optimization*, 26(4):2091–2119, 2016.

[14] Bennet Gebken, Sebastian Peitz, and Michael Dellnitz. A descent method for equality and inequality constrained multiobjective optimization problems. In *Numerical and Evolutionary Optimization*, pages 29–61. Springer, 2019.

[15] Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. In *Proc. Advances in Neural Information Processing Systems*, volume 34, pages 29630–29642, virtual, 2021.

[16] L. M. Graña Drummond, A. N. Iusem, and B. F. Svaiter. On first order optimality conditions for vector optimization. *Acta Mathematicae Applicatae Sinica, English Series*, 19(3), September 2003.

[17] L. M. Graña Drummond and B.F. Svaiter. A steepest descent method for vector optimization. *Journal of Computational and Applied Mathematics*, 175(2):395–414, March 2005.

[18] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[19] Johannes Jahn. *Vector Optimization: Theory, Applications, and Extensions*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[20] Tatsuya Komatsu, Yusuke Fujita, Jaesong Lee, Lukas Lee, Shinji Watanabe, and Yusuke Kida. Better intermediates improve CTC inference. *arXiv preprint arXiv:2204.00176*, 2022.

[21] Panagiotis Kyriakis, Jyotirmoy Deshmukh, and Paul Bogdan. Pareto policy adaptation. In *Proc. International Conference on Learning Representations*, virtual, 2021.

[22] Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, and Qingfu Zhang. Pareto set learning for expensive multi-objective optimization. In *Proc. Advances in Neural Information Processing Systems*, volume 35, New Orleans, LA, December 2022.

[23] Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. *arXiv preprint arXiv:2402.19078*, 2024.

[24] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. In *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[25] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021.

[26] Suyun Liu and Luis Nunes Vicente. The Stochastic Multi-gradient Algorithm for Multi-objective Optimization and its Application to Supervised Machine Learning. *Annals of Operations Research*, pages 1–30, 2021.

[27] Xingchao Liu, Xin Tong, and Qiang Liu. Profiling Pareto Front With Multi-Objective Stein Variational Gradient Descent. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021.

[28] David G. Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*, volume 116 of *International Series in Operations Research & Management Science*. Springer US, New York, NY, 2008.

[29] Sohvi Luukkonen, Helle W. van den Maagdenberg, Michael T.M. Emmerich, and Gerard J.P. van Westen. Artificial intelligence in multi-objective drug design. *Current Opinion in Structural Biology*, 79:102537, 2023.

[30] Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *Proc. International Conference on Machine Learning*, virtual, 2020.

[31] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proc. International Conference on Machine Learning*, pages 6755–6764, virtual, 2020.

[32] Kaisa Miettinen. *Nonlinear Multiobjective Optimization*, volume 12. Springer US, Boston, MA, 1998.

[33] Michinari Momma, Chaosheng Dong, and Jia Liu. A multi-objective/multi-task learning framework induced by pareto stationarity. In *Proc. International Conference on Machine Learning*, Baltimore, MD, 2022.

[34] Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the pareto front with hypernetworks. In *Proc. International Conference on Learning Representations*, virtual, April 2020.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[36] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015.

[37] Javier Peña, Juan C. Vera, and Luis F. Zuluaga. New characterizations of hoffman constants for systems of linear constraints. *Mathematical Programming*, 187(1):79–109, 2021.

[38] Hoang Phan, Ngoc Tran, Trung Le, Toan Tran, Nhat Ho, and Dinh Phung. Stochastic multiple target sampling gradient descent. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

[39] A F M Saif, Lisha Chen, Xiaodong Cui, Songtao Lu, Brian Kingsbury, and Tianyi Chen. M2ASR: Multilingual multi-task automatic speech recognition via multi-objective optimization. In *Interspeech 2024*, pages 1240–1244, 2024.

[40] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, December 2018.

[41] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis Vlahavas. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011:1–9, 2011.

[42] Yijun Yang, Jing Jiang, Tianyi Zhou, Jie Ma, and Yuhui Shi. Pareto policy pool for model-based offline reinforcement learning. In *Proc. International Conference on Learning Representations*, virtual, 2021.

[43] Jane J Ye and Qiji J Zhu. Multiobjective optimization problem with variational inequality constraints. *Mathematical Programming*, 96(1):139–160, 2003.

[44] Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[45] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2020.

[46] Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, Tingjun Hou, Jian Wu, et al. Sample-efficient multi-objective molecular optimization with gflownets. In *Proc. Advances in Neural Information Processing Systems*, volume 36, New Orleans, LA, 2023.

Appendix for " FERERO: A Flexible Framework for Preference-Guided Multi-Objective Learning "

Table of Contents

A Notations

A summary of notations used in this work is listed in Table 4 for ease of reference.

Recall that given vectors v, w, we use $v < w$ and $v \leq w$ to denote $v_i < w_i$ for all i, and $v_i \leq w_i$ for all i, respectively. We use $v \lneq w$ to denote $v \leq w$ and $v \neq w$, and define $>, \geq, \gneq$ analogously. In the proof, we use $\|\cdot\|$ to denote the ℓ_2-norm, and $\|\cdot\|_1$ to denote the ℓ_1-norm. We use $|\cdot|_{\mathrm{ab}}$ to denote the operator that takes element-wise absolute value of a matrix. We use $\mathbf{1}$ and 0 to denote the all-one and all-zero vectors, respectively. Their dimensions are specified only when they are not clear in the context. We use $[v, w]$ to represent column concatenation of matrices or vectors, and use $[v; w]$ to represent row concatenation of matrices or vectors.

B Related Works and Comparison

In this section, we provide a detailed review and comparison of additional related works in multi-task/objective learning, vector optimization, and Pareto front approximation.

Table 4: Notations and their descriptions.

Notations	Descriptions
$\theta \in \mathbb{R}^q$	Model parameter, or decision variable
ξ	Stochastic samples during training
$f_{\xi,m}(\theta)$, $f_m(\theta)$	A scalar-valued objective function evaluated on data point ξ, with $f_{\xi,m} : \mathbb{R}^q \to \mathbb{R}$, or on dataset D, f_m, with $f_m := \frac{1}{\lvert D\rvert} \sum_{\xi \in D} f_{\xi,m}(\theta)$
$\nabla f_m(\theta)$	Gradient of $f_m(\theta)$, with $\nabla f_m : \mathbb{R}^q \to \mathbb{R}^q$
$F_\xi(\theta)$, $F(\theta)$	A vector-valued objective function evaluated on data point ξ, with $F_\xi : \mathbb{R}^q \to \mathbb{R}^M$, or on dataset D, with $F := \frac{1}{\lvert D\rvert} \sum_{\xi \in D} F_\xi(\theta)$
$\nabla F(\theta)$	Gradient of $F(\theta)$, with $\nabla F : \mathbb{R}^q \to \mathbb{R}^{q \times M}$
α	Step size to update model parameter θ
γ	Step size to update multiplier λ

B.1 Extended discussion of related works

In this section, we provide an extended discussion of the works that are closely related to ours.

Pareto front approximation. Pareto front approximation aims to find multiple different solutions whose objective values approximate the Pareto front. Scalarization-based methods can be used to approximate the Pareto front by enumerating different weights of the objectives. However, they cannot find solutions on the nonconvex part of the Pareto front [32]. Decomposition-based methods partition the objective space into different subsets with constraints that represent different trade-off preferences, and solve the constrained multi-objective optimization subproblems with gradient-based or evolutionary algorithms [24, 15]. Probabilistic inference methods update a set of models following a distribution that converges to Pareto stationary [27, 38]. The expected update direction of the models typically follows the steepest common descent direction for all objectives. Pareto set learning methods use a neural network to learn a mapping from user preferences to corresponding models. The learned neural network is able to generate different models with different input user preferences [34, 42, 21, 22].

B.2 A detailed comparison with existing works

Preferences as linear constraints of objectives. Different constraints S partition the objectives into sub-regions, as shown in Figure 1. Many preferences can be modeled by linear equality or inequality constraints [24, 30, 33]. For example, below we list different choices of C for different methods in Figure 1.

(a) $B_g = [0, I_{2:M}]^\top \in \mathbb{R}^{M \times M}, b = -[0, \epsilon_2, \ldots, \epsilon_M]^\top$;

(b) $B_h \in \mathbb{R}^{(M-1) \times M}, b = 0$;

In Figure 1a, the preferences are based on the function values of f_1 controlled by different thresholds, corresponding to the inequality constraints defined by (a). In Figure 1b, the constraints are that the objectives $F(\theta)$ should lie on one of the preference vectors v, therefore should satisfy the equality constraint $B_h F(\theta) = 0$.

Detailed comparison with the most relevant works. Below we provide a fine-grained comparison with some existing works in Table 5, as an extension of Table 1.

In terms of preference modeling, the scalarization-based methods such as Linear Scalarization and Smooth Tchebycheff scalarization use weight of different objectives to model preferences. They are not flexible enough to capture preferences illustrated in Figure 1. PMTL uses a constrained multi-objective optimization formulation, with preferences modeled by inequalities. EPO models the preference by an r^{-1} ray, same as the example given in Figure 1b. (X)WC-MGDA uses a shifted ray not necessarily from the origin to model the preferences. In all of these works, they only model the absolute preferences that define the preferred objective values. In contrast, we also consider the relative preference that define the relative improvement directions of objectives.

In addition to the comparison in Table 1, our framework enjoys additional benefits including the ability to escape weak optimal solutions and to maintain scale-invariance. These abilities are attributed to the subprogram that is adaptive to the objective values, as detailed in Lemma 4.

Table 5: Comparison to existing PMOL methods, extension of Table 1.

Method	Handle nonconvex PF	General partial order	Single subprogram w/o computing active index	Scale invariance	Escape weak optimal	Provable CQ
Linear Scalarization	✗	✗	✓	✗	✗	-
(Smooth) Tchebycheff [23]	✓	✗	✓	✗	✗	-
PMTL [24]	✓	✗	✗	✗	✗	assume LICQ
EPO [30]	✓	✗	✗	✗	✓	✗
(X)WC-MGDA [33]	✓	✗	✗	✗	✓	✗
FERERO (ours)	✓	✓	✓	✓	✓	prove calmness

Below, we further summarize the reasons behind the benefits of our proposed method. We use "→" to indicate the reasons on the left and the corresponding benefits on the right.

- Flexible preference
  - relative preference (by general partial order) → allow controlled ascent
  - absolute preference (by constraints) →
    - handle nonconvex Pareto Front
    - equality constraints → align exactly to preference vector
    - constraints are linear functions of objectives → provable CQ
- Adaptive subprogram
  - adaptive to objectives →
    - scale invariance
    - ability to escape weak optimality
  - adaptive to constraints → single subprogram w/o computing active indices → non-asymptotic convergence

C Preliminaries

In this section we introduce preliminaries on the general cone-induced partial ordering and the corresponding optimality conditions for completeness since we use these concepts in our proofs. Then we discuss the relation between the Pareto optimality and the optimality induced by a general polyhedral cone.

C.1 General cone-induced partial ordering

In this section, we introduce basic definitions, lemmas, propositions, and theorems in vector optimization, including the cone-induced partial ordering, the minimum and weakly minimum associated with the partial ordering in real linear space, and necessary conditions for minimum. These concepts are defined in [19]. We restate them following our notations for completeness. We denote Z as a real linear space, C, S as subsets in Z, and w, x, y, z as points or elements in Z, 0_Z as the zero vector in the space Z.

Definition 3 (Cone). *Let C be a nonempty subset of a real linear space Z. The set C is called a* cone*, if $y \in C, \lambda \geq 0 \Longrightarrow \lambda y \in C$.*

Lemma 3 (Convex cone). *A cone C in a real linear space is convex if and only if $C + C \subset C$.*

Definition 4 (Partially ordered linear space). *A real linear space equipped with a partial ordering is a partially ordered linear space.*

Theorem 5. *(a) If $\leq$ is a partial ordering on Z, then the set $C := \{z \in Z \mid 0_Z \leq z\}$ is a convex cone. If, in addition, $\leq$ is antisymmetric, then C is pointed.*

(b) If C is a convex cone in Z, then the binary relation $\leq_C := \{(x, y) \in Z \times Z \mid y - x \in C\}$ is a partial ordering on Z. If, in addition, C is pointed, then $\leq_C$ is antisymmetric.

Definition 5 (Ordering cone). *A convex cone characterizing a partial ordering in a real linear space is an ordering cone.*

Definition 6 (Cone-induced partial ordering)**.** *Let C be a closed pointed convex cone of $\mathbb{R}^M$, with nonempty interior. The partial order in $\mathbb{R}^M$ induced by $C, \leq_C$ is defined by*

$$u \leq_C v, \ \text{ if } \ v - u \in C. \tag{C.1}$$

The relation induced by $\operatorname{int}(C)$ in $\mathbb{R}^M$, $<_C$ is defined by

$$u <_C v, \ \text{ if } \ v - u \in \operatorname{int}(C). \tag{C.2}$$

Definition 7 (C-minimum and C-weakly minimum)**.** *Let S be a nonempty subset of a partially ordered linear space with an ordering cone C, then*
(a) an element $z \in S$ is called a C-minimum of the set S, if $(\{z\} - C) \cap S \subset \{z\} + C$, in other words, there exists no other $z' \in S$ with $z' \leq_C z$ and $z' \neq z$;
(b) an element $z \in S$ is called a C-weakly minimum of the set S, if $(\{z\} - \operatorname{int}(C)) \cap S = \emptyset$, where $\operatorname{int}(C) \neq \emptyset$ is the algebraic interior of C, in other words, there exists no other $z' \in S$ with $z' <_C z$ and $z' \neq z$.

Definition 8 (C-stationary)**.** *A point $\theta \in \mathbb{R}^q$ is C-stationary if there is no first-order common descent direction $d \in \mathbb{R}^q$ that $\nabla F(\theta)^\top d \in -\operatorname{int}(C)$, i.e., $\operatorname{range}(\nabla F(\theta)^\top) \cap (-\operatorname{int}(C)) = \emptyset$.*

C.2 Necessary and sufficient conditions for C-optimality

Note that, when $C = \mathbb{R}^M_+ := \{z \in \mathbb{R}^M \mid z_m \geq 0 \text{ for all } m \in [M]\}$, C-minimum and C-weakly minimum in Definition 7 are Pareto minimum and weakly Pareto minimum, respectively. Recall that $F : \mathbb{R}^q \to \mathbb{R}^M$ is a continuously differentiable function. The problem we consider is to find the unconstrained C-minimizers of F, denoted as $\min_C F(\theta)$ with $\theta \in \mathbb{R}^q$. We then proceed to introduce the relation between C-stationarity and Pareto stationarity in this section.

Proposition 1. *Let C be a closed convex pointed cone.*
1) Suppose $C \subseteq \mathbb{R}^M_+$. If θ is Pareto stationary, θ is C-stationary. In other words, C-stationarity is a necessary condition for Pareto stationarity.
2) Suppose $\mathbb{R}^M_+ \subseteq C$. if θ is C-stationary, θ is Pareto stationary. In other words, C-stationarity is a sufficient condition for Pareto stationarity.

Proof of Proposition 1. 1) By definition, if θ is Pareto stationary, then $\operatorname{range}(\nabla F(\theta)^\top) \cap (-\operatorname{int}(\mathbb{R}^M_+)) = \emptyset$. Since $C \subseteq \mathbb{R}^M_+$, then $-\operatorname{int}(C) \subseteq -\operatorname{int}(\mathbb{R}^M_+)$, and we have

$$\operatorname{range}(\nabla F(\theta)^\top) \cap (-\operatorname{int}(C)) \subseteq \operatorname{range}(\nabla F(\theta)^\top) \cap (-\operatorname{int}(\mathbb{R}^M_+)) = \emptyset. \tag{C.3}$$

Therefore, θ is C-stationary.

Following similar arguments, 2) can also be proved. □

D Proof of Auxiliary Lemmas

In this section, we provide proof of the main theoretical results in this paper.

D.1 Lagrangian of the subprogram

Proof of subprogram reformulation. Define the Lagrangian function

$$\begin{aligned} L(c, d, \lambda_f, \lambda_g, \lambda_h) :=& c + \frac{1}{2}\|d\|^2 + \lambda_f^\top \left(A\nabla F(\theta)^\top d - c(\mathbf{1}^\top AF(\theta))^{-1} AF(\theta)\right) \\ &+ \lambda_g^\top \left(B_g \nabla F(\theta)^\top d + c_g G(\theta)\right) + \lambda_h^\top \left(B_h \nabla F(\theta)^\top d + c_h H(\theta)\right) \end{aligned} \tag{D.1}$$

where $\lambda_f \in \mathbb{R}^M_+$, $\lambda_g \in \mathbb{R}^{M_g}_+$, $\lambda_h \in \mathbb{R}^{M_h}$. By the first-order optimality condition w.r.t. d and c, we can obtain that

$$d^* + \nabla F(\theta)(A^\top \lambda_f^* + B_g^\top \lambda_g^* + B_h^\top \lambda_h^*) = 0; \tag{D.2}$$

$$\mathbf{1}^\top AF(\theta) - {\lambda_f^*}^\top AF(\theta) = 0. \tag{D.3}$$

Combining the last equation with $\lambda_f \in \mathbb{R}^M_+$, we obtain $\lambda_f^* \in \Omega_{\lambda_f}$. Plugging the above results into the Lagrangian function gives

$$[\lambda_f^*; \lambda_g^*; \lambda_h^*] \in \underset{[\lambda_f;\lambda_g;\lambda_h]\in\Omega_\lambda}{\arg\min} \frac{1}{2}\|\nabla F(\theta)(A^\top \lambda_f + B_g^\top \lambda_g + B_h^\top \lambda_h)\|^2 - c_g \lambda_g^\top G(\theta) - c_h \lambda_h^\top H(\theta) \tag{D.4}$$

which leads to the dual form in (2.3). Since (2.1) is a constrained convex optimization problem where Slater's condition holds, therefore, the duality gap is zero. □

D.2 First-order necessary optimality conditions

We then discuss the first-order necessary optimality conditions for problem (PMOL). We begin the discussion with the geometric notions of improving and feasible directions.

Improving directions. The improvement directions are defined as generalized common descent directions so that the iterates strictly improve or dominate the previous iterates based on C_A, i.e., $F(\theta_t) - F(\theta_{t+1}) \in \text{int}(C_A)$. Denote $d_t \in \mathbb{R}^q$ as an update direction at iteration t, and $\alpha_t > 0$ as the step size at the t-th iteration. The general update equation given update direction d_t is $\theta_{t+1} = \theta_t + \alpha_t d_t$. Based on first-order Taylor expansion, the amount of improvement at iteration t can be approximately expressed as $F(\theta_t) - F(\theta_{t+1}) \approx -\alpha_t \nabla F(\theta_t)^\top d_t \in \text{int}(C_A)$. We term such directions the general C_A-improving directions. The cone of C_A-improving directions at x is

$$D_{C_A} = \{d \in \mathbb{R}^q \mid \nabla F(\theta)^\top d \in -\text{int}(C_A)\}. \tag{D.5}$$

When $A = I_M$, they are common descent directions.

Feasible directions. Similar to the concept in constrained single objective optimization, the feasible directions are those that ensure $F(\theta_t + \alpha_t d_t) \in S$. We rewrite problem (PMOL) with explicit C_A-induced partial ordering as

$$\min{}_{C_A} \ F(\theta) \ \text{ s.t. } \ G(\theta) \leq 0, \ H(\theta) = 0. \tag{PMOL}$$

where $G : \mathbb{R}^q \to \mathbb{R}^{M_g}, H : \mathbb{R}^q \to \mathbb{R}^{M_h}$ are linear functions of F, and are differentiable. Let $I = \{i \mid G_i(\theta) = 0\}$ be the index set of the active inequality constraints in $G(\theta)$, and $G_I(\theta) = [\cdots, G_i(\theta), \cdots]^\top$ for $i \in I$. A subset of the feasible directions described by the gradients of the equality and active inequality constraints at θ is given by

$$D_g = \{d \in \mathbb{R}^q \mid \nabla G_I(\theta)^\top d < 0\}, \quad D_H = \{d \in \mathbb{R}^q \mid \nabla H(\theta)^\top d = 0\}. \tag{D.6}$$

A necessary optimality condition is that there exists no feasible and improving directions at θ, i.e., $D_{C_A} \cap D_g \cap D_h = \emptyset$. An algebraic description of the necessary optimality conditions for (PMOL) is summarized below.

Proposition 2 (First-order necessary optimality conditions for (PMOL))**.** *Let $C_A \coloneqq \{y \in \mathbb{R}^M \mid Ay \geq 0\}$ that satisfies $\text{int}(C_A) \neq \emptyset$. If $\bar{\theta}$ solves (PMOL) locally, then there exists $\lambda_f \in \mathbb{R}^M_+$, $\lambda_g \in \mathbb{R}^{M_g}_+$, $[\lambda_f; \lambda_g] \neq 0$, and $\lambda_h \in \mathbb{R}^{M_h}$ that*

$$\nabla F(\bar{\theta}) A^\top \lambda_f + \nabla G(\bar{\theta}) \lambda_g + \nabla H(\bar{\theta}) \lambda_h = 0, \ \text{ and } \ \lambda_g^\top [-G(\bar{\theta})]_+ = 0 \tag{D.7}$$

Proof of Proposition 2. The geometric description $D_{C_A} \cap D_g \cap D_h = \emptyset$ is equivalent to that the linear system below w.r.t. d is inconsistent

$$\begin{bmatrix} A\nabla F(\bar{\theta})^\top \\ \nabla G_I(\bar{\theta})^\top \end{bmatrix} d < 0 \ \text{ and } \ \nabla H(\bar{\theta})^\top d = 0. \tag{D.8}$$

By the Motzkin's transposition theorem, system (D.8) being inconsistent is equivalent to that the following linear system w.r.t. p, λ_h has a solution with $p \gneq 0$

$$\begin{bmatrix} \nabla F(\bar{\theta}) A^\top & \nabla G_I(\bar{\theta}) \end{bmatrix} p + \nabla H(\bar{\theta}) \lambda_h = 0. \tag{D.9}$$

Letting $p = [\lambda_f; \lambda_{g,I}]$, where $\lambda_{g,I} = [\cdots; \lambda_{g,i}; \cdots], i \in I$, and $\lambda_{g,i'} = 0$, for all $i' \notin I$ completes the proof. □

Remark 6. *Notice that, Proposition 2 provides a Fritz John (FJ)-type first-order necessary optimality condition, which has been discussed in prior works such as [43, Theorem 1.2] with additional variational inequality constraints, and [16, Section 3, (2)-(5)] with inequalities constraints only. In the FJ-type necessary optimality condition, the multiplier λ_f associated with the objective $F(\theta)$ can be zero if $|I| \geq 1$, which is undesirable. We need additional constraint qualifications to ensure the condition in (D.7) with $\lambda_f \neq 0$ is also a necessary optimality condition, i.e., the KKT condition, which is equivalent to $\mu_0 = 1$, and without considering the variational inequality constraints in [43, Theorem 1.2]. The constraint qualification is discussed in detail in Appendix D.3.2.*

D.3 Properties of PMOL

In this section, we discuss the properties of PMOL and their proofs. These include the properties of the subprogram in Lemma 1, and the calmness CQ of PMOL in Lemma 2.

D.3.1 Proof of Lemma 1: properties of the subprogram

Lemma 4 (Additional properties of the subprogram)**.** *For the subprogram (2.3), the following properties hold:*
1. The solution $d^*(\theta)$ is unique.
2. If θ is a local weak optimal solution with $AF(\theta) > 0$, then $d^*(\theta) = 0$, $\psi(\theta) = 0$. Otherwise, if θ is not a local weak optimal solution, then $d^*(\theta) \neq 0$, $\psi(\theta) < 0$, and when θ is feasible,

$$2\psi(\theta) \leq -\|d^*(\theta)\|^2 < 0. \tag{D.10}$$

3. (Ability to escape weak optimal solutions). Let θ be a weak optimal solution, with $(AF(\theta))_m = 0$ for some $m \in [M]$. If there exists feasible and non-strictly improving directions at θ with $A\nabla F(\theta)^\top d \lneq 0$, then $d^*(\theta) \neq 0$, $\psi(\theta) < 0$. Otherwise, if there exists no feasible and non-strictly improving directions at θ with $A\nabla F(\theta)^\top d \lneq 0$, then $d^*(\theta) = 0$, $\psi(\theta) = 0$.
4. (Scale invariance) Suppose there are only equality constraints, i.e., $M_g = 0$, and $M_h = M - 1$, B_h is full row rank and is selected such that $B_h(F(\theta_1) - F(\theta_2)) = 0$ with $F(\theta_1), F(\theta_2)$ being two different reference points in the objective space. For all $\theta \in \mathbb{R}^q$ that are feasible, i.e., $H(\theta) = 0$, when $A = I$, the normalized solution $d^*(\theta)/\|d^*(\theta)\|$ does not change when the objective $F(\theta)$ is scaled by an arbitrary positive diagonal matrix.

Proof of Lemma 4. For **Property-1**, the uniqueness of $d^*(\theta)$ follows from the strict convexity of the objective function w.r.t. the direction d.

For **Property-2**, in the first case if θ is a local optimal solution, by definition, there exists no feasible and improving directions d such that $A\nabla F(\theta)^\top d < 0$. Let $\Omega_d(\theta)$ be the set of $d \in \mathbb{R}^q$ that satisfy the constraints in (2.1), i.e.,

$$\Omega_d(\theta) \coloneqq \{d \in \mathbb{R}^q \mid B_g \nabla F(\theta)^\top d + c_g G(\theta) \leq 0, B_h \nabla F(\theta)^\top d + c_h H(\theta) = 0\}. \tag{D.11}$$

Then, since $AF(\theta) > 0$, for all $d \in \Omega_d(\theta)$,

$$\max_{m \in [M]} (A\nabla F(\theta)^\top d)_m \geq 0 \tag{D.12}$$

$$\text{and} \quad \max_{m \in [M]} (A\nabla F(\theta)^\top d)_m / (AF(\theta))_m \geq 0. \tag{D.13}$$

And since $AF(\theta) > 0$, it holds that

$$\begin{aligned} \psi(\theta) &\coloneqq \min_{(d,c) \in \Omega_d(\theta) \times \mathbb{R}} c + \frac{1}{2}\|d\|^2 \\ &= \min_{d \in \Omega_d(\theta)} \max_{m \in [M]} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d\|^2 \geq 0 \end{aligned} \tag{D.14}$$

with $\psi(\theta) = 0$ attainable by taking $d = 0 \in \Omega_d(\theta)$. The first case of Property-2 is proved.

In the second case, if θ is not a local weak optimal solution, then there exists $d \in \Omega_d(\theta)$ such that $A\nabla F(\theta)^\top d < 0$. Taking $\sigma = -\max_{m \in [M]} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/\big((AF(\theta))_m \|d\|^2\big)$, and $d_\sigma = \sigma d$, then

$$\psi(\theta) \coloneqq \min_{(d,c) \in \Omega_d(\theta) \times \mathbb{R}} c + \frac{1}{2}\|d\|^2$$

$$
\begin{aligned}
&= \min_{d\in\Omega_d(\theta)} \max_{m\in[M]} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d\|^2 \\
&= \max_{m\in[M]} (A\nabla F(\theta)^\top d^*(\theta))_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d^*(\theta)\|^2 \\
&< \max_{m\in[M]} (A\nabla F(\theta)^\top d_\sigma)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d_\sigma\|^2 \\
&= \sigma \max_{m\in[M]} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\sigma^2\|d\|^2 = -\frac{1}{2}\sigma^2\|d\|^2 < 0. \quad \text{(D.15)}
\end{aligned}
$$

Thus $d^*(\theta) \neq 0$. Recall that

$$
d^*(\theta) = -\nabla F(\theta)\Big(A^\top \lambda_f^* + B_g^\top \lambda_g^* + B_h^\top \lambda_h^*\Big) \quad \text{(D.16)}
$$

where by the feasibility and optimality conditions,

$$
\begin{aligned}
{\lambda_h^*}^\top \big(B_h \nabla F(\theta)^\top d^*(\theta) + c_h H(\theta)\big) &= 0, && \text{(D.17a)} \\
{\lambda_g^*}^\top \big(B_g \nabla F(\theta)^\top d^*(\theta) + c_g G(\theta)\big) &= 0, && \text{(D.17b)} \\
{\lambda_f^*}^\top \big(A\nabla F(\theta)^\top d^*(\theta) - c^*(\mathbf{1}_M^\top AF(\theta))^{-1} AF(\theta)\big) &= 0. && \text{(D.17c)}
\end{aligned}
$$

Combining the above with (D.16), we have

$$
\begin{aligned}
\|d^*(\theta)\|^2 &= -\, d^*(\theta)^\top \nabla F(\theta)\Big(A^\top \lambda_f^* + B_g^\top \lambda_g^* + B_h^\top \lambda_h^*\Big) \\
&= -\, d^*(\theta)^\top \nabla F(\theta) A^\top \lambda_f^* + c_h {\lambda_h^*}^\top H(\theta) + c_g {\lambda_g^*}^\top G(\theta) \\
&\leq -\, c^*(\theta)(\mathbf{1}^\top AF(\theta))^{-1} {\lambda_f^*}^\top AF(\theta) = -c^*(\theta) \quad \text{(D.18)}
\end{aligned}
$$

where the last inequality uses the fact that θ is feasible, and $G(\theta) \leq 0$, $H(\theta) = 0$.

Then it holds that

$$
2\psi(\theta) = 2c^*(\theta) + \|d^*(\theta)\|^2 \leq -\|d^*(\theta)\|^2 < 0. \quad \text{(D.19)}
$$

Therefore, Property-2 holds.

For **Property-3**, let $I \subseteq [M]$ be the set such that $(AF(\theta))_m = 0$ for all $m \in I$, then (2.1) is equivalent to

$$
\begin{aligned}
\psi(\theta) = \min_{(d,c)\in\mathbb{R}^q\times\mathbb{R}}\ & c + \frac{1}{2}\|d\|^2 && \text{SP1w} \\
\text{s.t.}\ & (A\nabla F(\theta)^\top d)_m - c(\mathbf{1}^\top AF(\theta))^{-1}(AF(\theta))_m \leq 0, \ \text{ for all } m \in [M]\setminus I \\
& (A\nabla F(\theta)^\top d)_m \leq 0, \ \text{ for all } m \in I \\
& B_g \nabla F(\theta)^\top d + c_g G(\theta) \leq 0 \\
& B_h \nabla F(\theta)^\top d + c_h H(\theta) = 0
\end{aligned}
$$

In the first case, if there exists feasible and non-strictly improving directions at θ with $A\nabla F(\theta)^\top d \lneq 0$, then such $d \neq 0$, $d \in \Omega_d$. Following similar arguments as (D.15) by taking $\sigma = -\max_{m\in[M]\setminus I}(A\nabla F(\theta)^\top d)_m(\mathbf{1}^\top AF(\theta))/\big((AF(\theta))_m\|d\|^2\big)$, and $d_\sigma = \sigma d$, then

$$
\begin{aligned}
&\psi(\theta) \coloneqq \min_{(d,c)\in\Omega_d(\theta)\times\mathbb{R}} c + \frac{1}{2}\|d\|^2 \\
&= \min_{d\in\Omega_d(\theta)} \max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d\|^2 \\
&< \max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d_\sigma)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d_\sigma\|^2 \\
&= \sigma \max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d)_m (\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\sigma^2\|d\|^2 = -\frac{1}{2}\sigma^2\|d\|^2 < 0. \quad \text{(D.20)}
\end{aligned}
$$

And the corresponding $d^*(\theta) \neq 0$.

In the second case, if there exists no feasible and non-strictly improving directions at θ, then for all $d \in \Omega_d(\theta)$,

$$\max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d)_m \geq 0 \tag{D.21}$$

$$\text{and} \quad \max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d)_m/(AF(\theta))_m \geq 0. \tag{D.22}$$

And since $(AF(\theta))_m > 0$ for all $m \in [M] \setminus I$, it holds that

$$\begin{aligned}\psi(\theta) &\coloneqq \min_{(d,c)\in\Omega_d(\theta)\times\mathbb{R}} c + \frac{1}{2}\|d\|^2 \\ &= \min_{d\in\Omega_d(\theta)} \max_{m\in[M]\setminus I} (A\nabla F(\theta)^\top d)_m(\mathbf{1}^\top AF(\theta))/(AF(\theta))_m + \frac{1}{2}\|d\|^2 \geq 0\end{aligned} \tag{D.23}$$

with $\psi(\theta) = 0$ if and only if $d = 0 \in \Omega_d(\theta)$.

Combining the above arguments, Property-3 is proved.

For **Property-4**, let $d^*(\theta)$ be the solution to the original problem (2.1) without inequality constraints. Using the fact that $H(\theta) = 0$, and letting $\lambda = A^\top\lambda_f + B_h^\top\lambda_h = \lambda_f + B_h^\top\lambda_h$, then the original dual problem can be written as

$$\begin{aligned} &d^*(\theta) = -\nabla F(\theta)\lambda^* \\ &\text{s.t. } \lambda^* \in \arg\min_{\lambda\in\Omega_{\tilde{\lambda}}(\theta)} \varphi(\lambda;\theta) \coloneqq \frac{1}{2}\|\nabla F(\theta)\lambda\|^2 \end{aligned} \tag{D.24}$$

where $\Omega_{\tilde{\lambda}}(\theta) = \big(\Omega_{\lambda_f}(\theta)\big) + B_h^\top\big(\mathbb{R}^{M_h}\big)$, and $\Omega_{\lambda_f}(\theta) = \{\lambda_f \in \mathbb{R}_+^M \mid {\lambda_f}^\top F(\theta) = \mathbf{1}^\top F(\theta)\}$.

Suppose the objective is scaled by a positive diagonal matrix $\Lambda \in \mathbb{R}^{M\times M}$, then the scaled subprogram has a dual given by

$$\begin{aligned} &d^*(\theta) = -\nabla F(\theta)\Lambda\lambda^* \\ &\text{s.t. } \lambda^* \in \arg\min_{\lambda\in\Omega_{\tilde{\lambda}}(\theta;\Lambda)} \varphi(\lambda;\theta) \coloneqq \frac{1}{2}\|\nabla F(\theta)\Lambda\lambda\|^2 \end{aligned} \tag{D.25}$$

where $\Omega_{\tilde{\lambda}}(\theta;\Lambda) = \big(\Omega_{\lambda_f}(\theta;\Lambda)\big) + {B'_h}^\top\big(\mathbb{R}^{M_h}\big)$, and $\Omega_{\lambda_f}(\theta;\Lambda) = \{\lambda_f \in \mathbb{R}_+^M \mid {\lambda_f}^\top\Lambda F(\theta) = \mathbf{1}^\top\Lambda F(\theta)\}$. Letting $\lambda' = \Lambda\lambda$, then

$$\begin{aligned} &d^*(\theta) = -\nabla F(\theta)\lambda'^* \\ &\text{s.t. } \lambda'^* \in \arg\min_{\lambda'\in\Omega_{\tilde{\lambda}'}(\theta;\Lambda)} \varphi(\lambda;\theta) \coloneqq \frac{1}{2}\|\nabla F(\theta)\lambda'\|^2 \end{aligned} \tag{D.26}$$

where $\Omega_{\tilde{\lambda}'}(\theta;\Lambda) = \Lambda\big(\Omega_{\lambda_f}(\theta;\Lambda)\big) + \Lambda{B'_h}^\top\big(\mathbb{R}^{M_h}\big)$. The set $\Lambda\big(\Omega_{\lambda_f}(\theta;\Lambda)\big)$ can be written as

$$\begin{aligned}\Lambda\big(\Omega_{\lambda_f}(\theta;\Lambda)\big) &= \{\Lambda\lambda_f \mid \lambda_f \in \mathbb{R}_+^M, {\lambda_f}^\top\Lambda F(\theta) = \mathbf{1}^\top\Lambda F(\theta)\} \\ &= \{\lambda'_f \in \mathbb{R}_+^M \mid F(\theta)^\top{\lambda_f}' = \mathbf{1}^\top\Lambda F(\theta)\}.\end{aligned} \tag{D.27}$$

Notice that,

$$F(\theta)^\top{\lambda_f}' = \mathbf{1}^\top\Lambda F(\theta) = \mathbf{1}^\top F(\theta)c_s \tag{D.28}$$

where $c_s = \mathbf{1}^\top\Lambda F(\theta)/(\mathbf{1}^\top F(\theta))$. Therefore, $\Lambda\big(\Omega_{\lambda_f}(\theta;\Lambda)\big) = c_s\big(\Omega_{\lambda_f}(\theta)\big)$.

Also note that, $B_h \in \mathbb{R}^{(M-1)\times M}$ is full row rank, and is selected based on $F(\theta)$, which satisfies

$$B_h(F(\theta_1) - F(\theta_2)) = 0 \tag{D.29}$$

where $F(\theta_1), F(\theta_2)$ are two reference points which fully defines the kernel of B_h. Similarly, when $F(\theta)$ is scaled by Λ, the corresponding B'_h satisfies

$$B'_h\Lambda(F(\theta_1) - F(\theta_2)) = 0. \tag{D.30}$$

This further implies

$$\Lambda {B'_h}^\top(\mathbb{R}^{M_h}) = \mathrm{range}(\Lambda {B'_h}^\top) = \ker(B'_h\Lambda)^\perp = \ker(B_h)^\perp = B_h(\mathbb{R}^{M_h}) = c_s B_h(\mathbb{R}^{M_h}). \quad \text{(D.31)}$$

Combining with $\Lambda\big(\Omega_{\lambda_f}(\theta;\Lambda)\big) = c_s\big(\Omega_{\lambda_f}(\theta)\big)$, it holds that

$$\Omega_{\tilde{\lambda}'}(\theta;\Lambda) = c_s\Omega_{\tilde{\lambda}}(\theta). \quad \text{(D.32)}$$

Therefore, the solution of $\tilde{\lambda}$ and λ' is only subject to a scaling factor, which does not change the direction of $d^*(\theta)$. This proves Property-4, the scale invariance. □

D.3.2 Proof of Lemma 2: calmness of PMOL

Example 1. *Let $F:\mathbb{R}^q\to\mathbb{R}^2$. Consider the problem below as a special case of (PMOL), given by*

$$\min_{\mathbb{R}^2_+} F(\theta) \ \text{ s.t. } \ f_2(\theta) = \min f_2(\theta). \quad \text{(D.33)}$$

For $\bar{\theta} = \arg\min_{\theta\in\mathbb{R}^q} f_2(\theta)$, we have $\nabla f_2(\bar{\theta}) = 0$, and $\bar{\theta}$ satisfies (D.7) with $\lambda = [0,1]^\top \neq 0$ and $\lambda_h = 1$. However, $\nabla H(\bar{\theta}) = \nabla f_2(\bar{\theta}) = 0$ violates the LICQ, the Slater's CQ, and the MFCQ.

Below we restate the definition of the Calmness condition for PMOL [43], which generalizes the calmness condition in single-objective optimization.

Definition 9 (Calmness condition for PMOL [43, Restatement of Definition 4.5])**.** *Let $\bar{\theta}$ be a local solution to (PMOL). We say the PMOL problem satisfies the calmness condition at $\bar{\theta}$ provided that there exists $\epsilon > 0$ and a Lipschitz function $\phi:\mathbb{R}^{M_g+M_h}\to\mathbb{R}^M$ satisfying $\phi(0,0)=0$ such that there exists no $(\theta,p,q)\in[(\bar{\theta},0,0)+\epsilon\mathcal{B}]/\{(\bar{\theta},0,0)\}$ satisfying*

$$G(\theta) + p \le 0, \quad \text{(D.34a)}$$
$$H(\theta) + q = 0, \quad \text{(D.34b)}$$
$$F(\theta) - F(\bar{\theta}) + \phi(p,q) \in -\mathrm{int}(C_A). \quad \text{(D.34c)}$$

Our proof relies on the following general version of Hoffman error bound, which bounds the distance of a point to a nonempty solution set defined by constraints by a measure of the constraint violation of the point.

Lemma 5 (Relative form of Hoffman error bound [37, Proposition 5])**.** *Given $B_h\in\mathbb{R}^{k_H\times M}, b_h\in\mathbb{R}^{k_H}$, $B_g\in\mathbb{R}^{k_G\times M}, b_g\in\mathbb{R}^{k_G}$, define $\Sigma(p,q) \coloneqq \{y\in\mathbb{R}^M \mid B_g y + b_g \le p, B_h y + b_h = q\}$, and $\mathrm{dom}\,\Sigma \coloneqq \{(p,q)\mid \Sigma(p,q)\neq\emptyset\}$. Let $\Omega_R\subseteq\mathbb{R}^M$ be a reference polyhedron (e.g., one defined by the intersection of half-spaces). Then for all $u\in\Omega_R$, and $(p,q)\in\mathrm{dom}\,\Sigma$, there exists a relative Hoffman constant c_{hof} depending only on B_g, B_h, Ω_R such that*

$$\mathrm{dist}(u, \Sigma(p,q)\cap\Omega_R) \le c_{\mathrm{hof}}(B_g, B_h \mid \Omega_R)\left\|\begin{bmatrix}(B_g u + b_g - p)_+\\ B_h u + b_h - q\end{bmatrix}\right\| \quad \text{(D.35)}$$

where $(B_g u + b_g - p)_+ \coloneqq \max\{0, B_g u + b_g - p\}$ which replaces each negative component of $B_g u + b_g - p$ by zero, and $\mathrm{dist}(u,\Omega) \coloneqq \inf_{u'\in\Omega}\|u-u'\|$.

Proof of Lemma 2. We first construct $\phi(p,q) = \overline{c_{\mathrm{hof}}}\|[p^\top, q^\top]^\top\|A^{-1}\mathbf{1}_M$, where $\overline{c_{\mathrm{hof}}}$ is the Hoffman constant upper bound in Lemma 5. Then $\phi(0,0)=0$, and $\phi(p,q)$ is Lipschitz because

$$\begin{aligned}\|\phi(p,q)-\phi(p',q')\| &\le \overline{c_{\mathrm{hof}}} M\|A^{-1}\|\left|\left\|\begin{bmatrix}p\\q\end{bmatrix}\right\| - \left\|\begin{bmatrix}p'\\q'\end{bmatrix}\right|\right| \\ &\le \overline{c_{\mathrm{hof}}} M\|A^{-1}\|\left\|\begin{bmatrix}p-p'\\q-q'\end{bmatrix}\right\|. \end{aligned} \quad \text{(D.36)}$$

Next we prove the PMOL calmness condition holds by contradiction. Suppose for every $\epsilon>0$, there exists $(\hat{\theta},p,q)\in[(\bar{\theta},0,0)+\epsilon\mathcal{B}]/\{(\bar{\theta},0,0)\}$ satisfying (D.34).

Define $\Omega_{F_1} \coloneqq \{F(\theta)\in\Sigma(0,0)\mid\theta\in\mathbb{R}^q\}\neq\emptyset$, there exists $\tilde{\theta}\in\mathbb{R}^q$ such that $F(\tilde{\theta})\in\Omega_{F_1}$ and $\|F(\tilde{\theta})\|<\infty$. We then consider the following two cases:

Case 1: $F(\hat{\theta}) \in \Sigma(0,0)$. In this case, $(\hat{\theta}, p, q) = (\hat{\theta}, 0, 0) \neq (\bar{\theta}, 0, 0)$, thus $\hat{\theta} \neq \bar{\theta}$. Take $\tilde{\theta} = \hat{\theta} \neq \bar{\theta}$.
Case 2: $F(\hat{\theta}) \notin \Sigma(0,0)$. Take $\tilde{\theta}$ such that $F(\tilde{\theta}) \in \Omega_{F_1}$, then $F(\tilde{\theta}) \neq F(\hat{\theta})$.

In both cases, let Ω_R be the convex hull of $\{F(\hat{\theta}), F(\tilde{\theta})\}$, i.e., $\Omega_R = \mathrm{conv}(\{F(\tilde{\theta}), F(\hat{\theta})\})$. Then Ω_R is a line segment (or reduces to a point in *case 1*), thus a polyhedron. Since $\Sigma(0,0)$ is a line, $F(\tilde{\theta}) \in \Sigma(0,0) \cap \Omega_R$, thus $\Sigma(0,0) \cap \Omega_R = \Omega_R = \{F(\tilde{\theta})\}$ in *case 1*, and $\Sigma(0,0) \cap \Omega_R = \{F(\tilde{\theta})\}$ in *case 2*. Therefore, in both cases,

$$\|F(\tilde{\theta}) - F(\hat{\theta})\| = \mathrm{dist}(F(\hat{\theta}), \Sigma(0,0) \cap \Omega_R) \tag{D.37}$$

where $\mathrm{dist}(F, \Omega) \coloneqq \inf_{F' \in \Omega} \|F - F'\|$.

We also have

$$\begin{aligned}\mathrm{dist}(F(\hat{\theta}), \Sigma(0,0) \cap \Omega_R) \overset{(a)}{\leq} & c_{\mathrm{hof}}(\Omega_R) \left\| \begin{bmatrix} (B_g F(\hat{\theta}) + b_g)_+ \\ B_h F(\hat{\theta}) + b_h \end{bmatrix} \right\| \\ \overset{(b)}{\leq} & \overline{c_{\mathrm{hof}}} \left\| \begin{bmatrix} (-p)_+ \\ -q \end{bmatrix} \right\| \leq \overline{c_{\mathrm{hof}}} \left\| \begin{bmatrix} p \\ q \end{bmatrix} \right\|\end{aligned} \tag{D.38}$$

where (a) follows from Lemma 5; (b) follows from (D.34) that $0 \leq (B_g F(\hat{\theta}) + b_g)_+ \leq (-p)_+$, $B_h F(\hat{\theta}) + b_h = -q$, and that $c_{\mathrm{hof}}(\Omega_R) \leq \overline{c_{\mathrm{hof}}}$ for different bounded Ω_R. Multiplying $\|A\| \mathbf{1}_M$ on both sides of the above inequality yields

$$\|A\| \mathrm{dist}(F(\hat{\theta}), \Sigma(0,0) \cap \Omega_R) \mathbf{1}_M \leq A\phi(p,q). \tag{D.39}$$

It can then be derived that

$$\begin{aligned}AF(\tilde{\theta}) - AF(\hat{\theta}) \leq & \|AF(\tilde{\theta}) - AF(\hat{\theta})\| \mathbf{1}_M \leq \|A\| \|F(\tilde{\theta}) - F(\hat{\theta})\| \mathbf{1}_M \\ \leq & \|A\| \mathrm{dist}(F(\hat{\theta}), \Sigma(0,0) \cap \Omega_R) \mathbf{1}_M \leq A\phi(p,q).\end{aligned} \tag{D.40}$$

By rearranging the above inequality and applying (D.34c), we have that

$$AF(\tilde{\theta}) \leq AF(\hat{\theta}) + A\phi(p,q) < AF(\bar{\theta}) \tag{D.41}$$

which contradicts to that $\bar{\theta}$ is a global solution to (PMOL).

Therefore, the PMOL calmness condition in Definition 9 is satisfied. □

E Proof of Theorem 1: convergence of Algorithm 1

Recall that, we let $\lambda = [\lambda_f; \lambda_g; \lambda_h] \in \mathbb{R}^{M+M_g+M_h}$, $A_{ag} = [A; B_g; B_h] \in \mathbb{R}^{(M+M_g+M_h) \times M}$, and use the following concise notation

$$\begin{aligned}& d^*(\theta) = -\nabla F(\theta) A_{ag}^\top \lambda^*(\theta) \\ & \text{s.t. } \lambda^*(\theta) \in \underset{\lambda \in \Omega_\lambda(\theta)}{\arg\min}\, \varphi(\lambda; \theta) \coloneqq \frac{1}{2} \|\nabla F(\theta) A_{ag}^\top \lambda\|^2 - c_g \lambda_g^\top G(\theta) - c_h \lambda_h^\top H(\theta)\end{aligned} \tag{E.1}$$

where $\Omega_\lambda(\theta) = \Omega_{\lambda_f}(\theta) \times \mathbb{R}_+^{M_g} \times \mathbb{R}^{M_h}$, and $\Omega_{\lambda_f}(\theta) = \{\lambda_f \in \mathbb{R}_+^M \mid {\lambda_f}^\top AF(\theta) = \mathbf{1}^\top AF(\theta)\}$.

In the following discussion in this section, we first present the supporting lemmas and their proofs, then provide the proof of Theorem 1.

E.1 Auxiliary lemmas

Lemma 6 is a result from the smoothness of $F(\theta)$, and thus the smoothness of $G(\theta)$ and $H(\theta)$, whose smoothness constants depend on B_g and B_h, respectively.

Lemma 6. *Suppose Assumptions 1, 2 hold. Then for all $\theta, \theta' \in \mathbb{R}^q$, and all $\lambda_f \in \mathbb{R}^M$, we have*

$$\lambda_f^\top AF(\theta_{t+1}) - \lambda_f^\top AF(\theta_t) \leq \alpha_t \lambda_f^\top A \nabla F(\theta_t)^\top d_t + \frac{\ell_{f,1} \|A^\top \lambda_f\|_1}{2} \alpha_t^2 \|d_t\|^2 \tag{E.2}$$

$$G(\theta_{t+1}) - G(\theta_t) \leq \alpha_t \nabla G(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2} \alpha_t^2 \|B_g^\top\|_{\infty,1} \|d_t\|^2 \mathbf{1} \tag{E.3}$$

$$H(\theta_{t+1}) - H(\theta_t) \leq \alpha_t \nabla H(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2} \alpha_t^2 \|B_h^\top\|_{\infty,1} \|d_t\|^2 \mathbf{1}. \tag{E.4}$$

Proof. By Assumption 2, it holds that $\lambda_f^\top AF(\theta)$ is $\|A^\top\lambda_f\|_1\ell_{f,1}$-smooth. By the definition of smoothness, we have

$$\lambda_f^\top AF(\theta_{t+1}) \le \lambda_f^\top AF(\theta_t) + \alpha_t\lambda_f^\top A\nabla F(\theta_t)^\top d_t + \frac{\ell_{f,1}\|A^\top\lambda_f\|_1}{2}\alpha_t^2\|d_t\|^2. \tag{E.5}$$

Let $B_{g,m}$ and $B_{h,m}$ be the m-th row of B_g and B_h, respectively, then by the $\ell_{f,1}$-smoothness of $F(\theta)$, $B_{g,m}F(\theta)$ is $\ell_{f,1}\|B_{g,m}\|_1$-smooth for all $m \in [M_g]$. Also because $\|B_{g,m}\|_1 \le \|B_g^\top\|_{\infty,1}$ where $\|B_g^\top\|_{\infty,1} = \max_{m\in M_g}\|\|B_{g,m}\|_1\|$, $g_m(\theta)$ is $\ell_{f,1}\|B_g^\top\|_{\infty,1}$-smooth for all $m \in [M_g]$. By the definition of smoothness, it holds that

$$G(\theta_{t+1}) - G(\theta_t) \le \alpha_t\nabla G(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_g^\top\|_{\infty,1}\|d_t\|^2\mathbf{1}. \tag{E.6}$$

Following similar arguments as the above for $G(\theta)$, (E.4) can be proved. □

Lemma 7. *For the subprogram (2.3) or equivalently (E.1), it holds that for any $\lambda \in \Omega_\lambda(\theta)$,*

$$\langle\nabla F(\theta)A_{ag}^\top\lambda, \nabla F(\theta)A_{ag}^\top\lambda^*(\theta)\rangle - [0^\top, c_gG(\theta)^\top, c_hH(\theta)^\top](\lambda - \lambda^*(\theta)) \ge \|\nabla F(\theta)A_{ag}^\top\lambda^*(\theta)\|^2. \tag{E.7}$$

Proof of Lemma 7. Since $\varphi(\lambda;\theta)$ is a convex function w.r.t. λ, by the first order optimality condition, it holds that for all $\lambda \in \Omega_\lambda(\theta)$

$$\langle\nabla_\lambda\varphi(\lambda^*(\theta);\theta), \lambda - \lambda^*(\theta)\rangle \ge 0 \tag{E.8}$$

which can be further written as

$$\lambda^\top A_{ag}\nabla F(\theta)^\top\nabla F(\theta)A_{ag}^\top\lambda^*(\theta) - [0^\top, c_gG(\theta)^\top, c_hH(\theta)^\top](\lambda - \lambda^*(\theta)) \ge \|\nabla F(\theta)A_{ag}^\top\lambda^*(\theta)\|^2. \tag{E.9}$$

This completes the proof. □

We next prove Lemma 8, which can be viewed as a descent lemma for $[G(\theta)]_+$ and $|H(\theta)|_{\text{ab}}$ based on the smoothness of $G(\theta)$ and $H(\theta)$, as well as proper hyperparameter choices. This is crucial for proving the convergence result in Theorem 1. One key technical challenge in proving the lemma is that even though $G(\theta)$ and $H(\theta)$ are smooth, $[G(\theta)]_+$ and $|H(\theta)|_{\text{ab}}$ are not. We address this challenge by exploiting the fact that $\nabla G(\theta_t)^\top d^*(\theta_t) \le -c_gG(\theta_t)$ and $\nabla H(\theta_t)^\top d^*(\theta_t) = -c_gH(\theta_t)$, as well as choosing α_t properly depending on c_g and c_h.

Lemma 8. *Let $\epsilon \ge 0$ be a constant. Define $[y]_+ := \max\{y, 0\}$ which replaces each negative component of y by zero, and $|y|_{\text{ab}}$ replaces each component of y by its absolute value. Let $\{\theta_t\}$ be the sequence produced by Algorithm 1 with the update $\theta_{t+1} = \theta_t + \alpha_t d_t$, where d_t satisfies the constraints of the subprogram (2.1) up to an error of ϵ, i.e.,*

$$[\nabla G(\theta_t)^\top d_t + c_gG(\theta_t)]_+ \le \epsilon\mathbf{1}, \tag{E.10}$$

$$|\nabla H(\theta_t)^\top d_t + c_hH(\theta_t)|_{\text{ab}} \le \epsilon\mathbf{1}. \tag{E.11}$$

If $\alpha_t \le \min\{c_g^{-1}, c_h^{-1}\}$, then it holds that

$$[G(\theta_{t+1})]_+ - [G(\theta_t)]_+ \le -\alpha_tc_g[G(\theta_t)]_+ + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_g^\top\|_{\infty,1}\|d_t\|^2\mathbf{1} + \epsilon\mathbf{1} \tag{E.12}$$

$$|H(\theta_{t+1})|_{\text{ab}} - |H(\theta_t)|_{\text{ab}} \le -\alpha_tc_h|H(\theta_t)|_{\text{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_h^\top\|_{\infty,1}\|d_t\|^2\mathbf{1} + \epsilon\mathbf{1}. \tag{E.13}$$

Proof. By the smoothness of $G(\theta)$ in Lemma 6 and $\nabla G(\theta)^\top d + c_gG(\theta) \le [\nabla G(\theta)^\top d + c_gG(\theta)]_+ \le \epsilon\mathbf{1}$, it holds that

$$\begin{aligned} G(\theta_{t+1}) - G(\theta_t) &\le \alpha_t\nabla G(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_g^\top\|_{\infty,1}\|d_t\|^2\mathbf{1} + \epsilon\mathbf{1} \\ &\le -\alpha_tc_gG(\theta_t) + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_g^\top\|_{\infty,1}\|d_t\|^2\mathbf{1} + \epsilon\mathbf{1}. \end{aligned} \tag{E.14}$$

For all $m \in [M_g]$, since $G(\theta_t) \leq [G(\theta_t)]_+$, it holds that

$$g_m(\theta_{t+1}) - [g_m(\theta_t)]_+ \leq g_m(\theta_t) - [g_m(\theta_t)]_+ - \alpha_t c_g g_m(\theta_t) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_g^\top\|_{\infty,1}\|d_t\|^2 + \epsilon \quad \text{(E.15)}$$

$$\leq -[-g_m(\theta_t)]_+ - \alpha_t c_g g_m(\theta_t) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_g^\top\|_{\infty,1}\|d_t\|^2 + \epsilon. \quad \text{(E.16)}$$

It can be further derived that

$$-[-g_m(\theta_t)]_+ - \alpha_t c_g g_m(\theta_t) = \begin{cases} -\alpha_t c_g g_m(\theta_t), g_m(\theta_t) \geq 0 \\ (1-\alpha_t c_g) g_m(\theta_t), g_m(\theta_t) < 0 \end{cases}$$

$$\leq -\alpha_t c_g [g_m(\theta_t)]_+ \quad \text{(E.17)}$$

where the last inequality holds since $1 - \alpha_t c_g \geq 0$. Plugging this inequality back into (E.16), yields that when $g_m(\theta_{t+1}) \geq 0$,

$$[g_m(\theta_{t+1})]_+ - [g_m(\theta_t)]_+ \leq -\alpha_t c_g [g_m(\theta_t)]_+ + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_g^\top\|_{\infty,1}\|d_t\|^2 + \epsilon. \quad \text{(E.18)}$$

When $g_m(\theta_{t+1}) < 0$, we have

$$[g_m(\theta_{t+1})]_+ - [g_m(\theta_t)]_+ \leq -[g_m(\theta_t)]_+ \leq -\alpha_t c_g [g_m(\theta_t)]_+$$

$$\leq -\alpha_t c_g [g_m(\theta_t)]_+ + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_g^\top\|_{\infty,1}\|d_t\|^2 + \epsilon. \quad \text{(E.19)}$$

Combining (E.18) and (E.19) proves (E.12).

By the smoothness of $H(\theta)$ and $|\nabla H(\theta)^\top d + c_h H(\theta)|_{\text{ab}} \leq \epsilon \mathbf{1}$, we have

$$|H(\theta_{t+1})|_{\text{ab}} \leq |H(\theta_t) - \alpha_t c_h H(\theta_t)|_{\text{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 \mathbf{1} + \epsilon \mathbf{1}$$

$$= (1 - \alpha_t c_h)|H(\theta_t)|_{\text{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 \mathbf{1} + \epsilon \mathbf{1} \quad \text{(E.20)}$$

where the last equality holds because $1 - \alpha_t c_h \geq 0$, which proves (E.13). □

E.2 Proof of Theorem 1

Proof of Theorem 1. To consider both objective function minimization and constraint satisfaction, we define a Lyapunov function below with a constant vector $\lambda = (\lambda_f, \lambda_g, \lambda_h) \in \Omega_\lambda(\theta)$, where $\lambda_f = \mathbf{1}$, $\lambda_g \in \mathbb{R}_+^{M_g}$, $\lambda_h \in \mathbb{R}^{M_h}$, and $\lambda_g > \lambda_g^*(\theta_t)$, $\lambda_h > \lambda_h^*(\theta_t)$ for all $t \in [T]$.

$$\mathbb{V}_t := \underbrace{\lambda_f^\top A F(\theta_t)}_{\mathbb{V}_{f,t}} + \underbrace{\lambda_g^\top [G(\theta_t)]_+}_{\mathbb{V}_{g,t}} + \underbrace{\lambda_h^\top |H(\theta_t)|_{\text{ab}}}_{\mathbb{V}_{h,t}}. \quad \text{(E.21)}$$

Note that $\mathbb{V}_t \geq 0$ for all t since $AF(\theta) \geq 0, \lambda_f \geq 0$.

For notation simplicity, we let $d_t^* = d^*(\theta_t)$. From Assumption 2, the smoothness of the objectives, and Lemma 6, based on the update $\theta_{t+1} = \theta_t + \alpha_t d_t$, it holds that

$$\mathbb{V}_{f,t+1} - \mathbb{V}_{f,t} \overset{(a)}{\leq} \alpha_t \lambda_f^\top A \nabla F(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A^\top\|_{\infty,1}\|d_t\|^2 \lambda_f^\top \mathbf{1}$$

$$\overset{(b)}{\leq} \alpha_t \lambda_f^\top A \nabla F(\theta_t)^\top d_t^* + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A^\top\|_{\infty,1}\|d_t^*\|^2 \lambda_f^\top \mathbf{1} + \epsilon \mathbf{1}$$

$$\overset{(c)}{\leq} -\alpha_t \|d_t^*\|^2 + \alpha_t \big(c_g \lambda_g^*(\theta_t)^\top G(\theta_t) + c_h \lambda_h^*(\theta_t)^\top H(\theta_t)\big) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A^\top \mathbf{1}\|_1 \|d_t^*\|^2 + \epsilon \mathbf{1} \quad \text{(E.22)}$$

where (a) follows Lemma 6; (b) follows from that d_t is an ϵ-optimal solution to the subprogram; (c) follows from Lemma 7 by setting $\lambda = [\mathbf{1}^\top, 0^\top, 0^\top]^\top \in \Omega_\lambda(\theta)$.

From Lemma 8, for $\alpha_t \leq \min\{c_g^{-1}, c_h^{-1}\}$, it holds that

$$\mathbb{V}_{g,t+1} - \mathbb{V}_{g,t} \leq -\alpha_t c_g \lambda_g^\top [G(\theta_t)]_+ + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_g^\top\|_{\infty,1}\|d_t\|^2 \lambda_g^\top \mathbf{1} + \epsilon \lambda_g^\top \mathbf{1} \quad \text{(E.23)}$$

$$\mathbb{V}_{h,t+1} - \mathbb{V}_{h,t} \leq -\alpha_t c_h \lambda_h^\top |H(\theta_t)|_{\text{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 \lambda_h^\top \mathbf{1} + \epsilon \lambda_h^\top \mathbf{1}. \tag{E.24}$$

Combining the above inequalities for $\mathbb{V}_{f,t}, \mathbb{V}_{g,t}, \mathbb{V}_{h,t}$, we have

$$\begin{aligned}
\mathbb{V}_{t+1} - \mathbb{V}_t \leq & -\alpha_t \|d_t^*\|^2 + \alpha_t \big(c_g \lambda_g^*(\theta_t)^\top G(\theta_t) + c_h \lambda_h^*(\theta_t)^\top H(\theta_t)\big) \\
& + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \|d_t^*\|^2 - \alpha_t c_g \lambda_g^\top [G(\theta_t)]_+ - \alpha_t c_h \lambda_h^\top |H(\theta_t)|_{\text{ab}} + \epsilon \lambda^\top \mathbf{1} \\
\leq & -\alpha_t \|d_t^*\|^2 - \alpha_t c_g (\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ - \alpha_t c_g \lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+ \\
& - \alpha_t c_h (\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\text{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \|d_t^*\|^2 + \epsilon \lambda^\top \mathbf{1}
\end{aligned} \tag{E.25}$$

where the last inequality holds because $\lambda_g^*(\theta_t)^\top G(\theta_t) = \lambda_g^*(\theta_t)^\top [G(\theta_t)]_+ - \lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+$. Taking telescoping sum of the above inequality from $t = 0, \ldots, T-1$ and rearranging, we have

$$\begin{aligned}
& \sum_{t=0}^{T-1} \alpha_t \Big(1 - \frac{1}{2}\|A_{ag}^\top \lambda\|_1 \ell_{f,1} \alpha_t\Big) \|d_t^*\|^2 + \alpha_t c_g (\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + \alpha_t c_g \lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+ \\
& + \alpha_t c_h (\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\text{ab}} \leq \mathbb{V}_0 - \mathbb{V}_T + T\epsilon \lambda^\top \mathbf{1} \leq \mathbb{V}_0 + T\epsilon \|\lambda\|_1.
\end{aligned} \tag{E.26}$$

Recall that $\alpha_t \leq 1/(\ell_{f,1}\|A_{ag}^\top \lambda\|_1)$. Plugging this into the above inequality yields

$$\begin{aligned}
& \sum_{t=0}^{T-1} \frac{1}{2}\alpha_t \|d_t^*\|^2 + \alpha_t c_g (\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + \alpha_t c_g \lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+ \\
& + \alpha_t c_h (\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\text{ab}} \leq \mathbb{V}_0 + T\epsilon \|\lambda\|_1.
\end{aligned} \tag{E.27}$$

Taking $\alpha_t = \Theta(1)$, then

$$\begin{aligned}
& \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{2}\|d^*(\theta_t)\|^2 + c_g (\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + c_g \lambda_g^*(\theta_t)^\top [-G(\theta_t)]_+ \\
& + c_h (\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\text{ab}} = \mathcal{O}\Big(\frac{1}{T} + \epsilon\Big).
\end{aligned} \tag{E.28}$$

The proof is complete. □

Next we show that the subprogram converges with a projected gradient descent (PGD) algorithm on λ with K iterations.

Lemma 9 (Convergence of the subprogram with projected gradient descent)**.** *At the t-th iteration, given θ_t, let $\{\lambda_{t,k}\}_k$ be the sequence generated by the projected gradient descent algorithm to solve the subprogram $\min_{\lambda \in \Omega_\lambda(\theta_t)} \varphi(\lambda; \theta_t)$, then*

$$\varphi(\lambda_{t,K}; \theta_t) - \min_{\lambda \in \Omega_\lambda(\theta_t)} \varphi(\lambda; \theta_t) \leq \frac{\|\lambda_{t,0} - \lambda^*(\theta_t)\|^2}{2\gamma K}. \tag{E.29}$$

Proof. The result follows from the convergence result of projected gradient descent for convex objective functions. Note that at each iteration t, given θ_t, $\Omega_\lambda(\theta_t)$ is fixed. □

Lemma 10. *Suppose Assumption 3 holds. Due to the $\ell_{\varphi_\lambda,1}$-smoothness and the convexity of the subprogram, it holds for all $\lambda \in \Omega_\lambda(\theta)$ that*

$$\|\nabla_\lambda \varphi(\lambda; \theta) - \nabla_\lambda \varphi(\lambda^*(\theta); \theta)\|^2 \leq 2\ell_{\varphi_\lambda,1}\big(\varphi(\lambda; \theta) - \varphi(\lambda^*(\theta); \theta)\big). \tag{E.30}$$

Proof. Since the objectives $f_m(\theta)$ are Lipschitz continuous for all $m \in [M]$, the subprogram objective $\varphi(\lambda; \theta)$ is $\ell_{\varphi,\lambda,1}$-smooth w.r.t. λ. By Proposition 1 (b) in [44], it holds that

$$\frac{1}{2\ell_{\varphi_\lambda,1}}\|\nabla_\lambda \varphi(\lambda; \theta) - \nabla_\lambda \varphi(\lambda^*(\theta); \theta)\|^2 + \langle \nabla_\lambda \varphi(\lambda^*(\theta); \theta), \lambda - \lambda^*(\theta)\rangle \leq \varphi(\lambda; \theta) - \varphi(\lambda^*(\theta); \theta). \tag{E.31}$$

By the convexity of $\varphi(\lambda;\theta)$ w.r.t. λ, for all $\lambda \in \Omega_\lambda(\theta)$,

$$\langle \nabla_\lambda \varphi(\lambda^*(\theta);\theta), \lambda - \lambda^*(\theta)\rangle \geq 0. \tag{E.32}$$

Combining the above two inequalities proves the result. □

Corollary 7 (Convergence of Algorithm 1 with K-iteration PGD for the subprogram)**.** *Suppose Assumptions 1, 2 hold. Let $\{\theta_t\}$ be the sequence produced by Algorithm 1 with the update $\theta_{t+1} = \theta_t + \alpha_t d_t$, where d_t is the ϵ-optimal solution to the subprogram (2.1) obtained by K-iteration PGD for the subprogram on λ. Define $\lambda := (\lambda_f, \lambda_g, \lambda_h) \in \Omega_\lambda(\theta)$ with $\lambda_g \geq \lambda_g^*(\theta)$, $\lambda_h \geq \lambda_h^*(\theta)$ for all $\theta \in \mathbb{R}^q$. If the step size $\alpha_t \leq 1/(\ell_{f,1}\|A_{ag}^\top \lambda\|_1)$ and $\alpha_t = \Theta(1)$, then*

$$\sum_{t=T}^{T-1} \frac{1}{2}\|d_t^*\|^2 + c_g(\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + c_h(\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\mathrm{ab}} = \mathcal{O}(1). \tag{E.33}$$

Proof. For $t = 0, \ldots, T-1$, we take $K = T^2$, applying Lemma 9, we have

$$\varphi(\lambda_{t,K};\theta_t) - \min_{\lambda \in \Omega_\lambda(\theta_t)} \varphi(\lambda;\theta_t) \leq \frac{\|\lambda_{t,0} - \lambda^*(\theta_t)\|^2}{2\gamma T^2}. \tag{E.34}$$

From Lemma 10, the above inequality implies

$$\|\nabla\varphi(\lambda_t;\theta_t) - \nabla\varphi(\lambda^*(\theta_t);\theta_t)\|^2 \leq 2\ell_{\varphi_\lambda,1}\big(\varphi(\lambda_t;\theta_t) - \varphi(\lambda^*(\theta_t);\theta_t)\big) \leq \frac{\ell_{\varphi_\lambda,1}\|\lambda_{t-1} - \lambda^*(\theta_t)\|^2}{\gamma T^2}. \tag{E.35}$$

Plugging in the gradient $\nabla\varphi(\lambda_t;\theta_t)$, we have

$$\begin{aligned} &\|A\nabla F(\theta_t)^\top (d_t - d_t^*)\|^2 + \|\nabla G(\theta_t)^\top (d_t - d_t^*)\|^2 + \|\nabla G(\theta_t)^\top (d_t - d_t^*)\|^2 \\ \leq &\frac{\ell_{\varphi_\lambda,1}\|\lambda_{t-1} - \lambda^*(\theta_t)\|^2}{\gamma T^2} \leq \frac{4\ell_{\varphi_\lambda,1} c_\lambda^2}{\gamma T^2}. \end{aligned} \tag{E.36}$$

Let $\epsilon = \frac{4\ell_{\varphi_\lambda,1} c_\lambda^2}{\gamma T^2}$, from Theorem 1, it holds that

$$\begin{aligned} \mathbb{V}_{t+1} - \mathbb{V}_t \leq &-\alpha_t\|d_t^*\|^2 + \alpha_t c_g\big(\lambda_g^*(\theta_t) - \lambda_g\big)^\top [G(\theta_t)]_+ + \alpha_t c_h\big(\lambda_h^*(\theta_t) - \lambda_h\big)^\top |H(\theta_t)|_{\mathrm{ab}} + \epsilon^{\frac{1}{2}} \\ &+ \frac{1}{2}\gamma\alpha_t\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2 + \frac{\ell_{f,1}}{2}\alpha_t^2\|A_{ag}^\top\lambda\|_1\|d_t^*\|^2. \end{aligned} \tag{E.37}$$

Taking telescoping sum of the above inequality from $t = 0, \ldots, T-1$, rearranging, and letting $\alpha_t \leq 1/(\|\lambda\|_1 \ell_{f,1}\|A_{ag}^\top\|_{\infty,1})$, we have

$$\sum_{t=T}^{T-1} \frac{1}{2}\alpha_t\|d_t^*\|^2 + \alpha_t c_g(\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + \alpha_t c_h(\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\mathrm{ab}} \leq \mathbb{V}_T + T\epsilon^{\frac{1}{2}}. \tag{E.38}$$

Letting $\alpha_t = \Theta(1), \gamma = \Theta(1)$ yields

$$\sum_{t=T}^{T-1} \frac{1}{2}\|d_t^*\|^2 + c_g(\lambda_g - \lambda_g^*(\theta_t))^\top [G(\theta_t)]_+ + c_h(\lambda_h - \lambda_h^*(\theta_t))^\top |H(\theta_t)|_{\mathrm{ab}} = \mathcal{O}(1). \tag{E.39}$$

The proof is complete. □

F Proof of Theorems 2 and 3: convergence of Algorithm 2

F.1 Auxiliary lemmas

Lemma 11 (Smoothness of φ)**.** *Suppose Assumptions 1 and 3 hold. $\varphi(\lambda;\theta)$ is $\ell_{\varphi_\lambda,1}$-smooth w.r.t. λ, with $\ell_{\varphi_\lambda,1} = M\|A_{ag}\|^2\ell_f^2$.*

Proof. The Hessian of $\varphi(\lambda;\theta)$ w.r.t. λ can be computed by

$$\nabla^2_\lambda \varphi(\lambda;\theta) = A_{ag}\nabla F(\theta)^\top \nabla F(\theta) A_{ag}^\top.$$

By Assumption 3, the Lipschitz continuity of F, it holds that

$$\|\nabla^2_\lambda \varphi(\lambda;\theta)\| \le \|A_{ag}\nabla F(\theta)^\top \nabla F(\theta) A_{ag}^\top\| \le \|\nabla F(\theta) A_{ag}^\top\|^2 \le M\|A_{ag}\|^2 \ell_f^2.$$

The result is proved. □

Lemma 12 ($\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|$ bounded by $\|d_t\|$)**.** *Suppose Assumptions 1 and 3 hold. For $\{\theta_t\}$ produced by Algorithm 2, we have*

$$\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\| \le \|A^\top\|_{\infty,1}\ell_f\|d_t\|. \tag{F.1}$$

Proof. The gradient of $\varphi(\lambda_t;\theta_t)$ w.r.t. λ_f can be computed by

$$\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t) = A\nabla F(\theta_t)^\top \nabla F(\theta_t) A_{ag}^\top \lambda_t = -A\nabla F(\theta_t)^\top d_t. \tag{F.2}$$

By Assumption , it holds that

$$\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\| \le \|A^\top\|_{\infty,1}\ell_f\|d_t\|. \tag{F.3}$$

□

Lemma 13. *Let $\lambda_t = [\lambda_{f,t};\lambda_{h,t}]$. Consider the sequence $\{\lambda_t\}_{t=1}^T$ generated by the update (3.1). Then for all $\lambda \in \Omega_\lambda(\theta_t)$ with $\lambda = (\lambda_f, \lambda_h)$, it holds that*

$$\begin{aligned}
&2\gamma_t\langle \lambda_{f,t} - \lambda_f, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle \le \|\lambda_{f,t} - \lambda_f\|^2 - \|\lambda_{f,t+1} - \lambda_f\|^2 + \gamma_t^2\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2;\\
&2\gamma_t\langle \lambda_{h,t} - \lambda_h, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle = \|\lambda_{h,t} - \lambda_h\|^2 - \|\lambda_{h,t+1} - \lambda_h\|^2 + \gamma_t^2\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2.
\end{aligned} \tag{F.4}$$

Proof. By the update of $\lambda_{f,t}$, and the non-expansiveness of projection, let $\lambda_f = \mathbf{1} \in \Omega_{\lambda_f}(\theta)$ for all $\theta \in \mathbb{R}^q$, we have

$$\begin{aligned}
&\|\lambda_{f,t+1} - \lambda_f\|^2 \le \|\lambda_{f,t} - \gamma_t\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t) - \lambda_f\|^2\\
=&\|\lambda_{f,t} - \lambda_f\|^2 - 2\gamma_t\langle \lambda_{f,t} - \lambda_f, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle + \gamma_t^2\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2.
\end{aligned} \tag{F.5}$$

Rearranging the above inequality gives

$$\begin{aligned}
&2\gamma_t\langle \lambda_{f,t} - \lambda_f, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle\\
\le&\|\lambda_{f,t} - \lambda_f\|^2 - \|\lambda_{f,t+1} - \lambda_f\|^2 + \gamma_t^2\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2
\end{aligned} \tag{F.6}$$

which proves the first inequality.

By the update of $\lambda_{h,t}$, for all constant $\lambda_h \in \mathbb{R}^{M_h}$, we have

$$\begin{aligned}
&\|\lambda_{h,t+1} - \lambda_h\|^2 = \|(\lambda_{h,t} - \gamma_t\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)) - \lambda_h\|^2\\
=&\|\lambda_{h,t} - \lambda_h\|^2 + \gamma_t^2\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2 - 2\gamma_t\langle \lambda_{h,t} - \lambda_h, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle.
\end{aligned} \tag{F.7}$$

Rearranging the above inequality gives

$$\begin{aligned}
&2\gamma_t\langle \lambda_{h,t}, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle - 2\gamma_t\langle \lambda_h, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle\\
=&\|\lambda_{h,t} - \lambda_h\|^2 - \|\lambda_{h,t+1} - \lambda_h\|^2 + \gamma_t^2\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2.
\end{aligned} \tag{F.8}$$

□

Corollary 8. *Let $\lambda_t = [\lambda_{f,t};\lambda_{h,t}]$. Consider the sequence $\{\lambda_t\}_{t=1}^T$ generated by the update (3.1). Then for all $\lambda \in \Omega_\lambda(\theta_t)$ with $\lambda = (\lambda_f, \lambda_h)$, it holds that*

$$2\gamma_t\big(\varphi(\lambda_t;\theta_t) - \varphi(\lambda;\theta_t)\big) \le \|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2 + \gamma_t^2\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2. \tag{F.9}$$

Proof of Corollary 8. The result follows from combining the two inequalities in Lemma 13, and applying the convexity property of φ w.r.t. λ. □

F.2 Analysis with the same merit function: proof of Theorem 2

In this section, we provide analysis with the same merit function as Theorem 1. The proof follows similar ideas of the proofs of Theorem 3 and Theorem 5 in [6].

We first define the following auxiliary functions to assist our analysis. Note that the functions are only used for analysis but not for the algorithm update.

$$\varphi_\rho(\lambda;\theta) \coloneqq \varphi(\lambda;\theta) + \frac{\rho}{2}\|\lambda\|^2, \quad \lambda_\rho^*(\theta) \coloneqq \operatorname*{arg\,min}_{\lambda\in\Omega_\lambda(\theta)} \varphi_\rho(\lambda;\theta). \tag{F.10}$$

We then present the following Lemmas that are useful for the proof of convergence of Algorithm 2.

Lemma 14. *Suppose Assumption 3 holds, and $\lambda^*(\theta)$ and $\lambda_\rho^*(\theta)$ are bounded for $\theta \in \{\theta_t\}_{t=0}^{T-1}$ produced by Algorithm 2, i.e., $\|\lambda^*(\theta)\| \le c_{\overline{\lambda}}$, $\|\lambda_\rho^*(\theta)\| \le c_{\overline{\lambda}}$. Then on the trajectory of Algorithm 2, with $\theta \in \{\theta_t\}_{t=0}^{T-1}$, we have*

$$\varphi(\lambda_\rho^*(\theta);\theta) - \varphi(\lambda^*(\theta);\theta) \le \frac{\rho}{2} c_{\overline{\lambda}}. \tag{F.11}$$

Proof of Lemma 14. The proof follows the proof of [6, Lemma 13]. □

Corollary 9. *Suppose Assumption 3 holds, and $\lambda^*(\theta)$ and $\lambda_\rho^*(\theta)$ are bounded for $\theta \in \{\theta_t\}_{t=0}^{T-1}$ produced by Algorithm 2, i.e., $\|\lambda^*(\theta)\| \le c_{\overline{\lambda}}$, $\|\lambda_\rho^*(\theta)\| \le c_{\overline{\lambda}}$. Then on the trajectory of Algorithm 2, with $\theta \in \{\theta_t\}_{t=0}^{T-1}$, we have*

$$\|\nabla_{\lambda_h}\varphi(\lambda;\theta)\|^2 \le 2\ell_{\varphi_\lambda,1}\big(\varphi(\lambda;\theta) - \varphi(\lambda_\rho^*(\theta);\theta)\big) + \ell_{\varphi_\lambda,1}\rho c_{\overline{\lambda}}. \tag{F.12}$$

Proof of Corollary 9. By applying Lemma 10, and that $\nabla_{\lambda_h}\varphi(\lambda^*(\theta);\theta) = 0$, we have

$$\begin{aligned}\|\nabla_{\lambda_h}\varphi(\lambda;\theta)\|^2 =& \|\nabla_{\lambda_h}\varphi(\lambda;\theta) - \nabla_{\lambda_h}\varphi(\lambda^*(\theta);\theta)\|^2 \le \|\nabla_\lambda\varphi(\lambda;\theta) - \nabla_\lambda\varphi(\lambda^*(\theta);\theta)\|^2 \\ \overset{\text{Lemma 10}}{\le}& 2\ell_{\varphi_\lambda,1}\big(\varphi(\lambda;\theta) - \min_{\lambda\in\Omega_\lambda(\theta)}\varphi(\lambda;\theta)\big).\end{aligned} \tag{F.13}$$

Applying Lemma 14, we can further derive

$$\begin{aligned}\varphi(\lambda;\theta) - \min_{\lambda\in\Omega_\lambda(\theta)}\varphi(\lambda;\theta) =& \varphi(\lambda;\theta) - \varphi(\lambda^*(\theta);\theta) + \varphi(\lambda_\rho^*(\theta);\theta) - \varphi(\lambda_\rho^*(\theta);\theta) \\ \overset{\text{Lemma 14}}{\le}& \varphi(\lambda;\theta) - \varphi(\lambda_\rho^*(\theta);\theta) + \frac{\rho}{2}c_{\overline{\lambda}}.\end{aligned} \tag{F.14}$$

Combining (F.13) and (F.14) yields the result. □

Lemma 15 (Continuity of $\lambda_\rho^*(\theta)$)**.** *For $\lambda_\rho^*(\theta)$ defined in (F.10), and $\Omega_\lambda(\theta) = \Omega_\lambda$, the following holds*

$$\begin{aligned}\|\lambda_\rho^*(\theta) - \lambda_\rho^*(\theta')\| \le& \rho^{-1}\|\nabla_\lambda^2\varphi(\lambda_\rho^*(\theta);\theta) - \nabla_\lambda^2\varphi(\lambda_\rho^*(\theta');\theta')\| \\ \le& 2\rho^{-1}\ell_{f,1}\ell_f\|A_{ag}^\top\|_{\infty,1}^2\|\theta - \theta'\|.\end{aligned} \tag{F.15}$$

Proof of Lemma 15. The proof follows the proof of [6, Lemma 12]. □

Lemma 16. *Suppose Assumptions 1, 2, 3 hold. Let $\{\theta_t\}, \{\lambda_t\}$ be the sequences produced by Algorithm 2 with step sizes $\alpha_t = \alpha > 0$, $\gamma_t = \gamma > 0$. Assume $\|\lambda^*(\theta_t)\|, \|\lambda_\rho^*(\theta_t)\|, \|\lambda_t\| \le c_{\overline{\lambda}}$. Then for any $\rho > 0$, it holds that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\varphi(\lambda_t;\theta_t) - \varphi(\lambda_\rho^*(\theta_t);\theta_t) \le \frac{2c_{\overline{\lambda}}^2}{\gamma T}(1 + 2\rho^{-1}\alpha T\ell_{f,1}\ell_f^2\|A_{ag}^\top\|_{\infty,1}^3) + \frac{\gamma}{2T}\sum_{t=0}^{T-1}\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2. \tag{F.16}$$

Proof of Lemma 16. The proof follows the proof techniques of [6, Lemma 15].

First, applying Corollary 8 and $\gamma_t = \gamma$ yields

$$2\gamma\big(\varphi(\lambda_t;\theta_t) - \varphi(\lambda_\rho^*(\theta_t);\theta_t)\big) \le \|\lambda_t - \lambda_\rho^*(\theta_t)\|^2 - \|\lambda_{t+1} - \lambda_\rho^*(\theta_t)\|^2 + \gamma^2\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2. \tag{F.17}$$

Taking telescoping sum of the above inequality and rearranging, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\varphi(\lambda_t;\theta_t) - \varphi(\lambda_\rho^*(\theta_t);\theta_t) \le& \frac{1}{2\gamma T}\Big(\underbrace{\sum_{t=0}^{T-1}\|\lambda_t - \lambda_\rho^*(\theta_t)\|^2 - \|\lambda_{t+1} - \lambda_\rho^*(\theta_t)\|^2}_{I_1}\Big)\\ &+ \frac{\gamma}{2T}\sum_{t=0}^{T-1}\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2 \end{aligned}\tag{F.18}$$

where I_1 can be further bounded by

$$\begin{aligned} I_1 \le& \|\lambda_0 - \lambda_\rho^*(\theta_0)\|^2 - \|\lambda_T - \lambda_\rho^*(\theta_{T-1})\|^2 + \sum_{t=0}^{T-2}\|2\lambda_{t+1} - \lambda_\rho^*(\theta_{t+1}) - \lambda_\rho^*(\theta_t)\|\|\lambda_\rho^*(\theta_{t+1}) - \lambda_\rho^*(\theta_t)\|\\ \le& 4c_\lambda^2 + 4c_\lambda\sum_{t=0}^{T-2}\|\lambda_\rho^*(\theta_{t+1}) - \lambda_\rho^*(\theta_t)\| \le 4c_\lambda^2 + 8c_\lambda\sum_{t=0}^{T-2}\rho^{-1}\alpha_t\ell_{f,1}\ell_f\|A_{ag}^\top\|_{\infty,1}^2\|d_t\| \end{aligned}$$

where the last inequality follows from Lemma 15 and the update of θ_t.

Finally, taking $\alpha_t = \alpha$, plugging the above bound for I_1 back into (F.18), and bounding $\|d_t\|$ by Assumption 3 and that $\|\lambda_t\| \le c_{\bar\lambda}$ prove the result. □

Proof of Theorem 2. We consider the following Lyapunov function with a constant vector $\lambda = [\lambda_f; \lambda_h] \in \Omega_\lambda(\theta)$, where $\lambda_f = \mathbf{1}$, $\lambda_h \in \mathbb{R}^{M_h}$.

$$\mathbb{V}_t := \underbrace{\lambda_f^\top AF(\theta_t)}_{\mathbb{V}_{f,t}} + \underbrace{\underbrace{\lambda_h^\top H(\theta_t)}_{\mathbb{V}_{h,1,t}} + \underbrace{c_{V_h}\|H(\theta_t)\|_1}_{\mathbb{V}_{h,3,t}}}_{\mathbb{V}_{h,t}}. \tag{F.19}$$

Recall that $\lambda_t = [\lambda_{f,t}; \lambda_{h,t}]$, and the algorithm takes the update $\theta_{t+1} = \theta_t + \alpha_t d_t$ with $d_t = \nabla F(\theta_t)A_{ag}^\top\lambda_t$. From Assumption 2, the smoothness of the objectives, and Lemma 6, the function $\lambda_f^\top AF(\theta)$ is smooth, thus

$$\begin{aligned}\mathbb{V}_{f,t+1} - \mathbb{V}_{f,t} \le& \langle\nabla F(\theta_t)A^\top\lambda_f, \theta_{t+1} - \theta_t\rangle + \frac{\ell_{f,1}}{2}\|A^\top\lambda_f\|_1\|\theta_{t+1} - \theta_t\|^2\\ =& \alpha_t\langle\nabla F(\theta_t)A^\top\lambda_f, d_t\rangle + \frac{\ell_{f,1}}{2}\alpha_t^2\|A^\top\lambda_f\|_1\|d_t\|^2. \end{aligned}\tag{F.20}$$

By Lemma 13, taking $\gamma_t > 0$ and rearranging, we have

$$\begin{aligned}\langle\nabla F(\theta_t)A^\top\lambda_f, d_t\rangle \le& \frac{1}{2\gamma_t}\big(\|\lambda_{f,t} - \lambda_f\|^2 - \|\lambda_{f,t+1} - \lambda_f\|^2\big)\\ &+ \frac{1}{2}\gamma_t\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2 - \langle\lambda_{f,t}, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle. \end{aligned}\tag{F.21}$$

Combining (F.20) and (F.21), we have

$$\begin{aligned}\mathbb{V}_{f,t+1} - \mathbb{V}_{f,t} \le& \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_{f,t} - \lambda_f\|^2 - \|\lambda_{f,t+1} - \lambda_f\|^2\big) + \frac{\ell_{f,1}}{2}\alpha_t^2\|A^\top\lambda_f\|_1\|d_t\|^2\\ &+ \frac{1}{2}\alpha_t\gamma_t\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2 - \alpha_t\langle\lambda_{f,t}, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle. \end{aligned}\tag{F.22}$$

By the smoothness of $\lambda_h^\top H(\theta)$, and $\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t) = -\nabla H(\theta_t)^\top d_t - c_h H(\theta_t)$, it holds that

$$\mathbb{V}_{h,1,t+1} - \mathbb{V}_{h,1,t} \le \alpha_t\lambda_h^\top\nabla H(\theta_t)^\top d_t + \frac{\ell_{f,1}}{2}\alpha_t^2\|B_h^\top\lambda_h\|_1\|d_t\|^2$$

$$= -\alpha_t c_h \lambda_h^\top H(\theta_t) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top \lambda_h\|_1 \|d_t\|^2 - \alpha_t \langle \lambda_h, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle. \tag{F.23}$$

Bounding the last term in the above inequality by Lemma 13, and taking $\gamma_t > 0$, we have

$$\mathbb{V}_{h,1,t+1} - \mathbb{V}_{h,1,t} \le -\alpha_t c_h \lambda_h^\top H(\theta_t) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top \lambda_h\|_1 \|d_t\|^2 + \frac{1}{2}\alpha_t\gamma_t \|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2 - \alpha_t \langle \lambda_{h,t}, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_{h,t} - \lambda_h\|^2 - \|\lambda_{h,t+1} - \lambda_h\|^2\big). \tag{F.24}$$

Adding up (F.22) and (F.24) yields

$$\begin{aligned}
&\mathbb{V}_{f,t+1} - \mathbb{V}_{f,t} + \mathbb{V}_{h,1,t+1} - \mathbb{V}_{h,1,t} \le -\alpha_t \langle \lambda_t, \nabla_\lambda \varphi(\lambda_t;\theta_t)\rangle + \frac{1}{2}\gamma_t\alpha_t \|\nabla_\lambda \varphi(\lambda_t;\theta_t)\|^2 \\
&\quad - \alpha_t c_h \lambda_h^\top H(\theta_t) + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \|d_t\|^2 \\
&\le -\alpha_t \|d_t\|^2 + \alpha_t c_h (\lambda_{h,t} - \lambda_h)^\top H(\theta_t) + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) \\
&\quad + \frac{1}{2}\gamma_t\alpha_t \|\nabla_\lambda \varphi(\lambda_t;\theta_t)\|^2 + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \|d_t\|^2
\end{aligned} \tag{F.25}$$

where the last inequality uses the fact that $\langle \lambda_t, \nabla_\lambda \varphi(\lambda_t;\theta_t)\rangle = \|d_t\|^2 - c_h \lambda_{h,t}^\top H(\theta_t)$.

Using the fact that $\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t) = -\nabla H(\theta_t)^\top d_t - c_h H(\theta_t)$, and with similar arguments as (E.20) in Lemma 8, we can further derive that

$$\begin{aligned}
|H(\theta_{t+1})|_{\mathrm{ab}} &\le |H(\theta_t) - \alpha_t c_h H(\theta_t) - \alpha_t \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)|_{\mathrm{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 \mathbf{1} \\
&\le (1 - \alpha_t c_h)|H(\theta_t)|_{\mathrm{ab}} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 \mathbf{1} + \alpha_t |\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)|_{\mathrm{ab}}.
\end{aligned} \tag{F.26}$$

Therefore,

$$\mathbb{V}_{h,2,t+1} - \mathbb{V}_{h,3,t} \le -\alpha_t c_h c_{V_h} \|H(\theta_t)\|_1 + \frac{\ell_{f,1}}{2} c_{V_h} M_h \alpha_t^2 \|B_h^\top\|_{\infty,1}\|d_t\|^2 + \alpha_t c_{V_h} \|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|_1. \tag{F.27}$$

Combining (F.25) and (F.27), and by choosing step sizes α_t, γ_t, parameter c_{V_h} such that

$$\frac{\ell_{f,1}}{2} c_{V_h} M_h \alpha_t^2 \|B_h^\top\|_{\infty,1} + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \le \frac{1}{2}, \tag{F.28}$$

we have

$$\mathbb{V}_{t+1} - \mathbb{V}_t \le -\frac{1}{2}\alpha_t \|d_t\|^2 - \alpha_t c_h (c_{V_h} - \|\lambda_h - \lambda_{h,t}\|_1)\|H(\theta_t)\|_1 + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) + \frac{1}{2}\gamma_t\alpha_t\ell_\varphi^2 + \alpha_t c_{V_h}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|_1. \tag{F.29}$$

Taking telescoping sum of the above inequality over $t = 0, \ldots, T-1$, and applying that $\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|_1 \le \sqrt{M_h}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|$, we have

$$\sum_{t=0}^{T-1} \mathbb{V}_{t+1} - \mathbb{V}_t \le \sum_{t=0}^{T-1} -\frac{1}{2}\alpha_t\|d_t\|^2 - \alpha_t c_h (c_{V_h} - \|\lambda_h - \lambda_{h,t}\|_1)\|H(\theta_t)\|_1 + \frac{1}{2}\gamma_t\alpha_t\ell_\varphi^2 + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) + \alpha_t c_{V_h}\sqrt{M_h}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\| \tag{F.30}$$

where $\sum_{t=0}^{T-1}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|$ can be further bounded by applying Lemma 16 and Corollary 9 along with Jensen's inequality as follows

$$\Big(\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|\Big)^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2$$

$$\overset{\text{Corollary 9}}{\leq} \frac{1}{T}\sum_{t=0}^{T-1} 2\ell_{\varphi_\lambda,1}\big(\varphi(\lambda_t;\theta_t)-\varphi(\lambda_\rho^*(\theta_t);\theta_t)\big)+\ell_{\varphi_\lambda,1}\rho c_{\overline{\lambda}}$$

$$\overset{\text{Lemma 16}}{\leq} 4\ell_{\varphi_\lambda,1}c_{\overline{\lambda}}^2\frac{1}{\gamma T}(1+2\rho^{-1}\alpha T\ell_{f,1}\ell_f^2\|A_{ag}^\top\|_{\infty,1}^3)+\frac{\gamma}{2T}\sum_{t=0}^{T-1}\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2+\rho\ell_{\varphi_\lambda,1}c_{\overline{\lambda}} \tag{F.31}$$

where $\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2=\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2+\|\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2\lesssim\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2+\|d_t\|^2$. Plugging the above inequality back into (F.30), choosing $\rho=\Theta\Big((\frac{\alpha}{\gamma})^{\frac{1}{2}}\Big)$, and rearranging yield

$$\frac{1}{T}\sum_{t=0}^{T-1}\|d_t\|^2+\|H(\theta_t)\|_1=\mathcal{O}\Big(\frac{1}{\alpha T}+\frac{1}{(\gamma T)^{\frac{1}{2}}}+(\frac{\alpha}{\gamma})^{\frac{1}{4}}+\gamma\Big). \tag{F.32}$$

Choosing $\alpha=\Theta(T^{-\frac{5}{6}})$, $\gamma=\Theta(T^{-\frac{1}{6}})$ proves the result. □

F.3 Sharper analysis with a different merit function: proof of Theorem 3

Lemma 17. *Suppose Assumptions 1, 2, and 3 hold. If $\alpha_t\|H(\theta_t)\|\leq c_{\alpha,h}$, and $\alpha_t\|d_t\|\leq c_d$ for all $t\in[T]$, then for all $\{\theta_t\}_{t=0}^T$ produced by Algorithm 2, it holds that*

$$\|H(\theta_{t+1})\|^2-\|H(\theta_t)\|^2\leq\alpha_t 2H(\theta_t)^\top\nabla H(\theta_t)^\top d_t+\frac{1}{2}\alpha_t^2\ell_{H^2,1}\|d_t\|^2$$
$$\text{with}\ \ \ell_{H^2,1}=2M\ell_f^2+2(\alpha_t^{-1}+\ell_H)c_{d,h}\sqrt{M}\ell_{f,1}.$$

Proof. By the mean-value theorem, for all $t\in[T]$, there exists $\tilde{\theta}$ such that

$$\|H(\theta_{t+1})\|^2-\|H(\theta_t)\|^2\leq\alpha_t 2H(\theta_t)^\top\nabla H(\theta_t)^\top d_t+\frac{1}{2}\alpha_t^2\|\nabla^2H(\tilde{\theta})^\top H(\tilde{\theta})\|\|d_t\|^2. \tag{F.33}$$

The Hessian of $\|H(\tilde{\theta})\|^2$ can be upper bounded by

$$\|\nabla^2H(\tilde{\theta})^\top H(\tilde{\theta})\|\leq 2\|\nabla H(\tilde{\theta})\nabla H(\tilde{\theta})^\top\|+2\|\nabla^2H(\tilde{\theta})\|\|H(\tilde{\theta})\|\leq 2M\ell_f^2+2\|H(\tilde{\theta})\|\sqrt{M}\ell_{f,1}. \tag{F.34}$$

Since $H(\tilde{\theta})$ is ℓ_H-Lipschitz continuous with $\ell_H=\|B_h\|\ell_F$, and $\tilde{\theta}$ lies on the line segment of θ_t and θ_{t+1} with $\|\theta_{t+1}-\theta_t\|=\alpha_t\|d_t\|$, therefore,

$$\|H(\tilde{\theta})\|\leq\|H(\theta_t)\|+\alpha_t\ell_H\|d_t\|\leq\|H(\theta_t)\|+\ell_H c_d. \tag{F.35}$$

Plugging the above inequality into (F.33) yields

$$\|H(\theta_{t+1})\|^2-\|H(\theta_t)\|^2\leq\alpha_t 2H(\theta_t)^\top\nabla H(\theta_t)^\top d_t+\frac{1}{2}\alpha_t^2\|\nabla^2H(\tilde{\theta})^\top H(\tilde{\theta})\|\|d_t\|^2$$
$$\leq\alpha_t 2H(\theta_t)^\top\nabla H(\theta_t)^\top d_t+\frac{1}{2}\alpha_t^2\Big(2M\ell_f^2+2(\alpha_t^{-1}c_{\alpha,h}+\ell_H c_d)\sqrt{M}\ell_{f,1}\Big)\|d_t\|^2.$$

The proof is complete. □

Lemma 18. *Suppose Assumptions 1 and 3 hold. For $\{\theta_t\},\{\lambda_t\}$ produced by Algorithm 2, choose λ_0 to be bounded. Choose α_t such that $\alpha_t\|H(\theta_t)\|\leq c_{\alpha,h}$, and $\alpha_t\|d_t\|\leq c_d$ for all $t\in[T]$, and choose $\gamma_t=\Theta(\frac{\alpha_t}{T})$. Then $\|\lambda_t\|$ is bounded for all $t\in[T]$. Consequently, $\|d_t\|$ is bounded for all $t\in[T]$. If we further have $\|H(\theta_t)\|$ bounded for all $t\in[T]$, then $\|\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\|$ is bounded for all $t\in[T]$.*

Proof. By choosing $\gamma_t=\Theta(\frac{\alpha_t}{T})$, and $\|\lambda_0\|$ to be bounded, then for all $\tau\in[T]$, we have

$$\|\lambda_\tau\|\leq\|\lambda_0\|+\sum_{t=0}^{\tau-1}\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|=\|\lambda_0\|+\sum_{t=0}^{\tau-1}\gamma_t\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|$$

$$\lesssim \|\lambda_0\| + \frac{1}{T}\sum_{t=0}^{\tau-1} \alpha_t (\|\nabla F(\theta_t) A_{ag}^\top\| \|d_t\| + c_h \|H(\theta_t)\|)$$
$$\leq \|\lambda_0\| + \ell_F \|A_{ag}\| c_d + c_h c_{\alpha,h} = \Theta(1). \tag{F.36}$$

This proves that $\|\lambda_t\|$ is bounded for all $t \in [T]$.

Since $d_t = \nabla F(\theta_t) A_a g \lambda_t$, consequently we have

$$\|d_t\| = \|\nabla F(\theta_t) A_a g \lambda_t\| = \Theta(1). \tag{F.37}$$

Furthermore, if $\|H(\theta_t)\| \leq c_{ht}$ for all $t \in [T]$, invoking that $\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t) = -\nabla H(\theta_t)^\top d_t - c_h H(\theta_t)$, we have

$$\begin{aligned}\|\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t)\| =& \|\nabla H(\theta_t)^\top d_t + c_h H(\theta_t)\| \leq \|\nabla H(\theta_t)\| \|d_t\| + c_h \|H(\theta_t)\| \\ \leq& \ell_H \|d_t\| + c_h \|H(\theta_t)\| = \Theta(1).\end{aligned} \tag{F.38}$$

The proof is complete. □

Lemma 19. *Suppose Assumptions 1, 2, and 3 hold. For $\{\theta_t\}, \{\lambda_t\}$ produced by Algorithm 2, suppose $\|\lambda_{h,t}\| \leq c_\lambda < \infty$ is bounded for all $t \in [T]$, and $\frac{\alpha_0}{T\gamma_0} = c_{\alpha,\gamma} < \infty$. Choose α_t such that $\alpha_t \|H(\theta_t)\| \leq c_{\alpha,h}$ for all $t \in [T]$, and suppose $\|H(\theta_t)\|$ bounded for all $t \in [T]$. Define S_T as follows*

$$S_T = \frac{2}{T}\sum_{t=0}^{T-1} \alpha_t \big(c_h \lambda_{h,t}^\top H(\theta_t) - \nabla_{\lambda_h} \varphi(\lambda_t; \theta_t)^\top H(\theta_t)\big) + \frac{\alpha_0}{T\gamma_0}(\|\lambda_{h,T}\|^2 - \|\lambda_{h,0}\|^2). \tag{F.39}$$

Then we have that $\|S_T\| = \Theta(1)$ is bounded.

Proof. By the definition of S_T, Cauchy-Schwartz inequality, we have

$$\|S_T\| \leq \frac{2}{T}\sum_{t=0}^{T-1} \alpha_t c_h \|\lambda_{h,t}\| \|H(\theta_t)\| + \alpha_t \|\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t)\| \|H(\theta_t)\| + \frac{\alpha_0}{T\gamma_0} \|\lambda_{h,T}\|^2. \tag{F.40}$$

By Lemma 18 and that $\|H(\theta_t)\|$ is bounded, $\|\lambda_t\|$ and $\|\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t)\|$ are both bounded for all $t \in [T]$, therefore, $\|S_T\| = \Theta(1)$ is also bounded. □

Proof of Theorem 3. We consider the following Lyapunov function with a constant vector $\lambda = [\lambda_f; \lambda_h] \in \Omega_\lambda(\theta)$, where $\lambda_f = \mathbf{1}$, $\lambda_h \in \mathbb{R}^{M_h}$.

$$\mathbb{V}_t := \underbrace{\lambda_f^\top A F(\theta_t)}_{\mathbb{V}_{f,t}} + \underbrace{\underbrace{\lambda_h^\top H(\theta_t)}_{\mathbb{V}_{h,1,t}} + \underbrace{\frac{1}{2}\|H(\theta_t)\|^2}_{\mathbb{V}_{h,3,t}}}_{\mathbb{V}_{h,t}}. \tag{F.41}$$

Following the same arguments from (F.20)-(F.25), we have

$$\begin{aligned}\mathbb{V}_{f,t+1} - \mathbb{V}_{f,t} + \mathbb{V}_{h,1,t+1} - \mathbb{V}_{h,1,t} \leq& -\alpha_t \|d_t\|^2 + \alpha_t c_h (\lambda_{h,t} - \lambda_h)^\top H(\theta_t) + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) \\ &+ \frac{1}{2}\gamma_t \alpha_t \|\nabla_\lambda \varphi(\lambda_t; \theta_t)\|^2 + \frac{\ell_{f,1}}{2}\alpha_t^2 \|A_{ag}^\top \lambda\|_1 \|d_t\|^2.\end{aligned} \tag{F.42}$$

Next we proceed to bound $\mathbb{V}_{h,3,t+1} - \mathbb{V}_{h,3,t}$. By Lemma 17, it holds that

$$\mathbb{V}_{h,3,t+1} - \mathbb{V}_{h,3,t} \leq \alpha_t H(\theta_t)^\top \nabla H(\theta_t)^\top d_t + \frac{1}{4}\alpha_t^2 \ell_{H^2,1} \|d_t\|^2 \tag{F.43}$$

where $\ell_{H^2,1} = 2M\ell_f^2 + 2(\alpha_t^{-1} c_{\alpha,h} + \ell_H c_d)\sqrt{M}\ell_{f,1}$. Because $\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t) = -\nabla H(\theta_t)^\top d_t - c_h H(\theta_t)$, the term $H(\theta_t)^\top \nabla H(\theta_t)^\top d_t$ can be further written as

$$H(\theta_t)^\top \nabla H(\theta_t)^\top d_t = -H(\theta_t)^\top \big(\nabla_{\lambda_h} \varphi(\lambda_t; \theta_t) + c_h H(\theta_t)\big)$$

$$= - c_h\|H(\theta_t)\|^2 - H(\theta_t)^\top \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t). \tag{F.44}$$

Plugging (F.44) into (F.43) yields

$$\frac{1}{2}\|H(\theta_{t+1})\|^2 - \frac{1}{2}\|H(\theta_t)\|^2 \le -\alpha_t c_h\|H(\theta_t)\|^2 - \alpha_t H(\theta_t)^\top \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t) + \frac{1}{4}\alpha_t^2 \ell_{H^2,1}\|d_t\|^2. \tag{F.45}$$

Letting $\ell_{F_{ag},1} = \ell_{f,1}\|A_{ag}^\top \lambda\|_1$, and adding up (F.42) and (F.45), we have

$$\begin{aligned}\mathbb{V}_{t+1} - \mathbb{V}_t \le & -\alpha_t\|d_t\|^2 - \alpha_t c_h\|H(\theta_t)\|^2 + \alpha_t\Big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\Big)^\top H(\theta_t) \\ & + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) + \frac{1}{2}\gamma_t\alpha_t\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2 + \frac{1}{4}\alpha_t^2(2\ell_{F_{ag},1} + \ell_{H^2,1})\|d_t\|^2\end{aligned} \tag{F.46}$$

where $\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2$ can be further bounded via triangle inequality and Cauchy-Schwarz inequality, given by

$$\begin{aligned}\|\nabla_\lambda\varphi(\lambda_t;\theta_t)\|^2 =& \|A_{ag}\nabla F(\theta_t)^\top d_t - c_h H(\theta_t)\|^2 \\ \le& 2\|A_{ag}\nabla F(\theta_t)^\top d_t\|^2 + 2\|c_h H(\theta_t)\|^2 \\ \le& 2\|A_{ag}\|^2 M\ell_f^2\|d_t\|^2 + 2c_h^2\|H(\theta_t)\|^2.\end{aligned} \tag{F.47}$$

Plugging (F.47) back into (F.46) yields

$$\begin{aligned}\mathbb{V}_{t+1} - \mathbb{V}_t \le & -\alpha_t\|d_t\|^2 - \alpha_t c_h\|H(\theta_t)\|^2 + \alpha_t\Big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\Big)^\top H(\theta_t) \\ & + \frac{\alpha_t}{2\gamma_t}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big) + \underbrace{\frac{1}{4}\alpha_t^2(2\ell_{F_{ag},1} + \ell_{H^2,1})\|d_t\|^2}_{J_1} \\ & + \underbrace{\gamma_t\alpha_t\|A_{ag}\|^2 M\ell_f^2\|d_t\|^2}_{J_2} + \underbrace{\gamma_t\alpha_t c_h^2\|H(\theta_t)\|^2}_{J_3}\end{aligned} \tag{F.48}$$

where by choosing the step sizes $\alpha_t \le \frac{1}{2\ell_{F_{ag},1} + \ell_{H^2,1}}$, $\gamma_t \le \min\left\{\frac{1}{4\|A_{ag}\|^2 M\ell_f^2}, \frac{1}{2c_h}\right\}$, it holds that

$$J_1 \le \frac{1}{4}\alpha_t\|d_t\|^2, \quad J_2 \le \frac{1}{4}\alpha_t\|d_t\|^2, \quad J_3 \le \frac{1}{2}\alpha_t c_h\|H(\theta_t)\|^2. \tag{F.49}$$

Plugging (F.49) into (F.48), choosing $\frac{\alpha_t}{\gamma_t} = \frac{\alpha_0}{\gamma_0}$ for all $t \in [T]$, and rearranging, we have

$$\begin{aligned}\mathbb{V}_{t+1} - \mathbb{V}_t \le & -\frac{1}{2}\alpha_t\|d_t\|^2 - \frac{1}{2}\alpha_t c_h\|H(\theta_t)\|^2 + \alpha_t\Big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\Big)^\top H(\theta_t) \\ & + \frac{\alpha_0}{2\gamma_0}\big(\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big).\end{aligned} \tag{F.50}$$

Taking telescoping sum of the above inequality over $t = 0, \ldots, T-1$ yields

$$\begin{aligned}& \sum_{t=0}^{T-1}\alpha_t\Big(\|d_t\|^2 + c_h\|H(\theta_t)\|^2\Big) - \sum_{t=0}^{T-1} 2\alpha_t\Big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\Big)^\top H(\theta_t) \\ \le & 2(\mathbb{V}_0 - \mathbb{V}_T) + \frac{\alpha_0}{\gamma_0}(\|\lambda_0 - \lambda\|^2 - \|\lambda_T - \lambda\|^2) \\ \le & 2\mathbb{V}_{f,0} + 2\lambda_h^\top\big(H(\theta_0) - H(\theta_T)\big) + \|H(\theta_0)\|^2 + \frac{\alpha_0}{\gamma_0}(\|\lambda_0 - \lambda\|^2 - \|\lambda_T - \lambda\|^2)\end{aligned} \tag{F.51}$$

where the last inequality uses the definition of $\mathbb{V}_t$ and that $\mathbb{V}_{f,t} \ge 0$.

Choosing $\lambda_{f,0} = \lambda_f$, and rearranging the above inequality, we have

$$\sum_{t=0}^{T-1}\alpha_t\Big(\|d_t\|^2 + c_h\|H(\theta_t)\|^2\Big) \le 2\mathbb{V}_{f,0} + \|H(\theta_0)\|^2 + \frac{\alpha_0}{\gamma_0}(\|\lambda_{h,0} - \lambda_h\|^2 - \|\lambda_{h,T} - \lambda_h\|^2)$$

$$+\sum_{t=0}^{T-1} 2\alpha_t\big(c_h\lambda_{h,t}-\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\big)^\top H(\theta_t)+2\lambda_h^\top\Big(H(\theta_0)-H(\theta_T)-c_h\sum_{t=0}^{T-1}\alpha_t H(\theta_t)\Big). \tag{F.52}$$

Note that (F.52) holds for all $\lambda_h\in\mathbb{R}^{M_h}$. And λ_h here serves as an auxiliary multiplier to prove the convergence in Theorem 3. We then discuss how to choose a bounded λ_h such that the desired convergence result holds. A simple choice is that the additional terms related to λ_h and $\lambda_{h,t}$ on the right hand side of (F.52) amount to a value greater than zero, i.e.,

$$\begin{aligned}
&2\lambda_h^\top\Big(-c_h\sum_{t=0}^{T-1}\alpha_t H(\theta_t)\Big)-\frac{\alpha_0}{\gamma_0}(\|\lambda_{h,0}-\lambda_h\|^2-\|\lambda_{h,T}-\lambda_h\|^2)\\
=&2\lambda_h^\top\Big(-c_h\sum_{t=0}^{T-1}\alpha_t H(\theta_t)\Big)-\frac{\alpha_0}{\gamma_0}(\|\lambda_{h,0}\|^2-\|\lambda_{h,T}\|^2-2\langle\lambda_h,\lambda_{h,0}-\lambda_{h,T}\rangle)\\
\geq&-2\sum_{t=0}^{T-1}\alpha_t\big(c_h\lambda_{h,t}-\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\big)^\top H(\theta_t),
\end{aligned}\tag{F.53}$$

which is a linear inequality w.r.t. λ_h.

Next we discuss when (F.53) has at least one bounded solution w.r.t. λ_h. Rearranging (F.53) yields

$$\begin{aligned}
&2\lambda_h^\top\Big(\frac{\alpha_0}{T\gamma_0}(\lambda_{h,0}-\lambda_{h,T})-c_h\frac{1}{T}\sum_{t=0}^{T-1}\alpha_t H(\theta_t)\Big)\\
\geq&-\frac{2}{T}\sum_{t=0}^{T-1}\alpha_t\big(c_h\lambda_{h,t}-\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\big)^\top H(\theta_t)+\frac{\alpha_0}{T\gamma_0}(\|\lambda_{h,T}\|^2-\|\lambda_{h,0}\|^2).
\end{aligned}\tag{F.54}$$

For notation simplicity, define the following

$$S_1=2c_h\frac{1}{T}\sum_{t=0}^{T-1}\alpha_t H(\theta_t)-\frac{2\alpha_0}{T\gamma_0}(\lambda_{h,0}-\lambda_{h,T}), \tag{F.55}$$

$$S_2=\frac{2}{T}\sum_{t=0}^{T-1}\alpha_t\big(c_h\lambda_{h,t}^\top H(\theta_t)-\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)^\top H(\theta_t)\big)+\frac{\alpha_0}{T\gamma_0}(\|\lambda_{h,T}\|^2-\|\lambda_{h,0}\|^2). \tag{F.56}$$

By Lemma 19, $\|S_2\|$ is bounded. In general, $S_1\neq 0$ and with proper $\lambda_{h,0}$, $\|S_1\|=\Omega(1)$. Then there exists bounded $\lambda_h\in\mathbb{R}^{M_h}$ with $\|\lambda_h\|=\mathcal{O}(1)$ such that (F.54) holds, thus the last three terms in (F.52) amount to a value no greater than zero, which yields

$$\sum_{t=0}^{T-1}\alpha_t\Big(\|d_t\|^2+c_h\|H(\theta_t)\|^2\Big)\leq 2\mathbb{V}_{f,0}+\|H(\theta_0)\|^2. \tag{F.57}$$

Since $\|H(\theta_t)\|$ and $\|d_t\|$ are bounded for all $t\in[T]$, we can choose $\alpha_t=\Theta(1)$, $\gamma_t=\Theta(T^{-1})$, and it holds that

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\|d_t\|^2+c_h\|H(\theta_t)\|^2\Big)=\mathcal{O}\Big(\frac{1}{T}\Big). \tag{F.58}$$

We then summarize the best possible choices for α_t,γ_t. Choosing θ_0 such that $2\mathbb{V}_{f,0}+\|H(\theta_0)\|^2=c_0<\infty$. Recall that we require $\alpha_t\leq\frac{1}{2\ell_{F_{ag},1}+\ell_{H^2,1}}$. Rearranging this inequality with $\ell_{H^2,1}=2M\ell_f^2+2(\alpha_t^{-1}c_{\alpha,h}+\ell_H c_d)\sqrt{M}\ell_{f,1}$ yields

$$\alpha_t(2\ell_{F_{ag},1}+\ell_{H^2,1})=2\alpha_t\ell_{f,1}\|A_{ag}^\top\lambda\|_1+2\alpha_t M\ell_f^2+2(c_{\alpha,h}+\alpha_t\ell_H c_d)\sqrt{M}\ell_{f,1}\leq 1. \tag{F.59}$$

Recall that we choose α_t such that $\alpha_t\|d_t\| \le c_d$ and $\alpha_t\|H(\theta_t)\| \le c_{\alpha,h}$. We can choose the following specific constants to ensure the above inequality holds

$$c_d = c_{\alpha,h} = \frac{1}{4\sqrt{M}\ell_{f,1}}, \quad c_h = c_{\alpha,h}^{-1} c_0,$$

which implies that

$$\alpha_t \le \frac{1}{4(\ell_{f,1}\|A_{ag}^\top \lambda\|_1 + M\ell_f^2) + \ell_H}. \tag{F.60}$$

Note that $\lambda = [\lambda_f; \lambda_h]$, with $\lambda_f = \mathbf{1}$, and λ_h satisfying (F.53). If $\|\lambda_h\|_1$ is bounded by a constant, then $\|A_{ag}^\top \lambda\|_1 \le \|A_{ag}^\top\|_1 \|\lambda\|_1$ is bounded, and we can choose $\alpha_t = \Omega(1)$, i.e., α_t is lower bounded by a constant.

To summarize, the following inequalities should be satisfied by the step sizes

$$\alpha_t = \min\Big\{c_0, \frac{1}{4\sqrt{M}\ell_{f,1}\max\{1, \|H(\theta_t)\|, \|d_t\|\}}, \frac{1}{4(\ell_{f,1}\|A_{ag}^\top\lambda\|_1 + M\ell_f^2) + \ell_H}\Big\}, \tag{F.61}$$

$$\gamma_t \le \min\Big\{\frac{\alpha_0}{T}, \frac{1}{4\|A_{ag}\|^2 M\ell_f^2}, \frac{1}{2c_h}\Big\}, \text{ and } \gamma_t = \frac{\gamma_0}{\alpha_0}\alpha_t, \tag{F.62}$$

where $\ell_{F_{ag},1} = \ell_{f,1}\|A_{ag}^\top\lambda\|_1$. The proof is complete. □

G Stochastic Algorithms

In this section, we discuss the stochastic algorithms and their convergence guarantees.

G.1 Algorithm summary

The stochastic algorithm is summarized in Algorithm 3. Note that, instead of computing $\nabla F_{\xi_{t,1}}(\theta_t), \nabla F_{\xi_{t,2}}(\theta_t)$, which requires $2M$ gradient computation at each iteration, we compute $\nabla F_{\xi_{t,1}}(\theta_t), \nabla F_{\xi_{t,2}}(\theta_t)A_{ag}^\top\lambda_t$, which requires $M+1$ gradient computation per iteration. This saves nearly half of the per-iteration complexity compared to the most relevant existing stochastic algorithm for multi-objective optimization [6].

Algorithm 3 Preference-guided multi-objective algorithm with approximate update

1: Initialize $t = 0$, θ_0, step sizes α_t, γ_t, define A
2: **for** $t = 0, \dots, T-1$ **do**
3: Compute stochastic gradient $\nabla F_{\xi_{t,1}}(\theta_t), \nabla F_{\xi_{t,2}}(\theta_t)A_{ag}^\top\lambda_t$;
4: Compute an update direction $d_t = \nabla F_{\xi_{t,1}}(\theta_t)A_{ag}^\top\lambda_t$;
5: Choose the step size α_t by a predefined schedule;
6: Update θ_t by $\theta_{t+1} = \theta_t + \alpha_t d_t$;
7: Update λ_t by (3.4);
8: **end for**

G.2 Proof of Theorem 4: convergence of Algorithm 3

We first introduce the supporting lemmas, and then present the main proofs. Denote $\mathcal{F}_t$ as the σ-algebra generated by $\nabla F_{\xi_0}(\theta_0), \nabla F_{\xi_1}(\theta_1), \dots, \nabla F_{\xi_t}(\theta_t)$, where $\xi_t = \{\xi_{t,1}, \xi_{t,2}\}$. For simplicity, we let $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathcal{F}_t]$ in the proofs.

Lemma 20. *Let $\lambda_t = [\lambda_{f,t}; \lambda_{h,t}]$. Consider the stochastic sequence $\{\lambda_t\}_{t=0}^T$ produced by Algorithm 3. Then for all $\lambda \in \Omega_\lambda(\theta_t)$, it holds that*

$$\begin{aligned}
&2\gamma_t\mathbb{E}_t[\langle\lambda_{f,t} - \lambda_f, \nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle] \le \mathbb{E}_t[\|\lambda_{f,t} - \lambda_f\|^2 - \|\lambda_{f,t+1} - \lambda_f\|^2 + \gamma_t^2\|\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2];\\
&2\gamma_t\mathbb{E}_t[\langle\lambda_{h,t} - \lambda_h, \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle] \le \mathbb{E}_t[\|\lambda_{h,t} - \lambda_h\|^2 - \|\lambda_{h,t+1} - \lambda_h\|^2 + \gamma_t^2\|\tilde{\nabla}_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2].
\end{aligned} \tag{G.1}$$

Proof. By the update of λ, it holds that

$$\begin{aligned}&\|\lambda_{f,t+1}-\lambda_f\|^2 \le \|\lambda_{f,t}-\gamma_t\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)-\lambda_f\|^2\\ =&\|\lambda_{f,t}-\lambda_f\|^2-2\gamma_t\langle\lambda_{f,t}-\lambda_f,\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle+\gamma_t^2\|\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2.\end{aligned}\tag{G.2}$$

Taking expectation over the stochastic samples and rearranging the above inequality, we have

$$\begin{aligned}&2\gamma_t\mathbb{E}_t[\langle\lambda_{f,t}-\lambda_f,\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle]=2\gamma_t\mathbb{E}_t[\langle\lambda_{f,t}-\lambda_f,\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle]\\ \le&\mathbb{E}_t[\|\lambda_{f,t}-\lambda_f\|^2-\|\lambda_{f,t+1}-\lambda_f\|^2+\gamma_t^2\|\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2].\end{aligned}\tag{G.3}$$

Following similar arguments, it holds that

$$\begin{aligned}&2\gamma_t\mathbb{E}_t[\langle\lambda_{h,t}-\lambda_h,\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle]\\ \le&\mathbb{E}_t[\|\lambda_{h,t}-\lambda_h\|^2-\|\lambda_{h,t+1}-\lambda_h\|^2+\gamma_t^2\|\tilde{\nabla}_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2].\end{aligned}\tag{G.4}$$

The proof is complete. □

Proof of Theorem 4. Define the following Lyapunov functions

$$\mathbb{V}_t:=\underbrace{\lambda_f^\top AF(\theta_t)}_{\mathbb{V}_{f,t}}+\underbrace{\lambda_h^\top H(\theta_t)+\frac{1}{2}\|H(\theta_t)\|^2}_{\mathbb{V}_{h,t}}.\tag{G.5}$$

Let $\lambda_t=[\lambda_{f,t};\lambda_{h,t}]$. The algorithm takes the update $\theta_{t+1}=\theta_t+\alpha_t d_t$ with $d_t=\nabla F(\theta_t)A_{ag}^\top\lambda_t$. From Assumption 2, the smoothness of the objectives, and Lemma 6, the function $\lambda_f^\top AF(\theta)$ is $\ell_{AF,1}$-smooth with $\ell_{AF,1}=\ell_{f,1}\|A^\top\lambda_f\|_1$, thus

$$\begin{aligned}\mathbb{E}_t[\mathbb{V}_{f,t+1}-\mathbb{V}_{f,t}]\le&\mathbb{E}_t[\langle\nabla F(\theta_t)A^\top\lambda_f,\theta_{t+1}-\theta_t\rangle+\frac{\ell_{AF,1}}{2}\|\theta_{t+1}-\theta_t\|^2]\\ =&\alpha_t\mathbb{E}_t[\langle\nabla F(\theta_t)A^\top\lambda_f,d_t\rangle]+\frac{\ell_{AF,1}}{2}\alpha_t^2\mathbb{E}_t[\|d_t\|^2].\end{aligned}\tag{G.6}$$

By Lemma 20, taking $\gamma_t=\gamma$, and rearranging, we have

$$\begin{aligned}\mathbb{E}_t[\langle\nabla F(\theta_t)A^\top\lambda_f,d_t\rangle]\le&\frac{1}{2\gamma}\left(\mathbb{E}_t[\|\lambda_{f,t}-\lambda_f\|^2]-\mathbb{E}_t[\|\lambda_{f,t+1}-\lambda_f\|^2]\right)\\ &+\frac{1}{2}\gamma\mathbb{E}_t[\|\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2]-\mathbb{E}_t[\langle\lambda_{f,t},\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle].\end{aligned}\tag{G.7}$$

Combining (G.6) and (G.7) and taking total expectation, we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{f,t+1}]-\mathbb{E}[\mathbb{V}_{f,t}]\le&\frac{\alpha_t}{2\gamma}\mathbb{E}\left[\|\lambda_{f,t}-\lambda_f\|^2-\|\lambda_{f,t+1}-\lambda_f\|^2\right]+\frac{\ell_{AF,1}}{2}\alpha_t^2\mathbb{E}[\|d_t\|^2]\\ &+\frac{1}{2}\alpha_t\gamma\mathbb{E}[\|\tilde{\nabla}_{\lambda_f}\varphi(\lambda_t;\theta_t)\|^2]-\alpha_t\mathbb{E}[\langle\lambda_{f,t},\nabla_{\lambda_f}\varphi(\lambda_t;\theta_t)\rangle].\end{aligned}\tag{G.8}$$

Following similar arguments from (G.6)-(G.8),

$$\begin{aligned}&\mathbb{E}[\lambda_h^\top H(\theta_{t+1})]-\mathbb{E}[\lambda_h^\top H(\theta_t)]\le\alpha_t\mathbb{E}[\lambda_h^\top\nabla H(\theta_t)^\top d_t]+\frac{\ell_{H,1}}{2}\alpha_t^2\mathbb{E}[\|d_t\|^2]\\ \le&-\alpha_t c_h\mathbb{E}[\lambda_h^\top H(\theta_t)]-\alpha_t\mathbb{E}[\langle\lambda_h,\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle]+\frac{\ell_{H,1}}{2}\alpha_t^2\mathbb{E}[\|d_t\|^2]\\ \le&-\alpha_t c_h\mathbb{E}[\lambda_h^\top H(\theta_t)]+\frac{\ell_{H,1}}{2}\alpha_t^2\mathbb{E}[\|d_t\|^2]+\frac{1}{2}\alpha_t\gamma\mathbb{E}[\|\tilde{\nabla}_{\lambda_h}\varphi(\lambda_t;\theta_t)\|^2]\\ &-\alpha_t\mathbb{E}[\langle\lambda_{h,t},\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\rangle]+\frac{\alpha_t}{2\gamma}\mathbb{E}\left[\|\lambda_{h,t}-\lambda_h\|^2-\|\lambda_{h,t+1}-\lambda_h\|^2\right].\end{aligned}\tag{G.9}$$

Adding up (G.8) and (G.9) gives

$$\mathbb{E}[\mathbb{V}_{f,t+1}]-\mathbb{E}[\mathbb{V}_{f,t}]+\mathbb{E}[\lambda_h^\top H(\theta_{t+1})]-\mathbb{E}[\lambda_h^\top H(\theta_t)]$$

$$
\begin{aligned}
&\leq -\alpha_t \mathbb{E}[\langle \lambda_t, \nabla_\lambda \varphi(\lambda_t;\theta_t)\rangle] - \alpha_t c_h \mathbb{E}[\lambda_h^\top H(\theta_t)] + \frac{1}{2}\alpha_t \gamma \mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] \\
&\quad + \frac{\ell_{F_{ag},1}}{2}\alpha_t^2 \mathbb{E}[\|d_t\|^2] + \frac{\alpha_t}{2\gamma}\mathbb{E}\big[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big] \\
&= -\alpha_t \mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + \alpha_t c_h \mathbb{E}[\lambda_{h,t}^\top H(\theta_t)] - \alpha_t c_h \mathbb{E}[\lambda_h^\top H(\theta_t)] + \frac{1}{2}\alpha_t \gamma \mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] \\
&\quad + \frac{\ell_{F_{ag},1}}{2}\alpha_t^2 \mathbb{E}[\|d_t\|^2] + \frac{\alpha_t}{2\gamma}\mathbb{E}\big[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big]. \qquad \text{(G.10)}
\end{aligned}
$$

By Lemma 17, and that $\mathbb{E}[\nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)] = \mathbb{E}[-\nabla H(\theta_t)^\top d_t - c_h H(\theta_t)]$, we have

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}[\|H(\theta_{t+1})\|^2] - \frac{1}{2}\mathbb{E}[\|H(\theta_t)\|^2] \leq \alpha_t \mathbb{E}[H(\theta_t)^\top \nabla H(\theta_t)^\top d_t] + \frac{1}{4}\alpha_t^2 \ell_{H^2,1}\mathbb{E}[\|d_t\|^2] \\
&\leq -\alpha_t c_h \mathbb{E}[\|H(\theta_t)\|^2] - \alpha_t \mathbb{E}[H(\theta_t)^\top \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)] + \frac{1}{4}\alpha_t^2 \ell_{H^2,1}\mathbb{E}[\|d_t\|^2]. \qquad \text{(G.11)}
\end{aligned}
$$

Adding up (G.10) and (G.11) yields

$$
\begin{aligned}
\mathbb{E}[\mathbb{V}_{t+1}] - \mathbb{E}[\mathbb{V}_t] &\leq -\alpha_t \mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + \alpha_t c_h \mathbb{E}[\lambda_{h,t}^\top H(\theta_t)] - \alpha_t c_h \mathbb{E}[\lambda_h^\top H(\theta_t)] \\
&\quad + \frac{1}{2}\alpha_t \gamma \mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] + \frac{2\ell_{F_{ag},1} + \ell_{H^2,1}}{4}\alpha_t^2 \mathbb{E}[\|d_t\|^2] \\
&\quad - \alpha_t c_h \mathbb{E}[\|H(\theta_t)\|^2] - \alpha_t \mathbb{E}[H(\theta_t)^\top \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)] + \frac{\alpha_t}{2\gamma}\mathbb{E}\big[\|\lambda_t - \lambda\|^2 - \|\lambda_{t+1} - \lambda\|^2\big] \qquad \text{(G.12)}
\end{aligned}
$$

Taking telescoping sum of the above inequality, choosing $\alpha_t = \alpha$ and rearranging, we have

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \alpha\Big(\mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + c_h \mathbb{E}[\|H(\theta_t)\|^2]\Big) \\
\leq &\mathbb{E}[\mathbb{V}_0 - \mathbb{V}_T] + \frac{\alpha}{2\gamma}\mathbb{E}[\|\lambda_0 - \lambda\|^2 - \|\lambda_T - \lambda\|^2] - \alpha \sum_{t=0}^{T-1} \mathbb{E}\Big[\big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\big)^\top H(\theta_t)\Big] \\
&+ \frac{1}{2}\alpha\gamma \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] + \frac{2\ell_{F_{ag},1} + \ell_{H^2,1}}{4}\alpha^2 \sum_{t=0}^{T-1} \mathbb{E}[\|d_t\|^2]. \qquad \text{(G.13)}
\end{aligned}
$$

We choose $\lambda_{f,0} = \lambda_f$, and λ_h satisfying the following equation

$$
\begin{aligned}
&\frac{\alpha}{2\gamma}\mathbb{E}[\|\lambda_{h,0} - \lambda_h\|^2 - \|\lambda_{h,T} - \lambda_h\|^2] - \alpha \sum_{t=0}^{T-1} \mathbb{E}\Big[\big(c_h(\lambda_{h,t} - \lambda_h) - \nabla_{\lambda_h}\varphi(\lambda_t;\theta_t)\big)^\top H(\theta_t)\Big] \\
&+ \lambda_h^\top \mathbb{E}[H(\theta_0) - H(\theta_T)] = 0. \qquad \text{(G.14)}
\end{aligned}
$$

Plugging the above equation into (G.13), it holds that

$$
\begin{aligned}
&\sum_{t=0}^{T-1} \alpha\Big(\mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + c_h \mathbb{E}[\|H(\theta_t)\|^2]\Big) \leq \mathbb{E}[\mathbb{V}_{f,0}] + \frac{1}{2}\mathbb{E}[\|H(\theta_0)\|^2] \\
&+ \frac{1}{2}\alpha\gamma \sum_{t=0}^{T-1} \mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] + \frac{2\ell_{F_{ag},1} + \ell_{H^2,1}}{4}\alpha^2 \sum_{t=0}^{T-1} \mathbb{E}[\|d_t\|^2] \qquad \text{(G.15)}
\end{aligned}
$$

where $\mathbb{E}[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2] = \mathbb{E}[\mathbb{E}_t[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2]]$ can be further bounded by

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}_t[\|\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)\|^2]] \leq &2\mathbb{E}[\|A_{ag}\nabla F(\theta_t)^\top \nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + 2c_h^2 \mathbb{E}[\|H(\theta_t)\|^2] + \mathbb{E}[\mathrm{Var}_t[\tilde{\nabla}_\lambda \varphi(\lambda_t;\theta_t)]] \\
\leq &2\|A_{ag}\|^2 M \ell_f^2 \mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + 2c_h^2 \mathbb{E}[\|H(\theta_t)\|^2] + \sigma^2. \qquad \text{(G.16)}
\end{aligned}
$$

Similarly, $\mathbb{E}[\|d_t\|^2]$ can be further bounded by

$$
\mathbb{E}[\|d_t\|^2] \leq \mathbb{E}[\|\nabla F(\theta_t) A_{ag}^\top \lambda_t\|^2] + \sigma^2. \qquad \text{(G.17)}
$$

Plugging (G.16) and (G.17) back into (G.15), and setting $\alpha \leq 1/(2\ell_{F_{ag},1} + \ell_{H^2,1})$, $\gamma \leq \min\left\{\frac{1}{4\|A_{ag}\|^2 M\ell_f^2}, \frac{1}{2c_h}\right\}$, we have

$$\sum_{t=0}^{T-1} \alpha\Big(\mathbb{E}[\|\nabla F(\theta_t)A_{ag}^\top \lambda_t\|^2] + c_h\mathbb{E}[\|H(\theta_t)\|^2]\Big) \leq \mathbb{E}[\mathbb{V}_{f,0}] + \frac{1}{2}\mathbb{E}[\|H(\theta_0)\|^2] + \frac{1}{2}\alpha\gamma T\sigma^2 + \frac{2\ell_{F_{ag},1} + \ell_{H^2,1}}{4}\alpha^2 T\sigma^2 \tag{G.18}$$

Following similar arguments as (F.61) - (F.62), we require the step sizes

$$\alpha \leq \frac{1}{4(\ell_{f,1}\|A_{ag}^\top\lambda\|_1 + M\ell_f^2) + \ell_H} \tag{G.19}$$

$$\gamma \leq \min\left\{\frac{\alpha}{T}, \frac{1}{4\|A_{ag}\|^2 M\ell_f^2}, \frac{1}{2c_h}\right\} \tag{G.20}$$

where $\ell_{F_{ag},1} = \ell_{f,1}\|A_{ag}^\top\lambda\|_1$. This ensures λ_h is bounded and then one can choose $\alpha = \Theta(T^{-\frac{1}{2}})$, $\gamma = \Theta(T^{-\frac{3}{2}})$ to obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\|\nabla F(\theta_t)A_{ag}^\top\lambda_t\|^2 + \|H(\theta_t)\|^2\Big] = \mathcal{O}\Big(T^{-\frac{1}{2}}\Big). \tag{G.21}$$

□

H Implementation Details and Additional Experiment Results

In this section, we report the additional implementation details omitted from the main text in Appendix H.1 and the additional experimental results in Appendix H.2.

H.1 Implementation details

Computation. All experiments were conducted on a server with an Intel i9-7920X CPU, two NVIDIA A5000 GPUs and two NVIDIA A4500 GPUs.

For all the experiments reported in the main text except for the multi-lingual speech recognition experiment, we exactly follow the settings from [30]. The implementations of the baselines including LS, PMTL, and EPO are from the official code of the EPO paper in https://github.com/dbmptr/EPOSearch with their default hyperparameters. The results of XWC-MGDA are directly referenced from the paper due to lack of official implementation.

Synthetic data. For the results in both Figure 3 and Figure 4, the model parameter θ has dimension $q = 20$, the number of objectives is $M = 2$. The angles between the preference vectors and the horizontal axis are generated between $[\frac{1}{20}\pi, \frac{9}{20}\pi]$ with equal angular distance. This experiment does not involve stochastic optimization. For our method, we solve the subprogram using PGD with a step size 0.1 up to an error of 10^{-5} or with a maximum of 250 iterations. In the experiments, we set the parameter $c_h = 1$ for the subprogram if not otherwise specified.

In Figure 3, for all preferences and all methods, the initial model parameter θ_0 is randomly generated from a Gaussian distribution $\mathcal{N}(0, 1)$ for each dimension. In Table 6, we provide a summary of the hyperparameters for the baselines and our methods for the experiments in Figure 3.

Table 6: Summary of hyper-parameters for the synthetic data experiments in Figure 3.

	LS	MGDA	PMTL	EPO	Ours Figure 3e	Ours Figure 3f
step size α_t	0.1	0.2	0.2	0.1	0.05	0.05
max iterations	150	150	150	100	100	100

In Figures 4a-4c, the initial model parameters are randomly generated from a uniform distribution between $[-0.3, 0.3]$ for each dimension. In Figures 4d-4f, the initial model parameters are randomly generated from a uniform distribution between $[-0.5, -0.15]$ or $[0.15, 0.5]$ for each dimension. Table 7 summarizes the hyperparameters for the experiments in Figure 4.

Table 7: Summary of hyper-parameters for the synthetic data experiments in Figure 4.

	Figures 4a-4c			Figures 4d-4f		
	PMTL	EPO	Ours	PMTL	EPO	Ours
step size α_t	0.25	0.10	0.60	0.50	0.20	0.60
max iterations	100	60	10	200	120	200
c_h	-	-	1	-	-	0.01

Multi-patch image classification. For a fair comparison, we follow the same data splitting and processing procedures as [30] using their official code. In each of the three datasets, there are 120k samples for training and 20k samples for testing. There are two tasks on each dataset: 1) classifying the top-left image, and 2) classifying the bottom-right image.

For all methods, we use the SGD optimizer with batch size 256. Note that, for our stochastic method, we use batch size 128 for each batch in the double sampling. Thus the total number of samples taken at each iteration is also 256. The hyperparameters are summarized in Table 8. The results of XWC-MGDA are directly referenced from the paper.

Table 8: Summary of hyper-parameter choices for multi-patch image classification experiments.

	Multi-MNIST				Multi-Fashion				Multi-Fashion+MNIST			
	LS	PMTL	EPO	Ours	LS	PMTL	EPO	Ours	LS	PMTL	EPO	Ours
step size α_t	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3	1E-3
step size γ_t	-	-	-	1E-4	-	-	-	1E-4	-	-	-	1E-4
epochs	100	100	100	100	100	100	100	100	100	100	100	100
c_h	-	-	-	0.5	-	-	-	0.5	-	-	-	0.5

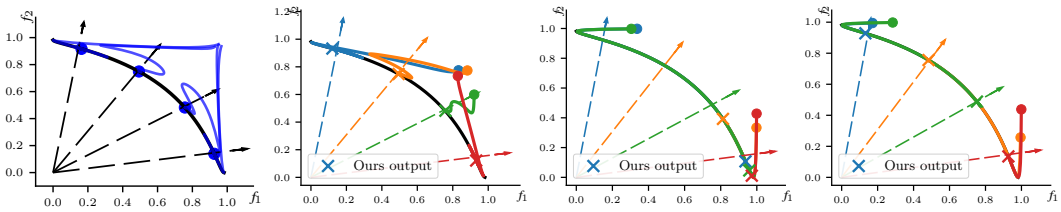
We use the Pymoo 0.6.1 library to compute the hypervolume. The Nadir points, i.e., the worst performance on single task baselines, used for the hypervolume computation are given in Table 9. For a fair comparison, the Nadir points we use are the same with [33] inferred from Figure 4 in the paper.

Table 9: Nadir points for the hypervolume computation

Dataset and metrics	Nadir points, metrics on objective $[1, \ldots, M]$
Multi-MNIST loss	[0.500, 0.450]
Multi-Fashion loss	[0.840, 0.800]
Multi-F+M loss	[0.625, 0.575]
Multi-MNIST accuracy	[0.830, 0.848]
Multi-Fashion accuracy	[0.840, 0.800]
Multi-F+M accuracy	[0.790, 0.785]
Emotion loss	[0.551, 0.636, 0.690, 0.539, 0.603, 0.570]

Multi-lingual speech recognition. We use two datasets, Librispeech and AISHELL v1. Librispeech is an English speech dataset that consists of 960 hours of labeled audio data. For our experiments, we use the "train-clean-100" subset of the Librispeech dataset for supervised training, which contains 100 hours of clean training data. Additionally, we use the full 960 hours of data for self-supervised training. AISHELL v1 is a 178-hour Mandarin speech corpus designed for various speech and speaker processing tasks. We use the full AISHELL v1 dataset for both self-supervised and supervised training. We combine these two datasets for our multi-lingual speech recognition experiments.

We use the conformer [18] model with 8 conformer blocks as the encoder. Each block contains 512 hidden units and 8 attention heads. Each attention head has dimension 64. The convolutional kernel

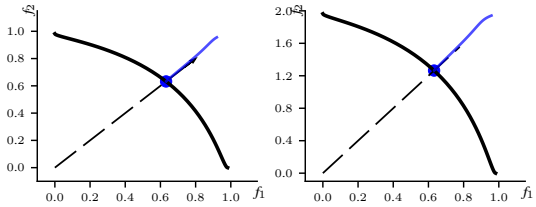


(a) Standard (b) Easy init. (c) Hard init. $A = I_M$ (d) Hard init. general A

Figure 8: Synthetic experiment results with Algorithm 2.

size is 31. Two classification heads are used. They contain two linear layers, one with 1000 output size for English, and another with 5000 output size for Chinese.

The loss functions we use include the Contrastive Predictive Coding (CPC) loss, and the Connectionist Temporal Classification (CTC) loss. The *CPC loss* [35] is a self-supervised loss to learn robust representations from unlabeled speech data. The CPC loss is designed to maximize the probability of a future sample given a contextual representation generated from the current speech sequence. The *CTC loss* is defined as the negative log-likelihood of the model parameter given the input sequence and the label sequence.



(a) Without scale (b) With scale

Figure 7: Scale invariance verification.

For all methods including the baselines, we use the step sizes $\alpha_{t,1} = 5 \times 10^{-4}$ for training backbone conformer parameters and $\alpha_{t,2} = 5 \times 10^{-5}$ for training classification head parameters. The step size $\gamma_t = 0.1$ and the parameter $c_h = 0.5$.

H.2 Additional experiment results

Synthetic data. We conduct several additional experiments on the synthetic objectives to further verify our theory. First, we conduct all the experiments on the synthetic objectives reported in the main text, using the single-loop approximate algorithm described in Algorithm 2. The results are plotted in Figure 8. The hyperparameters are the same unless otherwise specified.

From Figure 8a, we can see that Algorithm 2 with a one-step approximate update of λ_t also leads to convergence and preference alignment. However, different from the results obtained by exactly solving for $\lambda^*(\theta_t)$ at each iteration, the models on the optimization trajectories do not align exactly with the preference. Similar observations can be found in Figure 8b. In Figure 8c, which is a difficult case due to the initialization, $A = I_M$ does not work since it does not incorporate more general relative preference to allow controlled ascent update. This is addressed in Figure 8d, where a general A (the same as in prior experiments) is used. Compared with exactly solving for $\lambda^*(\theta_t)$ at each iteration, the approximate algorithm takes more iterations to converge, but has smaller per-iteration complexity, and smaller total time complexity.

Table 10: Summary of hyper-parameters for the synthetic data experiments in Figure 8.

	Figure 8a	Figure 8b	Figure 8c	Figure 8d
step size α_t	0.10	0.06	0.15	0.15
max iterations	100	100	250	250
c_h	6	6	0.1	0.1

We conduct another experiment to verify that the scale invariance can be preserved. We use the same objective as above, but scale the second one by 2. We use a fixed initialization $\theta_0 = 0.3 \cdot [\mathbf{1}_{q/2}; -\mathbf{1}_{q/2}]$ for this experiment. The other hyperparameters are the same as the default. We use both $F(\theta_0)$ and

$F(0)$ as the reference points and choose B_h such that $B_h(F(\theta_0) - F(0)) = 0$. Results in Figure 7 show that for different scales, the trajectory and the converging solution are the same.

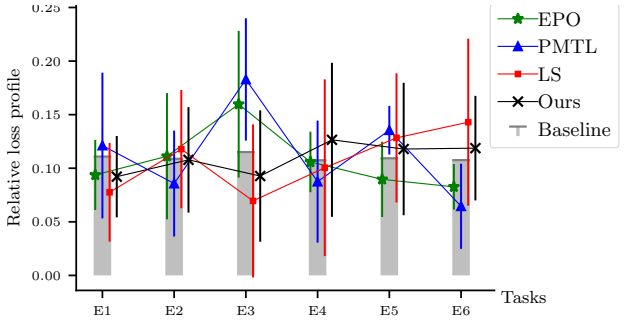


Figure 9: Relative loss profile for all methods on Emotions and Music dataset.

Table 11: Summary of hyper-parameter choices for emotion recognition experiments.

	LS	PMTL	EPO	Ours
step size α_t	1E-3	1E-3	1E-3	1E-3
step size γ_t	-	-	-	1E-4
batch size	50	50	50	50
epochs	200	200	200	200

Table 12: Summary of average run time in seconds (s) or minutes (m) and number of iterations or epochs of different methods on different datasets. We use Algorithm 1 for the synthetic experiments, and Algorithm 3 for the other two experiments.

Datasets	Metrics	LS	PMTL	EPO	FERERO
Synthetic, Figures 3(a-c)	Iterations	100	100	60	10
	Per-iteration run time	3.50E-4s	7.67E-4s	4.93E-3s	7.50E-4s
	Total run time	0.035s	0.0767s	0.296s	0.0075s
Synthetic, Figures 3(d-f)	Iterations	100	200	80	200
	Per-iteration run time	3.10E-4s	7.65E-4s	4.93E-3s	7.30E-4s
	Total run time	0.031s	0.153s	0.394s	0.146s
Multi-MNIST/Fashion/F+M	Epochs	100	100	100	100
	Per-epoch run time	3.54s	11.88s	9.66s	7.02s
	Total run time	5.9m	19.8m	16.1m	11.7m
Emotion	Epochs	200	200	200	200
	Per-epoch run time	9.5E-3s	0.496s	0.238s	0.039s
	Total run time	1.9s	99.1s	47.6s	7.70s

Emotion recognition. The task is to predict 6 types of emotions from 593 songs based on the Tellegen Watson-Clark model of affect. The 6 emotions include: amazed-surprised (E1), happy-pleased (E2), relaxing-calm (E3), quiet-still (E4), sad-lonely (E5), and angry-fearful (E6). Following [30], we use the fully connected neural network with 4 layers as the model architecture. The Sigmoid cross entropy loss is used as the objective for each task. And 10 preference vectors are generated uniformly. The hyperparameters used in this experiment are summarized in Table 11.

The results on the relative loss profile (RLP) are reported in Figure 9. Results show that all methods, including LS, work similarly well. EPO achieves the highest hypervolumes, and our proposed approach obtains the second-best hypervolumes. One reason could be that the Pareto front in this problem is convex.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1, introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Broader impacts and limitations section.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Assumptions 1, 2, 3 for the assumptions, and the Appendix D, and G for the proof.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 5 and Appendix H.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at https://github.com/lisha-chen/FERERO/.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5 and Appendix H.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 5. We use the standard deviations as the error bars for all experiments except the speech recognition experiments since the speech recognition experiments take much longer time to run.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5 and Appendix H.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We preserve anonymity.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See the end of the main paper in the Broader impacts and limitations section.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 5 and Appendix H.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects