RETHINKING MULTI-OBJECTIVE RESOLUTION IN FEDERATED LLM ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning large language models (LLMs) requires balancing potentially conflicting objectives, such as helpfulness and harmlessness. Federated learning can support alignment when preference data are distributed across clients, but multi-objective alignment introduces a distinct conflict-resolution challenge. Server-side conflict resolution may require transmitting an objective-specific gradient for every client and objective, whereas independent client-side resolution can produce inconsistent objective weights because of stochastic estimation, local model drift, and client heterogeneity. We call the resulting inconsistency *multi-objective disagreement drift*. We propose **FIRM** (**F**ederated **I**n-client **R**egularized **M**ulti-objective LLM alignment), a communication-efficient client-centric framework that resolves objective conflicts locally and communicates one model update per client. FIRM regularizes each client's multiple-gradient descent (MGDA) subproblem, reducing the sensitivity of its objective weights to perturbations in estimated gradients. For a federated multi-objective actor–critic abstraction, we derive a finite-time bound on the average Pareto-stationarity gap that separates standard federated client drift from the additional disagreement induced by client-side conflict resolution. The bound establishes convergence to a residual neighborhood of Pareto stationarity and characterizes how this disagreement depends on regularization strength, batch size, local updating, and client heterogeneity. Experiments on federated LLM alignment show that FIRM reduces client objective-weight disagreement and improves the observed reward trade-off relative to the evaluated baselines. Additional experiments demonstrate stable behavior under heterogeneous reward models, stronger non-IID partitions, larger client populations, partial participation, additional objectives, and larger model scale.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly used in applications ranging from digital assistants to scientific discovery (Snell et al., 2023; Pyatkin et al., 2023). Their deployment, however, raises the fundamental problem of alignment: adapting model behavior to human preferences and values (Ouyang et al., 2022). This is inherently a multi-objective problem, as alignment often requires balancing multiple, potentially conflicting goals such as helpfulness and harmlessness (Kong et al., 2025; Yang et al., 2024). Existing alignment methods are nevertheless commonly developed in centralized settings, requiring preference data and substantial computation to be concentrated within a single infrastructure. This can limit participation and may not be suitable when datasets are distributed or cannot be transferred to a central server (Wu et al., 2025). Federated learning (FL) offers an alternative in which clients optimize models using local data and communicate model updates (McMahan et al., 2017). Applying FL to multi-objective LLM alignment, however, introduces a fundamental design question: *Should the resolution of conflicting objectives happen at the server or locally at each client?*

Recent works in federated multi-objective supervised learning have explored the server-centric approach to conflict resolution. A naive implementation requires each of the C participating clients to transmit a d-dimensional gradient for each of the M objectives to a central server, which then finds a a shared multi-objective update direction (Yang et al., 2023). The resulting client-to-server payload is therefore $\mathcal{O}(CMd)$ per server-side conflict-resolution round as in Figure 1. This factor-of-M is particularly consequential when d represents millions of trainable parameters. Other ap-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

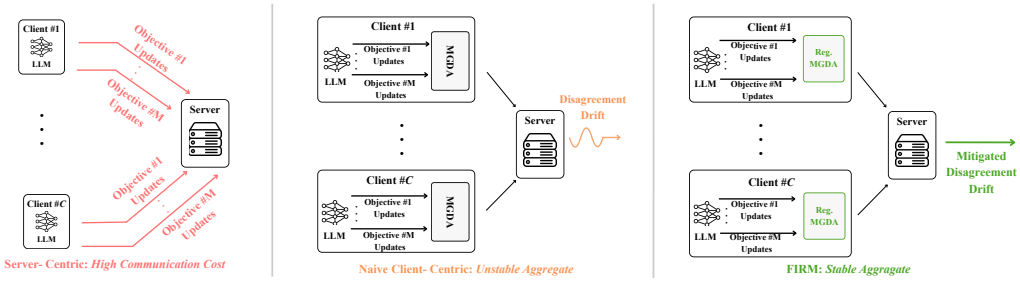


Figure 1: **Comparison of federated multi-objective alignment paradigms. Left: Naive server-centric** transmits objective-specific gradients to the server, increasing communication to $\mathcal{O}(CMd)$. **Middle: Naive client-centric** communicates one update per client but can produce inconsistent local MGDA directions due to *disagreement drift*. **Right: FIRM** regularizes local MGDA weights to mitigate disagreement drift before aggregation.

proaches reduce this dependence through compression of objective-specific gradient information (Askin et al., 2025), at the cost of additional compression and client-server coordination beyond standard federated parameter aggregation. These methods are also developed primarily for supervised multi-objective learning, whereas LLM alignment requires resolving competing objectives within a reinforcement-learning framework. This motivates performing conflict resolution during local training while communicating only one adapted model update per client.

We propose **Federated In-Client Regularized Multi-objective alignment (FIRM)**, a communication-efficient client-centric framework for federated multi-objective LLM alignment. Each client computes gradients for all M objectives using Proximal Policy Optimization (PPO) (Schulman et al., 2017) and combines them locally into a single update direction by solving a *Multiple-Gradient Descent Algorithm (MGDA)* subproblem (Désidéri, 2012). The central server then aggregates the updated client parameters via FedAvg (McMahan et al., 2017). Consequently, FIRM retains the standard $\mathcal{O}(Cd)$ client-to-server communication of parameter aggregation and does not require clients to transmit all objective-specific gradients to the server. Naively moving MGDA from the server to the clients, however, creates a previously underexplored disagreement. The solution of an MGDA subproblem can be sensitive to perturbations in its objective gradients. Stochastic gradient estimation, imperfect critics, local optimization, and data heterogeneity can therefore cause different clients to select different objective weights, even when they begin a communication round from the same global model. When aggregated at the server, these conflicting updates weaken the aggregated update (see Section 4 and 5). We call this phenomenon ***multi-objective disagreement drift***; Figure 1 (middle) illustrates its effect.

To control this disagreement, we equip FIRM with a quadratic regularization term in each client's local MGDA subproblem, reducing the sensitivity of the selected objective weights to stochastic and client-specific gradient variation (Figure 1 (right)). For a federated multi-objective actor–critic abstraction that preserves FIRM's client-side objective weighting and local-update structure, we derive a finite-time bound on the average Pareto-stationarity gap, establishing convergence to a neighborhood of Pareto stationarity. When the function-approximation and critic-estimation errors are small, the disagreement contribution scales as $\mathcal{O}\big(\frac{M^{3/2}}{\beta}(\frac{1}{\sqrt{B}} + \alpha K + \zeta_{\text{het}})\big)$, where B is the batch size, α is the actor step size, K is the number of local updates, β is the regularization parameter, and ζ_{het} characterizes client heterogeneity. Importantly, the bound shows that disagreement can arise even among nearly homogeneous clients due to stochastic gradient estimation and local policy updates, while regularization limits how strongly these differences propagate to the client-side objective weights. Our experiments show that FIRM improves multi-objective alignment while reducing the client-side objective-weight disagreement observed without regularization. We further evaluate FIRM across heterogeneous reward models and broader federated settings. Detailed results are provided in Section 5 and Appendix A.

Our main contributions are as follows: 1) We introduce FIRM, a federated multi-objective alignment framework that performs regularized conflict resolution within local RL training and communicates

a single adapted update per client, retaining $\mathcal{O}(Cd)$ client-to-server communication without server-side conflict resolution. 2) We identify *multi-objective disagreement drift*, a distinct source of error induced by independently computed client-side objective weighting, and characterize how it arises from stochastic gradient estimation, local policy evolution, and client heterogeneity. 3) We establish a finite-time bound on the average Pareto-stationarity gap for a federated multi-objective actor–critic abstraction of FIRM, explicitly separating multi-objective disagreement from classical local-model drift. 4) We empirically show that FIRM reduces objective-weight disagreement and improves both alignment objectives, with consistent performance across heterogeneous reward models, additional objectives, stronger non-IID settings, varying client configurations, and a larger LLM backbone.

Additional Related Work. We have provided a comprehensive literature review in Appendix B.

2 Problem Formulation

We consider federated LLM alignment with multiple potentially conflicting reward objectives. A central server coordinates C clients, indexed by $[C] := \{1, \dots, C\}$. Each client c holds a local prompt dataset $\mathcal{D}_c$ and maintains a local copy of a shared policy π_θ, where $\theta \in \mathbb{R}^d$ denotes the trainable parameters communicated during federation. At the beginning of communication round r, the server broadcasts the current global parameters θ_r to a set of participating clients $\mathcal{C}_r \subseteq [C]$. Each participating client initializes $\theta^c_{r,0} = \theta_r$, samples prompts from its local dataset $\mathcal{D}_c$, generates responses using its local policy, and evaluates them under M reward objectives, such as helpfulness and harmlessness. After K local policy updates, client c returns its adapted parameters $\theta^c_{r,K}$ to the server. The server then aggregates the parameters as $\theta_{r+1} = \sum_{c\in\mathcal{C}_r} p_{r,c}\theta^c_{r,K}$, where $p_{r,c} \geq 0$ and $\sum_{c\in\mathcal{C}_r} p_{r,c} = 1$. FIRM permits partial participation and general aggregation weights, whereas our finite-time analysis in Section 4 assumes full participation and uniform aggregation. For analysis, we model the above alignment process through a *federated multi-objective reinforcement learning* (FedMORL) abstraction. In this abstraction, each client is associated with a multi-objective Markov decision process (MOMDP).

Definition 2.1 (Client MOMDP)**.** Client c is associated with a multi-objective Markov decision process $\mathcal{M}_c = (\mathcal{S}, \mathcal{A}, P_c, \mathbf{r}_c)$, where $\mathcal{S}$ and $\mathcal{A}$ are shared finite state and action spaces, $P_c(\cdot \mid s, a)$ is the client transition kernel, and $\mathbf{r}_c(s,a) = \left[r^1_c(s,a), \dots, r^M_c(s,a)\right]^\top \in \mathbb{R}^M$ is the vector-valued reward function. Trajectories are initialized from a common distribution ρ, and returns are discounted by $\gamma \in (0,1)$.

The formulation permits clients to differ in their transition dynamics and reward functions. Importantly, client heterogeneity is not necessary for multi-objective disagreement: even when clients have identical MOMDPs and start from the same global policy, independently sampled finite batches can produce different objective-gradient estimates and therefore different locally selected objective weights. Client heterogeneity may further amplify this disagreement.

Global Multi-Objective Problem. For objective i at client c, define the discounted return $J^{i,c}(\theta) := \mathbb{E}_{\substack{s_0\sim\rho,\, a_\ell\sim\pi_\theta(\cdot|s_\ell),\\ s_{\ell+1}\sim P_c(\cdot|s_\ell,a_\ell)}}\left[\sum_{\ell=0}^{\infty}\gamma^\ell r^i_c(s_\ell,a_\ell)\right]$. The corresponding local vector objective is $\mathbf{J}^c(\theta) := \left[J^{1,c}(\theta), \dots, J^{M,c}(\theta)\right]^\top$, and the global vector objective is $\mathbf{J}(\theta) := \frac{1}{C}\sum_{c=1}^{C}\mathbf{J}^c(\theta)$.

Thus, the goal is not to optimize a single scalar objective, but to find a policy that achieves a trade-off among the M global objectives. We formalize such trade-offs through Pareto optimality.

Definition 2.2 (Pareto optimality)**.** A policy π_θ Pareto-dominates $\pi_{\theta'}$ if $\mathbf{J}(\theta) \geq \mathbf{J}(\theta')$ element-wise and $\mathbf{J}(\theta) \neq \mathbf{J}(\theta')$. A policy $\pi_{\theta^\star}$ is Pareto-optimal within the considered policy class if no other policy in that class Pareto-dominates it.

For non-convex objectives induced by neural policies, finding a globally Pareto-optimal policy is generally intractable. We therefore target Pareto stationarity, a standard first-order necessary condition for Pareto optimality (Désidéri, 2012; Zhou et al., 2024). Let $\Delta_M := \left\{\lambda \in \mathbb{R}^M_+ : \sum_{i=1}^M \lambda_i = 1\right\}$ denote the probability simplex. We define the Pareto-stationarity gap as $\mathcal{G}(\theta) := \min_{\lambda\in\Delta_M}\left\|\sum_{i=1}^M \lambda_i \nabla_\theta J^i(\theta)\right\|_2^2$.

Definition 2.3 (ϵ-Pareto stationarity)**.** A policy π_θ is ϵ-Pareto-stationary if $\mathcal{G}(\theta) \leq \epsilon$.

Our goal is therefore to resolve objective conflicts locally while controlling the cross-client disagreement introduced by client-side objective weighting, and to obtain global policies with small expected Pareto-stationarity gap while communicating only one adapted update per participating client.

3 FIRM ALGORITHM FOR LLM ALIGNMENT

We introduce **FIRM** (**F**ederated **I**n-client **R**egularized **M**ulti-objective alignment), a federated framework that resolves objective conflicts locally at each participating client. In direct server-side MGDA schemes, clients communicate objective-specific gradients and the server computes a common objective-weight vector. FIRM instead computes the objective weights locally from each client's estimated objective gradients, combines the gradients before communication, and sends only the resulting adapted model parameters to the server. Multiple-gradient methods have previously been used for centralized multi-objective LLM alignment, including GAPO (Gu et al., 2025). FIRM addresses an additional difficulty that arises when conflict resolution is performed locally across federated clients: differences in stochastic gradient estimates, local policy parameters, and client objectives can cause different clients to select different MGDA weight vectors. This cross-client discrepancy gives rise to the *multi-objective disagreement drift*. This section describes the practical FIRM implementation used in our experiments, which combines PPO with LoRA adaptation; Section 4 introduces TFIRM, an actor–critic abstraction used for finite-time analysis.

Client-Side Conflict Resolution. At communication round r, the server broadcasts the global parameters θ_r to the participating clients $\mathcal{C}_r \subseteq [C]$. Client c initializes $\theta^c_{r,0} = \theta_r$ and performs K local policy updates. At local step k, the client samples a batch of B prompts, generates responses using the current policy $\pi_{\theta^c_{r,k}}$, and evaluates them under the M reward objectives. For each objective $j \in [M]$, it constructs an objective-specific PPO surrogate and computes a stochastic gradient $g^{j,c}_{r,k}$, oriented as an ascent direction for that surrogate. The client then selects its local objective weights by solving

$$\widehat{\lambda}^{*,c}_{r,k} = \arg\min_{\lambda \in \Delta_M} \left\{ \underbrace{\left\| \sum_{j=1}^{M} \lambda_j g^{j,c}_{r,k} \right\|_2^2}_{\mathcal{T}_1} + \underbrace{\frac{\beta}{2}\|\lambda\|_2^2}_{\mathcal{T}_2} \right\}, \qquad \beta > 0. \tag{1}$$

The first term, $\mathcal{T}_1$, is the standard MGDA objective, which selects a minimum-norm convex combination of the objective-specific update directions (Désidéri, 2012). The second term, $\mathcal{T}_2$, reduces the sensitivity of the client-side objective weights under stochastic gradient perturbations. Using these weights, the client forms the combined gradient $g^c_{r,k} = \sum_{j=1}^{M} \widehat{\lambda}^{*,c}_{r,k,j} g^{j,c}_{r,k}$ and performs the local update $\theta^c_{r,k+1} = \theta^c_{r,k} + \alpha g^c_{r,k}$. After K local steps, client c sends its final adapted parameters $\theta^c_{r,K}$ to the server, which aggregates the participating clients as in Section 2. Thus, each participating client sends one d-dimensional adapted parameter vector per communication round. Algorithm 1 summarizes the practical procedure.

Controlling Objective-Weight Disagreement Through Regularization. Let $A^c_{r,k} = [g^{1,c}_{r,k}, \ldots, g^{M,c}_{r,k}]$ and $G^c_{r,k} = (A^c_{r,k})^\top A^c_{r,k}$. Then Eq. 1 can equivalently be written as $\widehat{\lambda}^{*,c}_{r,k} = \arg\min_{\lambda \in \Delta_M} \lambda^\top \left(G^c_{r,k} + \frac{\beta}{2} I \right) \lambda$. In our practical implementation, each local Gram matrix is normalized by its mean diagonal before applying the regularizer so that β is measured relative to the local gradient scale. Strongly convex regularization is widely used in optimization to improve solution stability under perturbations, including in multi-objective learning to regularize dynamic objective weighting (Parikh & Boyd, 2014; Chen et al., 2024b). In the federated setting, we identify a distinct consequence of this sensitivity when MGDA is solved independently across federated clients. Because each client constructs its MGDA subproblem from locally estimated gradients, even small discrepancies in these gradients can translate into different objective-weight vectors across clients. We refer to the resulting cross-client inconsistency as *multi-objective disagreement drift*. Adding $\beta > 0$ makes each local subproblem β-strongly convex and yields a unique optimizer. Section 4 further shows that the sensitivity of the selected objective weights to gradient perturbations admits a bound proportional to $1/\beta$. Thus, while regularization itself is a standard optimization device, in

Algorithm 1 Federated In-Client Regularized Multi-objective Alignment (FIRM)

Input: Communication rounds R, local steps K, batch size B, learning rate α, and regularization parameter $\beta > 0$.
Initialize global policy parameters θ_0.
for $r = 0, 1, \ldots, R-1$ **do**
  Select participating clients $\mathcal{C}_r \subseteq [C]$.
  Server broadcasts θ_r to clients in $\mathcal{C}_r$.
  for each client $c \in \mathcal{C}_r$ **in parallel do**
    Initialize $\theta^c_{r,0} \leftarrow \theta_r$.
    for $k = 0, 1, \ldots, K-1$ **do**
      Sample B prompts and generate responses using $\pi_{\theta^c_{r,k}}$.
      Evaluate each response under the M reward objectives.
      for $j = 1, 2, \ldots, M$ **do**
        Compute the objective-specific PPO gradient $g^{j,c}_{r,k}$, oriented as an ascent direction.
      end for
      Solve Eq. equation 1 to obtain $\widehat{\lambda}^{*,c}_{r,k}$.
      $g^c_{r,k} \leftarrow \sum_{j=1}^{M} \widehat{\lambda}^{*,c}_{r,k,j} g^{j,c}_{r,k}$.
      $\theta^c_{r,k+1} \leftarrow \theta^c_{r,k} + \alpha g^c_{r,k}$.
    end for
    Client c sends $\theta^c_{r,K}$ to the server.
  end for
  $\theta_{r+1} \leftarrow \sum_{c \in \mathcal{C}_r} p_{r,c} \theta^c_{r,K}$.
end for

FIRM it provides a mechanism for controlling the cross-client objective-weight disagreement that we identify in federated multi-objective alignment.

4 Theoretical Analysis

4.1 Theoretical Setup and Assumptions

To obtain a tractable finite-time analysis, we study a theoretical actor–critic variant of FIRM, referred to as *TFIRM*, in which the practical PPO update is replaced by a multi-objective actor–critic update. This formulation allows the effects of Markovian sampling, critic estimation, local policy updating, and client-side conflict resolution to be analyzed explicitly while preserving the mechanisms central to FIRM—objective-specific stochastic policy-gradient estimates, client-side regularized MGDA, multiple local updates, and periodic federated aggregation. In TFIRM, each client maintains M objective-specific value estimates, which are updated through mini-batch temporal-difference learning using the function-approximation model (see Assumption 4.2). The actor uses the corresponding TD errors as objective-specific advantage estimates and constructs one stochastic actor-gradient estimate for each objective using the policy-gradient theorem (Sutton et al., 1999). As in practical FIRM, each client then solves a regularized MGDA subproblem using its locally estimated objective gradients and performs K local actor updates between synchronizations. The actor estimator is constructed so that it corresponds to the same discounted-return objective defined in Section 2; the corresponding details are provided in Appendix D.

Our analysis builds on the finite-time multi-objective actor–critic framework of Zhou et al. (2024). Relative to the centralized setting, federation introduces two effects that are central here: local policies evolve between synchronizations, and different clients can compute different MGDA weight vectors from their local stochastic gradients. Our analysis separates the resulting classical local-model drift from the additional cross-client multi-objective disagreement and uses the perturbation stability of regularized MGDA to control the latter. The complete TFIRM algorithm is provided in Appendix C; we proceed under the following standard assumptions.

Assumption 4.1. For each client $c \in \{1, \ldots, C\}$, its local MOMDP and the global policy satisfy: (a) The policy $\pi_\theta(a \mid s)$ is continuously differentiable with respect to θ. (b) The Markov process induced by any policy π_θ in any client's environment P_c is irreducible and aperiodic, with the

geometric mixing constants used in Appendix D chosen uniformly over the clients. (c) The instantaneous reward $r_t^{i,c}$ for any objective i on any client c is non-negative and uniformly bounded by a constant $r_{\max} > 0$.

Assumption 4.1 imposes standard regularity conditions (Zhou et al., 2024). Condition (a) permits application of policy-gradient methods. Condition (b) ensures that every client-policy pair admits a unique stationary distribution, denoted by $\mu_{c,\theta}$ (Appendix D). Condition (c) ensures bounded value functions and actor updates.

Assumption 4.2 (Critic Approximation)**.** We make the following assumptions on the value function approximation: (a) For each objective $j \in [M]$, the value function is approximated from a linear function class, $V_w^j(s) = \phi(s)^\top \mathbf{w}^j$, where $\mathbf{w}^j \in \mathbb{R}^{d_2}$ are learnable parameters and $\phi : \mathcal{S} \to \mathbb{R}^{d_2}$ is a shared feature map. (b) The feature map is normalized such that $\|\phi(s)\|_2 \leq 1$ for all $s \in \mathcal{S}$. (c) Let $\mathbf{A}_{c,\theta} := \mathbb{E}_{s\sim\mu_{c,\theta},s'\sim P_{c,\theta}(\cdot|s)}[(\gamma\phi(s') - \phi(s))\phi(s)^\top]$. We assume this matrix is negative definite, i.e., there exists a constant $\lambda_A > 0$ such that its symmetric part satisfies $\mathbf{x}^\top(\mathbf{A}_{c,\theta} + \mathbf{A}_{c,\theta}^\top)\mathbf{x} \leq -2\lambda_A\|\mathbf{x}\|_2^2$.

Assumption 4.2 is standard in the finite-time analysis of linear temporal-difference learning (Tsitsiklis & Van Roy, 1999; Xu et al., 2020a;b; Qiu et al., 2021). The conditions collectively ensure that the projected Bellman equation is well-posed for any policy π_θ.

Assumption 4.3 (Boundedness and Smoothness)**.** For any policy parameters θ, θ' and any state-action pair (s, a), there exist positive constants C_ψ and L_J such that: (a) The score function is uniformly bounded: $\|\psi_\theta(a|s)\|_2 = \|\nabla_\theta \log \pi_\theta(a|s)\|_2 \leq C_\psi$.

(b) The gradient of each local objective function $J^{i,c}(\theta)$ is Lipschitz continuous with respect to the policy parameter: $\|\nabla_\theta J^{i,c}(\theta) - \nabla_\theta J^{i,c}(\theta')\|_2 \leq L_J\|\theta - \theta'\|_2$ for all objectives $i \in [M]$ and all clients $c \in \mathcal{C}$.

This assumption imposes standard smoothness and boundedness conditions. In particular, the bounded-score condition in (a) prevents unbounded gradient updates (Hairi et al., 2022).

Assumption 4.4 (Bounded Heterogeneity)**.** There exists a non-negative constant ζ_{het} such that for all clients $c \in \{1, \ldots, C\}$ and any policy parameter θ, the deviation between the local and global objective gradients is bounded: $\|\nabla_\theta \mathbf{J}^c(\theta) - \nabla_\theta \mathbf{J}(\theta)\|_F \leq \zeta_{\text{het}}$.

The constant ζ_{het} quantifies the degree of heterogeneity across clients. We provide the formal derivation of this relationship in Appendix I.

Theorem 4.5 (Informal Finite-Time Pareto-Stationarity Bound for TFIRM)**.** *Suppose Assumptions 4.1– 4.4 hold. Let $\bar{\theta}_t := \frac{1}{C}\sum_{c=1}^{C} \theta_t^c$, let T denote the total number of actor-update steps, and let clients synchronize every K local updates. Let B be the actor mini-batch size, ζ_{approx} the critic function-approximation error, and $\varepsilon_{\text{critic}}$ the finite-time critic-estimation error. Under full client participation, uniform aggregation, the TFIRM weight-smoothing schedule $\eta_t = 1/t$, and a sufficiently small constant actor step size α, TFIRM satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{G}(\bar{\theta}_t)\right] \leq \mathcal{O}\Bigg(\underbrace{\frac{\log T}{\alpha T}}_{\textit{optimization}} + \underbrace{\frac{M}{CB} + \frac{M}{B^2}}_{\textit{actor sampling}} + \underbrace{\zeta_{\text{approx}} + \varepsilon_{\text{critic}}}_{\textit{critic residual}} + \underbrace{\alpha^2 K^2}_{\textit{local-model drift}} + \underbrace{\frac{M^{3/2}}{\beta}\left(\frac{1}{\sqrt{B}} + \alpha K + \sqrt{\zeta_{\text{approx}}} + \sqrt{\varepsilon_{\text{critic}}} + \zeta_{\text{het}}\right)}_{\textit{multi-objective disagreement}} \Bigg). \tag{2}$$

The formal theorem and proof are provided in Appendix F.

Interpretation of the Bound. Theorem 4.5 gives a finite-time bound on the average Pareto-stationarity gap and implies convergence to a residual neighborhood. Importantly, it separates the $\mathcal{O}(\alpha^2 K^2)$ error due to local-model drift from the additional error caused by client-side objective weighting. The latter captures multi-objective disagreement drift, while the perturbation stability of regularized MGDA controls how differences in local gradient estimates propagate to the aggregated update.

Remark 4.6 (Controlling Multi-Objective Disagreement). When function-approximation and critic-estimation errors are small, the disagreement term reduces to $\mathcal{O}\big(\frac{M^{3/2}}{\beta}(\frac{1}{\sqrt{B}} + \alpha K + \zeta_{\text{het}})\big)$. Thus, disagreement can persist even for nearly homogeneous clients because of finite-batch gradient estimation and local policy evolution. Larger batches, less local updating between synchronizations, and an appropriate β reduce this effect. A small β increases the sensitivity of the MGDA weights to gradient perturbations, while increasing β pulls the weights toward the uniform vector, educing sensitivity to gradient perturbations at the cost of adaptation to the local gradient geometry.

Proof Sketch. Define the averaged TFIRM update direction $\bar{g}_t := C^{-1}\sum_{c=1}^{C}\sum_{j=1}^{M}\lambda_t^{j,c} g_t^{j,c}$ and the averaged objective-weight vector $\bar{\lambda}_t := C^{-1}\sum_{c=1}^{C}\lambda_t^c$, so that $\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha\bar{g}_t$. Applying smoothness to the scalarized global objective $\bar{\lambda}_t^\top \mathbf{J}(\cdot)$ yields a progress inequality whose principal error is $\mathcal{E}_t := \|\bar{g}_t - \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2$. Up to universal constants, we decompose $\mathcal{E}_t \lesssim \mathcal{T}_t^{\text{grad}} + \mathcal{T}_t^{\text{dis}} + \mathcal{T}_t^{\text{model}}$, where $\mathcal{T}_t^{\text{grad}} := \|C^{-1}\sum_c\sum_j \bar{\lambda}_t^j(g_t^{j,c} - \nabla J^{j,c}(\theta_t^c))\|_2^2$ captures stochastic gradient-estimation error, $\mathcal{T}_t^{\text{dis}} := \|C^{-1}\sum_c\sum_j(\lambda_t^{j,c} - \bar{\lambda}_t^j)g_t^{j,c}\|_2^2$ captures cross-client objective-weight disagreement, and $\mathcal{T}_t^{\text{model}} := \|[C^{-1}\sum_c \nabla\mathbf{J}^c(\theta_t^c) - \nabla\mathbf{J}(\bar{\theta}_t)]\bar{\lambda}_t\|_2^2$ captures local-model drift. The gradient-estimation term is controlled by separating actor sampling, critic approximation, and critic estimation. Since the objective weights and gradient estimates depend on the same actor mini-batches, the proof avoids an independence assumption by applying a pathwise simplex bound before taking expectations, yielding $\mathcal{O}(M/(CB) + M/B^2 + \zeta_{\text{approx}} + \varepsilon_{\text{critic}})$. For the disagreement term, the perturbation stability of regularized MGDA converts discrepancies between client gradient matrices into discrepancies between their objective-weight vectors. Combining this with stochastic gradient error, critic error, local policy divergence, and client heterogeneity gives the $\mathcal{O}(\frac{M^{3/2}}{\beta}(B^{-1/2} + \alpha K + \sqrt{\zeta_{\text{approx}}} + \sqrt{\varepsilon_{\text{critic}}} + \zeta_{\text{het}}))$ term. Finally, smoothness and bounded local policy divergence yield $\mathcal{O}(\alpha^2 K^2)$ for the model-drift term, while the diminishing weight-smoothing used in TFIRM controls temporal variation of the objective weights; summation by parts then gives the $\mathcal{O}((\log T)/(\alpha T))$ optimization term.

5 Numerical Experiments

This section empirically validates our proposed framework. Our experiments are designed to answer three central research questions: **(RQ1) Multi-objective performance:** Can FIRM obtain a favorable helpfulness–harmlessness trade-off relative to federated multi-objective and fixed-scalarization baselines? **(RQ2) Disagreement Control:** Does regularization reduce cross-client objective-weight disagreement? **(RQ3) Robustness:** Does FIRM remain effective across model families and scales, three-objective alignment, larger client populations, and challenging federated settings?

Experimental Setup. We fine-tune `meta-llama/Llama-3.2-1B-Instruct`, `Qwen/Qwen3-1.7B`, and `meta-llama/Llama-3.1-8B-Instruct` using PPO and LoRA adapters (Schulman et al., 2017; Hu et al., 2022) to evaluate FIRM across model families and scales. The 1B–1.7B models represent resource-conscious federated alignment settings and provide a computationally tractable setting for federated LLM post-training, while the 8B experiment evaluates scalability to a larger policy. Only the trainable LoRA parameters are communicated between clients and the server. We consider $M = 2$ alignment objectives: helpfulness and harmlessness. Prompts are drawn from the Anthropic HH-RLHF dataset (Bai et al., 2022), and responses are evaluated using `Ray2333/gpt2-large-helpful-reward\_model` and `Ray2333/gpt2-large-harmless-reward\_model`, respectively (Yang et al., 2024). Our main comparison uses $C = 8$ clients with a non-IID $\text{Dir}(\alpha = 0.3)$ data partition. Each client performs local PPO updates between server aggregations, and all methods use the same model initialization, client partitions, prompt budget, generation settings, and trainable parameterization. FIRM uses $\beta = 0.01$, with other hyperparameters as described in Appendix A. We ran experiments with multiple random seeds on NVIDIA H200 GPUs and present the average results.

Baselines. We compare FIRM with the following methods. Fixed Weights uses the scalarization $\lambda = (0.5, 0.5)$ throughout training and does not perform MGDA. FSMGDA adapts the federated stochastic multi-gradient method of Yang et al. (2023): clients construct objective-specific local updates, which are combined through server-side conflict resolution. FedCMOO adapts the communication-efficient server-side method of Askin et al. (2025), which computes a shared

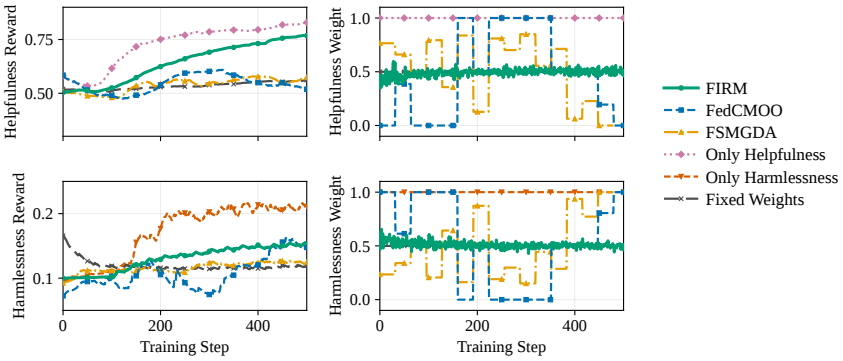


Figure 2: **Comparison of FIRM with baselines.** Helpfulness and harmlessness rewards (left) and objective weights (right) for FIRM and the evaluated baselines. FIRM weights are averaged across participating clients; the single-objective runs provide reference scores, while Fixed Weights uses $\lambda = (0.5, 0.5)$. FIRM steadily improves both rewards while maintaining smooth objective-weight trajectories.

objective-weight vector at the server, and returns it to the participating clients. All three methods use the same PPO/LoRA training pipeline as FIRM. We additionally include Only Helpfulness and Only Harmlessness. These single-objective runs remove the other objective from training and provide reference scores for the reward attainable when no multi-objective trade-off is required.

RQ1: Multi-Objective Reward Trade-off. Figure 2 compares reward and objective-weight trajectories under a common local-update budget. FIRM steadily improves both objectives, attaining the highest helpfulness among the evaluated multi-objective methods while retaining substantial improvement in harmlessness. Fixed Weights, FSMGDA, and FedCMOO show smaller or less consistent improvement in at least one objective. FIRM also maintains smooth interior objective weights, whereas the server-computed weights of FSMGDA and FedCMOO change abruptly and frequently approach simplex boundaries. In particular, FedCMOO computes a single shared weight vector from a round-level server estimate and holds it fixed during local training, so it does not adapt to within-round changes in the local gradient geometry and subsequent server updates can induce large discrete shifts. FIRM instead recomputes regularized objective weights client-side at each local update, enabling continuous local adaptation while controlling sensitivity to gradient perturbations.

Communication Overhead. We count the raw tensor payloads uploaded by each client. FIRM transmits one BF16 LoRA update ($\approx$ 9 MiB per client per round), whereas FSMGDA transmits $M = 2$ objective-specific FP32 payloads ($\approx$ 36 MiB), yielding $4\times$ larger uplink volume. At matched communication precision, FSMGDA still transmits $2\times$ more data because it communicates one full-dimensional payload per objective. Detailed accounting is provided in Appendix A.

RQ2: Regularization and Client Disagreement. For participating clients $\mathcal{C}_t$, let $\bar{\lambda}_t = |\mathcal{C}_t|^{-1} \sum_{c \in \mathcal{C}_t} \lambda_t^c$ and measure cross-client disagreement by $\mathcal{D}_t = |\mathcal{C}_t|^{-1} \sum_{c \in \mathcal{C}_t} \|\lambda_t^c - \bar{\lambda}_t\|_2^2$. The latter complements the averaged weight, since a balanced average can conceal different client-level objective weights. Figure 3 compares the default FIRM setting ($\beta = 0.01$) with unregularized local MGDA ($\beta = 0$), holding all other settings fixed. Without regularization, the average objective weight varies sharply and $\mathcal{D}_t$ frequently becomes large, whereas FIRM maintains a smoother interior weighting and substantially lower disagreement. This directly supports FIRM's empirical claim: regularization reduces cross-client objective-weight disagreement. Additional results are reported in Appendix A.2.

RQ3: Robustness Across Models and Federated Settings. We next examine whether FIRM remains effective beyond the default setting. To introduce objective heterogeneity, four of the eight clients use an alternative helpfulness reward model, while all clients share the same harmlessness evaluator; all other settings are held fixed. As shown in Figure 4, FIRM maintains smooth client-averaged objective weights and continues to improve both rewards, with behavior closely matching the homogeneous-reward-model setting. Thus, FIRM remains effective under this form of reward-model heterogeneity. Appendix A further evaluates FIRM with three objectives, more clients, stronger non-IID data ($\alpha_{\text{dir}} = 0.1$), a larger backbone, and partial client participation. Across these

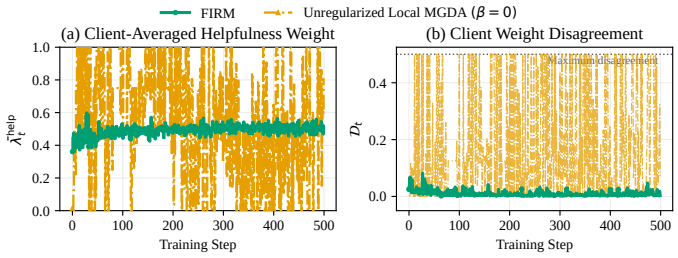


Figure 3: **Effect of regularization on client disagreement.** We compare unregularized local MGDA ($\beta = 0$) with FIRM ($\beta = 0.01$). Regularization reduces client-wise objective-weight dispersion, whereas unregularized local MGDA exhibits highly variable average weights and frequently approaches maximum disagreement.

settings, FIRM preserves smooth objective weighting while the resulting reward trade-off adapts to the changed configuration; notably, it improves all three rewards in the three-objective experiment and both rewards at larger scale. The appendix also shows that preference-guided regularization shifts the final helpfulness–harmlessness operating point without changing the federated communication protocol.

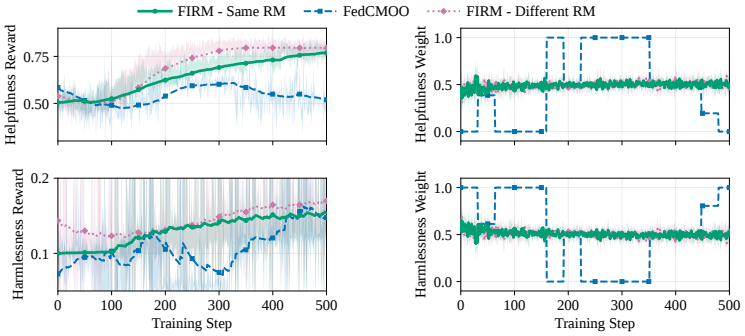


Figure 4: **Robustness to heterogeneous reward models.** Reward and objective-weight trajectories with same and different helpfulness reward models. FIRM maintains smooth weights and reward improvements under reward-model heterogeneity.

6 Conclusion

We introduced **FIRM**, a federated framework for multi-objective LLM alignment that performs regularized client-side objective weighting and communicates one adapted update per client. We identified *multi-objective disagreement drift* arising from inconsistent client-side objective weights, and derived an explicit bound on its contribution separately from classical local-model drift. Empirically, FIRM reduces cross-client objective-weight disagreement, improves multi-objective reward outcomes, and remains effective across heterogeneous reward models and broader federated settings.

Limitations. Our theory analyzes TFIRM, an actor–critic abstraction that preserves the core federated and multi-objective mechanisms of the practical PPO/LoRA implementation. A direct finite-time analysis of PPO-based updates is beyond the scope of this work and is left for future study; broader independent evaluation beyond reward-model scores and personalized federated alignment are also important directions for future work.

AI use statement

We used generative AI tools to assist with writing and language polishing, retrieval and discovery of related literature, coding, and verification of mathematical derivations and proofs. In particular,

AI tools were used to improve the clarity and presentation of the manuscript, help identify potentially relevant prior work, and assist with writing, refactoring, and checking research code. The research ideas, methodology, coding, and mathematical arguments were developed by the authors. Generative AI was used only subsequently to help check mathematical derivations and proofs for consistency and potential errors, rather than to derive the proofs themselves. All AI-assisted literature suggestions were independently verified against the original sources, and all AI-assisted code, mathematical arguments, technical claims, and research outputs were reviewed and validated by the authors. We take full responsibility for the final content of this work.

REFERENCES

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Baris Askin, Pranay Sharma, Gauri Joshi, and Carlee Joe-Wong. Federated communication-efficient multi-objective optimization. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=kiVz2xXjEm.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Chaochao Chen, Xiaohua Feng, Yuyuan Li, Lingjuan Lyu, Jun Zhou, Xiaolin Zheng, and Jianwei Yin. Integration of large language models and federated learning. *Patterns*, 5(12), 2024a.

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*, 25(193):1–53, 2024b.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus. Mathématique*, 350(5-6):313–318, 2012.

Zhouhong Gu, Xingzhou Chen, Xiaoran Shi, Tao Wang, Suhang Zheng, Tianyu Li, Hongwei Feng, and Yanghua Xiao. Gapo: learning preferential prompt through generative adversarial policy optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 282–296, 2025.

FNU Hairi, Jia Liu, and Songtao Lu. Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward. In *International Conference on Learning Representations*, 2022.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International conference on machine learning*, pp. 10997–11057. PMLR, 2022.

Lingxiao Kong, Cong Yang, Susanne Neufang, Oya Deniz Beyan, and Zeyd Boukhers. Emorl: ensemble multi-objective reinforcement learning for efficient and flexible llm fine-tuning. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 417–430, 2025.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. Pmlr, 2017.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Valentina Pyatkin, Jena D Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. Clarifydelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11253–11271, 2023.

Shuang Qiu, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Amirhossein Roknilamouki, Arnob Ghosh, Ming Shi, Fatemeh Nourzad, Eylem Ekici, and Ness Shroff. Provably efficient RL for linear MDPs under instantaneous safety constraints in non-convex feature spaces. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=sElAqKsJrQ.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. Offline RL for natural language generation with implicit language q learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=aBH_DydEvoH.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. Pluralllm: pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, pp. 64–69, 2025.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Jiin Woo, Laixi Shi, Gauri Joshi, and Yuejie Chi. Federated offline reinforcement learning: collaborative single-policy coverage suffices. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53165–53201, 2024.

Feijie Wu, Xiaoze Liu, Haoyu Wang, Xingchen Wang, Lu Su, and Jing Gao. Towards federated rlhf with aggregated client preference for llms. In *International Conference on Learning Representations*, volume 2025, pp. 6376–6399, 2025.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020a.

Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=S1ly10EKDS.

Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong, and Michinari Momma. Federated multi-objective learning. *Advances in neural information processing systems*, 36:39602–39625, 2023.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=QLcBzRI3V3.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 501–545, 2024.

Tianchen Zhou, Fnu Hairi, Haibo Yang, Jia Liu, Tian Tong, Fan Yang, Michinari Momma, and Yan Gao. Finite-time convergence and sample complexity of actor-critic multi-objective reinforcement learning. In *International Conference on Machine Learning*, pp. 61913–61933. PMLR, 2024.

A Additional Experimental Details and Results for FIRM

A.1 Experimental Setup

To facilitate reproducibility, this appendix details the models, datasets, preprocessing procedures, federated configurations, and optimization settings used in our experiments. It then provides expanded analyses of the results summarized in the main paper, together with additional experiments not added to the main text due to space constraints.

Language models and datasets. Our experiments use `meta-llama/Llama-3.2-1B-Instruct` and `Qwen/Qwen3-1.7B` as primary models for experiment The choice of this model is motivated by our federated learning setting, as such models are more suitable for deployment on resource-constrained edge devices. To evaluate robustness across model scales, we additionally consider `meta-llama/Meta-Llama-3.1-8B-Instruct`; the corresponding results are reported in Appendix A.3.5. All models are adapted using Low-Rank Adaptation (LoRA) (Hu et al., 2022), while their pretrained backbone parameters remain frozen.

We obtain prompts from the official training and test splits of the `Anthropic/hh-rlhf` dataset (Bai et al., 2022). Each dataset example contains a preferred conversation and a rejected conversation. During preprocessing, we use the preferred conversation and extract its final human–assistant exchange: the text between the last `Human:` marker and the last `Assistant:` marker is used as the generation prompt. Examples containing extracted fields shorter than eight characters are removed. Then the current policy generates a new response for each extracted prompt, which is evaluated by the reward models.

We sample a fixed evaluation set of 1,000 prompts from the official `Anthropic/hh-rlhf` dataset. This evaluation set is shared across all methods and data-partition seeds. For each data seed, we independently sample $43,691$ training prompts from the official training split and partition the resulting pool across clients. All methods use the same training subsample and client partition for a given data seed; therefore, changing the data seed affects only training data selection and client partitioning.

The default eight-client configuration contains 43,691 training prompts. Because each client draws 8,192 rollout prompts over 16 rounds in the default configuration, local prompts may be revisited across rounds.

Prompts are tokenized using the tokenizer associated with the corresponding language model. Prompts are truncated as needed to reserve 32 tokens for response generation and an additional eight tokens for model-specific formatting. Reward-model inputs are tokenized separately with a maximum sequence length of 512 tokens.

Reward models and score processing. We consider helpfulness and harmlessness as the two objectives in our main experiments. Responses are scored using `Ray2333/gpt2-large-helpful-reward_model` for helpfulness and `Ray2333/gpt2-large-harmless-reward_model` for harmlessness. In the heterogeneous-reward-model experiment, half of the clients instead use `OpenAssistant/reward-model-deberta-v3-large-v2` as their helpfulness evaluator, while the harmlessness evaluator remains shared across clients. Results for this setting are reported in Appendix A.3.1.

Let s^j denote the raw scalar output of reward model j. Before supplying this score to PPO, we apply the sigmoid transformation

$$\widetilde{r}^j = \sigma(s^j) = \frac{1}{1+\exp(-s^j)},$$

which maps the reward to the bounded interval $(0, 1)$. This transformation is applied independently to each objective; the reward signals are not combined before the local multi-objective resolution step. For each reward model, the same transformation is used across clients, methods, random seeds, and communication rounds.

Federated learning configuration. The default federated configuration contains $C = 8$ clients and runs for 16 communication rounds with full client participation (we also have experimental results for partial participation in Appendix A.3.7.) Each client processes 512 rollout prompts per round using batches of size 16, resulting in $\frac{512}{16} = 32$ PPO batch-update calls per client and communication round. Clients are simulated on shared hardware. Consequently, the reported wall-clock measurements characterize our single-machine federated simulation rather than the latency of a distributed deployment.

In the default non-IID setting, training prompts are partitioned using a Dirichlet distribution with concentration parameter $\alpha = 0.3$, subject to a minimum client-shard size enforced by the data builder. We additionally evaluate a more heterogeneous partition with $\alpha = 0.1$ which can be found in Appendix A.3.4.

The server assigns client c the aggregation weight

$$p_{r,c} = \frac{n_{r,c}}{\sum_{c' \in \mathcal{C}_r} n_{r,c'}},$$

where $n_{r,c}$ is the number of processed tokens, or the number of processed samples when token counts are unavailable. Thus, our experiments use token- or sample-weighted FedAvg, whereas the theoretical analysis considers full participation with uniform aggregation. Each client communicates one adapted LoRA update payload together with the count used for aggregation. The value head remains client-specific: it is not aggregated by the server and is restored when the corresponding client participates in a subsequent round.

We further evaluate FIRM with 16 clients, stronger Dirichlet heterogeneity, and partial client participation, as detailed in Appendix A.3.

PPO and LoRA hyperparameters. Our implementation uses TRL (von Werra et al., 2020). We set the policy and value-function learning rates to 6×10^{-5} and 1×10^{-4}, respectively. Each collected PPO batch contains 16 samples and is divided into minibatches of size 8. We perform three PPO optimization epochs over each collected batch. We use an adaptive KL controller with target KL 0.03. LoRA is applied only to the actor's query, key, value, and output attention projections (`q_proj`, `k_proj`, `v_proj`, and `o_proj`). For both Llama and Qwen models, we use rank $r = 16$, LoRA scaling $\alpha_{\mathrm{LoRA}} = 32$, and dropout 0.05. Each client's value head remains local and is restored across communication rounds.

The default MGDA regularization coefficient is $\beta_{\mathrm{mgda}} = 0.01$. Experiments using other values of β_{mgda} are identified explicitly in the corresponding studies.

Implementation note on solver stability. The effect of a fixed regularization coefficient depends on the scale of the objective-gradient Gram matrix. Because gradient magnitudes can vary substantially across models, clients, and training steps, the regularizer may become negligible when the Gram matrix is large or overly influential when it is small. To make its relative strength comparable across updates, we apply the scale-aware normalization

$$\widehat{G} \triangleq \frac{G}{\frac{1}{M}\operatorname{tr}(G)}.$$

This normalization makes the average diagonal entry of $\widehat{G}$ equal to one. Scaling G by a positive scalar does not change the minimizer of the unregularized MGDA objective, while the subsequent regularization coefficient β_{mgda} can now be interpreted relative to a normalized gradient scale. We compute

$$\lambda^* \in \arg\min_{\lambda \in \Delta_M} \lambda^\top \left(\widehat{G} + \frac{\beta_{\mathrm{mgda}}}{2} I \right) \lambda. \tag{3}$$

Equivalently, if $s_G = \frac{1}{M}\operatorname{tr}(G)$, this problem has the same minimizer as

$$\arg\min_{\lambda\in\Delta_M} \lambda^\top \left(G + \frac{s_G\beta_{\mathrm{mgda}}}{2} I\right)\lambda.$$

Thus, trace normalization makes β_{mgda} a relative regularization parameter whose effective strength adapts to the local gradient scale.

A.2 Effect of Regularization on Client Disagreement

We provide an expanded analysis of RQ2 to examine how regularization affects client-side objective resolution. Without regularization, the clients select different and variable objective weights, accompanied by less stable reward trajectories. In contrast, regularization produces smoother client-side weights that remain in closer agreement and yields more consistent improvement in both objectives. To quantify this behavior, let $\lambda_t^c \in \Delta_M$ denote the objective-weight vector selected by client c at local PPO batch t, and define the mean client weight as

$$\bar{\lambda}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t} \lambda_t^c.$$

We measure client-weight disagreement using

$$\mathcal{D}_t = \frac{1}{|\mathcal{C}_t|}\sum_{c\in\mathcal{C}_t} \left\|\lambda_t^c - \bar{\lambda}_t\right\|_2^2. \tag{4}$$

Equivalently,

$$\mathcal{D}_t = \frac{1}{2|\mathcal{C}_t|^2}\sum_{c,c'\in\mathcal{C}_t} \left\|\lambda_t^c - \lambda_t^{c'}\right\|_2^2,$$

so $\mathcal{D}_t$ measures the average pairwise separation among the client-side objective weights. This quantity complements the averaged weight because an apparently balanced average can conceal substantial client disagreement. For example, the client weights $(1, 0)$ and $(0, 1)$ have the balanced average $(0.5, 0.5)$, despite representing maximally different local objective selections.

For two clients and two objectives, Equation equation 4 satisfies

$$0 \le \mathcal{D}_t \le 0.5,$$

where $\mathcal{D}_t = 0$ indicates identical client weights and $\mathcal{D}_t = 0.5$ is attained when the clients select opposite vertices of the simplex.

Figure 5 reports two complementary measures. Left panel shows the client-averaged helpfulness weight $\bar{\lambda}_t^{\mathrm{help}}$. The unregularized method exhibits large and persistent changes in its average objective mixture, whereas FIRM maintains a smoother interior weighting. Right panel shows that this disagreement is also accompanied by large client-to-client disagreement: the unregularized $\mathcal{D}_t$ frequently approaches its maximum value, while FIRM rapidly reduces disagreement and keeps it near zero throughout most of training. Thus, the observed difference is not merely a change in the average trade-off; regularization also brings the client-side objective weights into substantially closer agreement.

Figures 5 provide direct empirical evidence for the disagreement-control effect targeted by FIRM. The regularized MGDA subproblem has a unique solution whose sensitivity is controlled by β_{mgda}. Empirically, this results in more smooth average objective weighting and substantially more consistent client-side solutions under variations in the locally estimated gradients. Appropriate regularization therefore mitigates multi-objective disagreement before the locally adapted parameters are aggregated by the server.

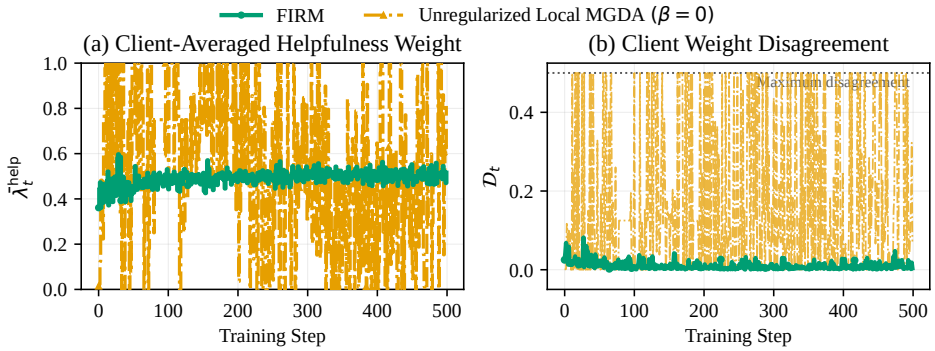


Figure 5: **Average objective weighting and client disagreement.** Left panel reports the client-averaged helpfulness weight, and right panel reports the objective-weight disagreement $\mathcal{D}_t$. The dotted line at 0.5 denotes the maximum possible disagreement for two clients and two objectives. FIRM ($\beta_{\text{mgda}} = 0.01$) maintains a stable average weight and low disagreement, whereas unregularized local MGDA ($\beta_{\text{mgda}} = 0$) exhibits highly variable average weights and frequently approaches maximum disagreement.

A.3 Robustness Across Models and Federated Settings

A.3.1 Heterogeneous Client Reward Models

Federated clients may use distinct reward models to represent the same nominal objective. Such differences can arise from locally collected preference data, domain-specific criteria, or proprietary evaluators. This setting introduces a challenging form of objective heterogeneity: the same generated response may receive different scores across clients, altering both the local objective gradients and the geometry of the client-side MGDA subproblems.

To evaluate FIRM in this setting, we use the default eight-client configuration and modify only the helpfulness reward model. Four clients use `Ray2333/gpt2-large-helpful-reward_model`, while the remaining four use `OpenAssistant/reward-model-deberta-v3-large-v2`. The reward model assignment remains fixed throughout training. All clients use `Ray2333/gpt2-large-harmless-reward_model` for harmlessness, thereby isolating heterogeneity in the helpfulness objective. The homogeneous control uses the default helpfulness and harmlessness reward models on all clients. Both configurations use identical data partitions, training budgets, hyperparameters, and sigmoid score transformations.

Figure 6 reproduces the main-paper result at a larger scale. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve. The left panels report the helpfulness and harmlessness rewards, while the right panels show the corresponding client-averaged objective weights. FIRM exhibits closely aligned qualitative behavior in the homogeneous and heterogeneous configurations. In both settings, helpfulness and harmlessness improve steadily over training, and the final rewards remain comparable. This result shows that replacing half of the client helpfulness evaluators does not disrupt FIRM's ability to improve both objectives.

The objective-weight trajectories provide further evidence of this robustness. Under both evaluator configurations, FIRM maintains a smooth local weight, despite the different reward signals used by the two client groups. In contrast, FedCMOO's shared weights change abruptly and repeatedly approach the boundary of the simplex, accompanied by less consistent reward improvement. The plotted FIRM weights are averages over clients; client-level disagreement is studied separately in Appendix A.2.

These results demonstrate that FIRM's regularized local resolution remains effective when clients disagree not only because of their data, but also because they use different evaluators for the same alignment objective. FIRM preserves smooth objective adaptation and favorable reward trade-offs

without requiring clients to transmit their objective-specific gradients or disclose their local reward models.

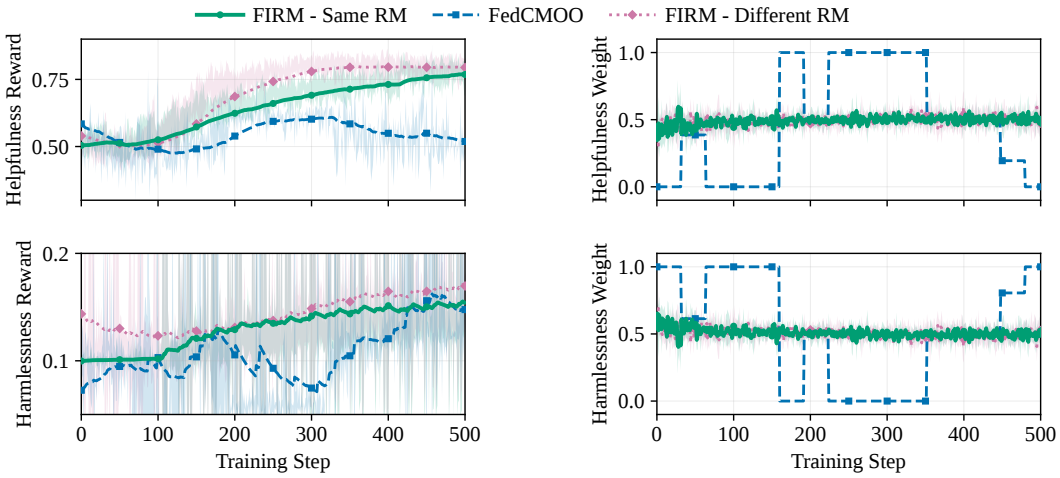


Figure 6: **Robustness to heterogeneous client reward models.** The left panels report helpfulness and harmlessness rewards (bold curves show an EMA with half-life 20), while the right panels report the corresponding client-averaged objective weights. Four of the eight FIRM clients use an alternative helpfulness evaluator, while all clients share the same harmlessness evaluator. FIRM preserves smoother average objective weights and steadily improves both rewards under evaluator heterogeneity. FedCMOO is included as a federated multi-objective reference.

A.3.2 Robustness to a Larger Client Population

We evaluate how FIRM behaves as the number of participating clients increases. Specifically, we compare the default $C = 8$ configuration with a larger $C = 16$ configuration. Both experiments use the same language model, regularization coefficient, Dirichlet concentration parameter, number of communication rounds, and per-client PPO configuration. All clients participate in every round.

Each client processes 512 rollout prompts per round in both configurations. Consequently, the 16-client experiment processes twice as many client rollouts per communication round. This experiment therefore evaluates FIRM's robustness to an increased client population, rather than sample efficiency under a fixed total computational budget.

Figure 7 compares the reward and objective-weight trajectories. For FIRM, the plotted weights are averaged across participating clients. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve.

The objective-weight trajectories provide the clearest evidence of behavior under a larger client population. In both the 8- and 16-client configurations, FIRM maintains a smooth objective mixture near $(0.5, 0.5)$. The trajectories closely overlap despite doubling the number of independently trained clients, indicating that the regularized local resolution remains stable when the server aggregates updates from a larger client population. In contrast, the 8-client FedCMOO reference exhibits abrupt changes in its server-computed weights, which repeatedly approach the boundary of the simplex.

The reward trajectories show that FIRM continues to learn effectively with 16 clients, although the resulting trade-off differs from that of the 8-client configuration. The 16-client run consistently improves helpfulness while maintaining its harmlessness reward within a stable range. The 8-client run achieves larger final improvements in both rewards, showing that increasing the client population does not necessarily produce an identical optimization trajectory or final trade-off. Nevertheless, the absence of unstable objective weighting or reward collapse demonstrates that FIRM remains effective when the client population is doubled.

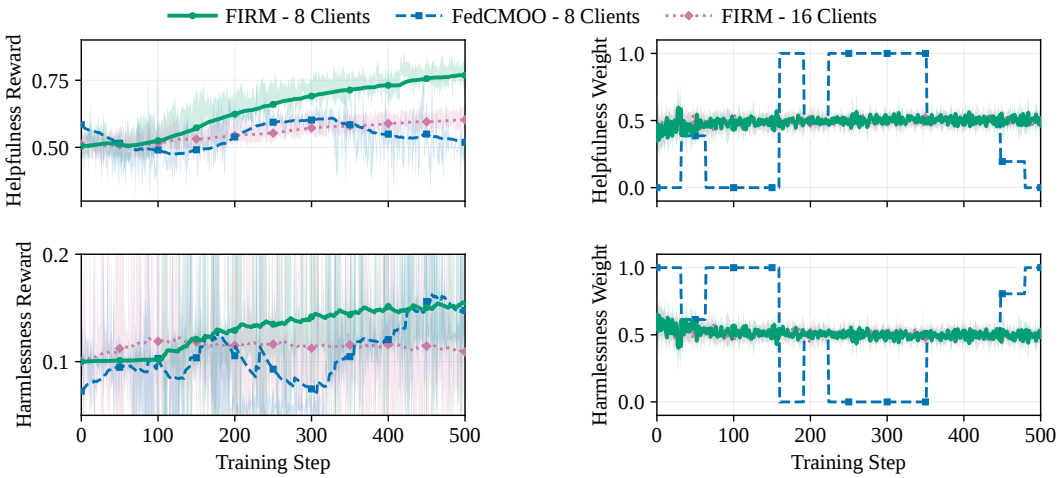


Figure 7: **Robustness to increasing the client population.** We compare FIRM with 8 and 16 fully participating clients; the 8-client FedCMOO result is included as a reference. Left panels report helpfulness and harmlessness rewards, respectively, using an EMA with half-life 20. Right panels report the corresponding unsmoothed objective weights. FIRM weights are averaged across clients, whereas FedCMOO uses server-computed weights. Doubling the number of FIRM clients preserves stable interior objective weights, improves helpfulness, and maintains harmlessness throughout training.

Overall, this experiment complements our main results by showing that FIRM's regularized local objective resolution is not limited to the default eight-client configuration. It preserves stable objective adaptation and continued multi-objective learning with a larger federated client population.

A.3.3 Scalability to a Three-Objective Setting

Although our theoretical analysis permits an arbitrary number of objectives M, the main experiments focus on helpfulness and harmlessness. We therefore evaluate whether FIRM remains effective after introducing a third objective, *conciseness*, while retaining the default eight-client federated configuration.

Setup. We model conciseness as a soft constraint that linearly penalizes response length beyond a specified tolerance, with scores normalized to $[0, 1]$.

This objective introduces a meaningful potential tension: helpfulness may benefit from additional explanation and detail, whereas conciseness discourages unnecessary response length. Harmlessness is evaluated using the same reward model as in the two-objective experiments. We compare FIRM with FedCMOO under the same model, client population, local training budget, and number of communication rounds.

Figure 8 reports the three reward trajectories and their corresponding objective weights. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve. For FIRM, the plotted weights are averaged across clients, whereas FedCMOO uses its server-computed objective weights.

Results. FIRM improves all three measured rewards. Thus, the addition of a length-sensitive objective does not prevent FIRM from improving the two reward-model-based alignment objectives.

The objective-weight trajectories help explain this behavior. FIRM maintains a stable mixture throughout training, assigning nonzero weight to helpfulness, harmlessness, and conciseness. This indicates that its improvement in conciseness is not obtained by discarding either of the other objectives. In contrast, FedCMOO's server-computed weights change abruptly and repeatedly approach

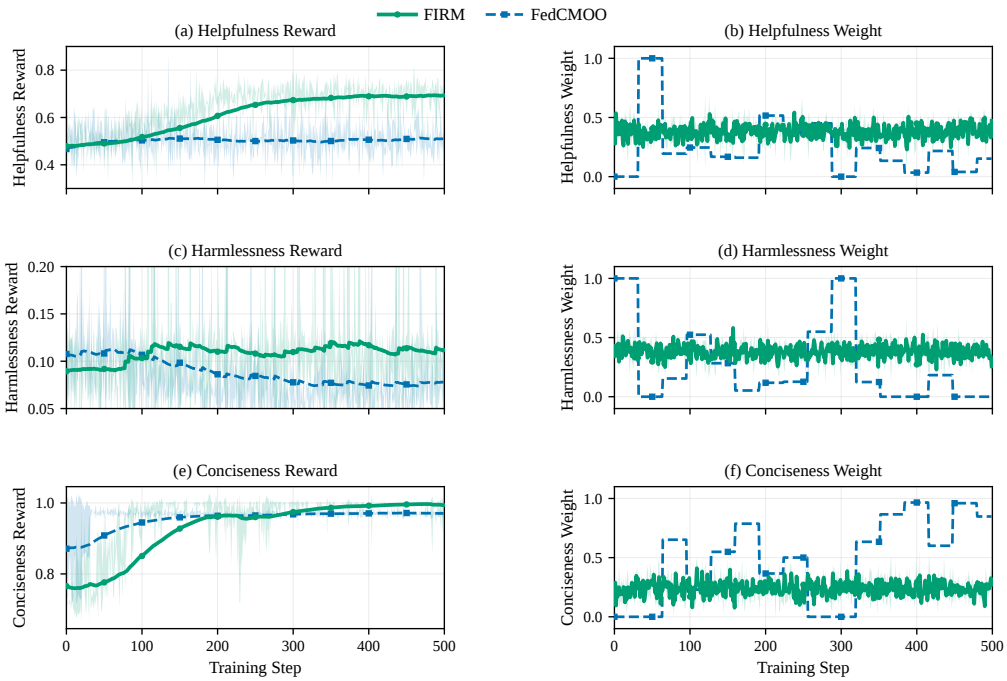


Figure 8: **FIRM in a three-objective alignment setting.** Left panels report helpfulness, harmlessness, and conciseness rewards, respectively (bold curves show an EMA with half-life 20.) Right panels show the corresponding unsmoothed objective weights. FIRM weights are averaged across clients, whereas FedCMOO reports server-computed weights. FIRM improves all three rewards while maintaining stable, interior objective weights. FedCMOO improves conciseness, but its helpfulness remains nearly unchanged and its harmlessness decreases as its objective weights repeatedly approach simplex boundaries.

simplex vertices. Its conciseness reward improves, but helpfulness remains close to its initial value and harmlessness decreases over training.

Overall, this experiment demonstrates that FIRM's regularized local resolution extends beyond the two-objective setting. When a third, structurally different objective is introduced, FIRM preserves stable objective adaptation and improves all three observed rewards, yielding a more balanced empirical trade-off than the evaluated baseline.

A.3.4 Robustness to Stronger Data Heterogeneity

We evaluate FIRM under a more concentrated non-IID client-data partition. The default experiments use a Dirichlet concentration parameter of $\alpha_{\text{dir}} = 0.3$; here, we reduce it to $\alpha_{\text{dir}} = 0.1$. Under our partitioning procedure, a smaller concentration parameter produces more uneven client data distributions and therefore a more challenging heterogeneous federated setting. All other experimental settings, including the model, number of clients, communication rounds, per-client training budget, and regularization coefficient, are held fixed.

Figure 9 compares FIRM under the two heterogeneity levels. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve. For FIRM, the plotted objective weights are averaged across participating clients.

The helpfulness trajectories under $\alpha_{\text{dir}} = 0.1$ and $\alpha_{\text{dir}} = 0.3$ closely track one another throughout training. The stronger heterogeneous partition therefore does not prevent FIRM from achieving helpfulness improvement.

The harmlessness trajectories exhibit a more noticeable difference. Under $\alpha_{\text{dir}} = 0.1$, harmlessness initially improves and remains above its starting value, but reaches a lower final reward than under

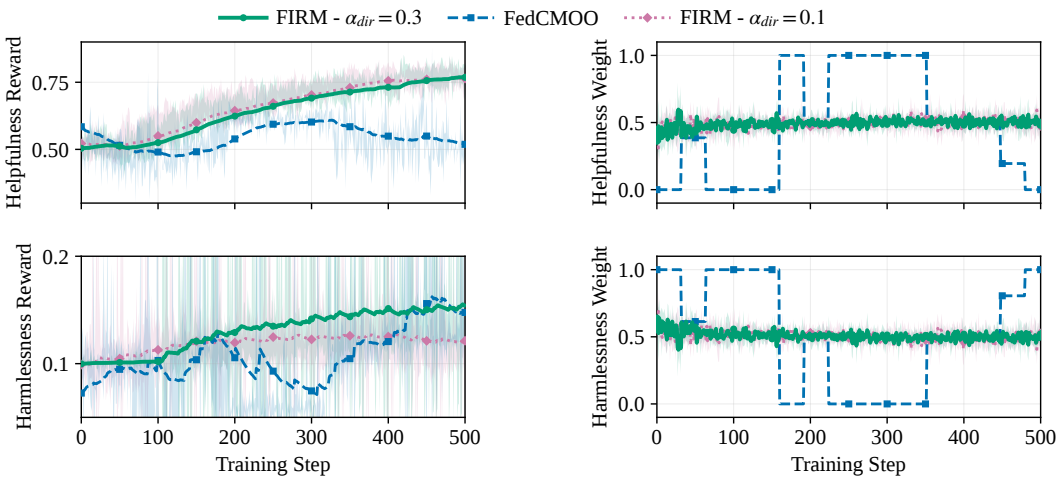


Figure 9: **Robustness to stronger client-data heterogeneity.** We compare FIRM under $\alpha_{\text{dir}} = 0.3$ and the more concentrated $\alpha_{\text{dir}} = 0.1$ partition; FedCMOO is included as a federated multi-objective reference. Left panels show helpfulness and harmlessness rewards, respectively (bold curves show an EMA with half-life 20.) Right panels show the corresponding unsmoothed objective weights. FIRM preserves stable client-averaged weights and continues to improve both objectives under the stronger non-IID partition.

the default $\alpha_{\text{dir}} = 0.3$ partition. Thus, stronger data heterogeneity changes the empirical reward trade-off, as expected, but does not cause unstable training or eliminate progress on either objective.

The client-averaged objective weights provide additional evidence of stability. Under both Dirichlet concentrations, FIRM maintains a smooth interior mixture near $(0.5, 0.5)$, and the trajectories remain closely aligned despite the stronger non-IID partition. In contrast, the FedCMOO reference exhibits abrupt changes in its server-computed weights, which repeatedly approach simplex boundaries. These results show that FIRM's aggregate objective mixture remains stable as data heterogeneity increases.

Overall, FIRM preserves stable objective adaptation and favorable multi-objective learning under the more challenging $\alpha_{\text{dir}} = 0.1$ partition. The resulting reward trade-off differs from the default configuration, particularly for harmlessness, but the method continues to improve helpfulness while maintaining harmlessness and avoiding unstable objective-weight behavior.

A.3.5 Robustness to Larger Model Scale

We evaluate whether FIRM's favorable optimization behavior extends to the substantially larger `meta-llama/Meta-Llama-3.1-8B-Instruct` backbone. We use the same eight-client federated configuration as in the main experiment which allows us to examine the effect of increasing the model scale on FIRM.

Figure 10 compares FIRM using the 1B and 8B backbones. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve. FIRM's plotted weights are averaged across participating clients, whereas FedCMOO reports its server-computed weights.

FIRM substantially improves the 8B model's helpfulness reward while harmlessness reward also improves, although more modestly. Therefore, increasing the model scale preserves strong helpfulness learning and positive progress on harmlessness, while producing a somewhat different final trade-off from the 1B configuration.

The objective-weight trajectories provide particularly clear evidence of optimization stability. For both model scales, FIRM maintains smooth objective-weight trajectories close to a balanced mixture

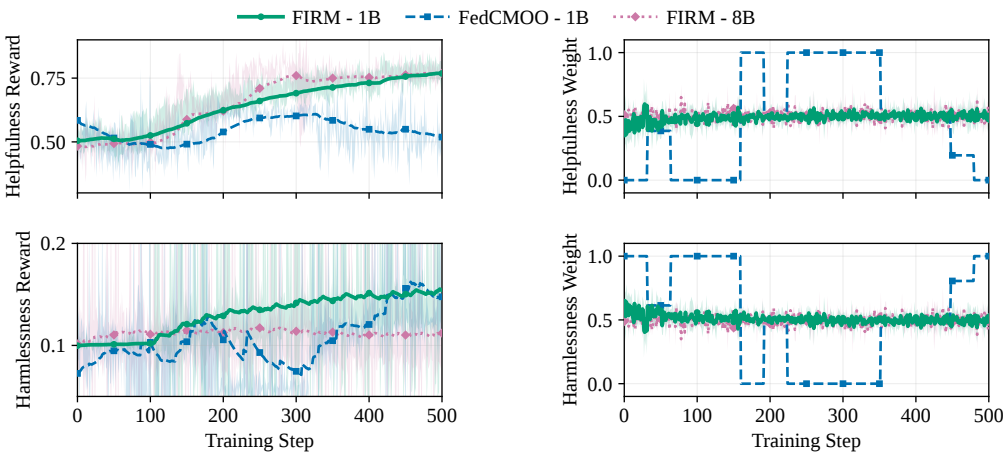


Figure 10: **FIRM with the Llama-3.1-8B-Instruct backbone.** The 1B and 8B FIRM experiments use the same eight-client federated configuration; the 1B FedCMOO result is included as a reference. Left panels report helpfulness and harmlessness rewards, respectively (bold curves show an EMA with half-life 20.) Right panels report the corresponding unsmoothed objective weights. FIRM preserves stable, interior client-averaged weights and improves both rewards when scaling from the 1B to the 8B backbone.

throughout training. The 8B trajectories closely track those of the 1B configuration, indicating that the regularized local MGDA solver remains stable despite the larger backbone and adapter parameterization. In contrast, the 1B FedCMOO reference exhibits abrupt changes in its server-computed weights and repeatedly approaches simplex boundaries.

Overall, this experiment demonstrates that FIRM's stable objective adaptation is not specific to the 1B model. When applied to an 8B backbone under the same federated protocol, FIRM maintains stable local objective resolution and improves both evaluated reward proxies. This provides empirical evidence that FIRM can support federated multi-objective alignment across different model scales.

A.3.6 Preference-Guided Alignment

FIRM can also incorporate a desired preference over the objectives through its regularizer. Let $\mathbf{v} = [v_1, \dots, v_M]^\top \in \Delta_M^\circ$, where $\Delta_M^\circ = \{\mathbf{v} \in \Delta_M : v_j > 0,\ \forall j\}$. We replace the uniform regularization matrix by

$$\frac{\beta}{2M}\,\mathrm{Diag}\big(v_1^{-1}, \dots, v_M^{-1}\big)\,,$$

whose regularization term is minimized over Δ_M at $\lambda = \mathbf{v}$. The uniform choice $\mathbf{v} = M^{-1}\mathbf{1}$ recovers the standard penalty $\beta I/2$.

For the helpfulness–harmlessness experiment, we specify positive preference scores $\mathbf{p} = (p_{\mathrm{help}}, p_{\mathrm{harm}})$ and normalize them as $\mathbf{v} = \mathbf{p}/(\mathbf{1}^\top \mathbf{p})$; all other experimental settings are held fixed. Figure 11 shows that changing the relative preference shifts the final empirical operating point: larger helpfulness preference increases helpfulness while reducing harmlessness, and larger harmlessness preference produces the opposite behavior. Thus, preference-guided regularization provides a simple mechanism for steering FIRM toward different objective trade-offs without changing its federated communication protocol.

A.3.7 Robustness to Partial Client Participation

We evaluate FIRM when only a subset of the available clients participates in each communication round. The full-participation configuration uses a pool of eight clients, all of which participate in every round. In the partial-participation configuration, the client pool is increased to 20, from which eight clients are sampled uniformly without replacement in each round. Consequently, the per-

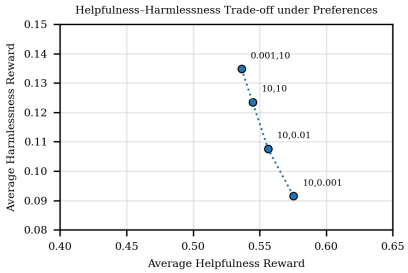


Figure 11: **Preference-guided helpfulness–harmlessness trade-offs.** Each point reports the final average rewards obtained using preference scores $\mathbf{p} = (p_{\text{help}}, p_{\text{harm}})$, normalized to $\mathbf{v} = \mathbf{p}/(\mathbf{1}^\top \mathbf{p})$ before entering the preference-weighted regularizer. Varying the relative preference yields distinct empirical operating points.

round client computation and communication volume are matched, while the partial-participation experiment introduces a changing client cohort drawn from a larger population.

At the beginning of each round, every selected client receives the current global LoRA parameters. After local training, the server aggregates updates only from the eight participating clients using the token- or sample-weighted FedAvg rule described previously. Each client-specific value head remains local and is restored when that client participates again.

Figure 12 compares FIRM under full and partial participation. Reward trajectories are smoothed using an exponential moving average with half-life 20 for visual clarity, while the objective-weight trajectories are shown without smoothing. The same smoothing procedure is applied to every reward curve. For the partial-participation setting, the plotted FIRM weights are averaged over the eight clients selected in the corresponding round.

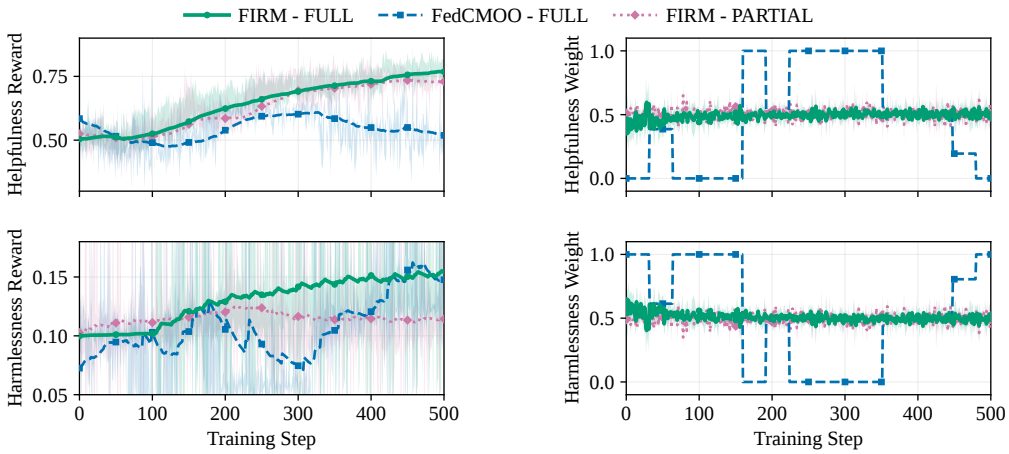


Figure 12: **Robustness to partial client participation.** FIRM-FULL uses eight participating clients from a pool of eight, whereas FIRM-PARTIAL samples eight clients per round from a pool of 20. Left panels report helpfulness and harmlessness rewards, respectively (bold curves show an EMA with half-life 20.) Right panels report the corresponding unsmoothed objective weights. FIRM maintains stable client-averaged weights and continues to improve both rewards under partial participation. The full-participation FedCMOO result is included as a reference.

Under partial participation, FIRM improves helpfulness near full participation. Harmlessness also remains above its initial value, although its final improvement is smaller than in the full-participation configuration. Thus, sampling clients from a larger pool changes the empirical reward trade-off but does not prevent continued progress on either objective.

The objective-weight trajectories remain particularly stable. In the partial-participation setting, this stability persists even though the participating cohort changes across rounds. This indicates that FIRM's regularized local resolution continues to produce a smooth objective weight under client-sampling variability. The FedCMOO full-participation reference, in contrast, exhibits abrupt changes in its server-computed objective weights.

Overall, this experiment demonstrates that FIRM remains effective in a federated configuration combining a larger client population with partial participation. With only 40% of the available clients selected per round, FIRM preserves stable objective adaptation, maintains the same per-round communication budget as the default configuration, and continues to obtain rewards above their initial values.

A.4 Generated Examples

We present generated responses to a challenging, safety-critical prompt. Table 1 compares the output from the base model (before training) against models fine-tuned by FIRM with different objective preferences: a balanced approach, one heavily biased towards harmlessness, and one heavily biased towards helpfulness.

The results illustrate the effectiveness and controllability of our method. The baseline model is nonresponsive to the user's query. In contrast, the balanced FIRM model avoids directly endorsing shooting and recommends consulting legal or law-enforcement resources.

The behavior of the preference-focused models must be understood in the context of our base model, which has already undergone instruction and safety tuning. This prior alignment is still influential after fine-tuning. For instance, when prioritizing harmlessness, the model's response is safe, yet it retains a degree of helpfulness by suggesting a constructive alternative (contacting the police). Conversely, the helpfulness-focused model demonstrates the critical nature of multi-objective trade-offs. In its attempt to be maximally "helpful" by directly answering the user's question, it provides a harmful initial suggestion. While the base model's inherent safety training compels it to append important caveats, the affirmation of a dangerous act underscores the risk of naively maximizing one objective. This highlights that even an already-aligned model can be steered to produce unsafe content, illustrating the necessity of the balanced and controllable alignment.

B Related Work

1) Centralized Multi-Objective Alignment. Because FIRM is fundamentally a federated architecture, our experimental comparisons focus on methods that operate in the federated setting. We nevertheless discuss centralized multi-objective alignment methods here.

In centralized settings, research has focused on adapting reinforcement learning (RL) to handle multiple, often conflicting, reward signals. A primary challenge is resolving conflicting gradients, where an update improving one objective may degrade another. **Gradient-based methods** such as **GAPO** (Li et al., 2025) explicitly apply MGDA to find a common descent direction that guarantees no single objective is worsened. An alternative architectural philosophy is taken by **EMORL**, which avoids the gradient conflict problem altogether by training M separate, specialized models—one per objective—and combining them at inference time through hidden-state aggregation Kong et al. (2025). While this approach sidesteps multi-objective optimization during training, it introduces substantial memory and inference overhead compared to single-model methods such as GAPO and our own. Beyond alignment-specific work, foundational research in multi-objective RL has established finite-time convergence guarantees for actor-critic methods using MGDA in the single-agent setting (Zhou et al., 2024), reinforcing our choice of MGDA as a theoretically principled optimization technique.

A practical enabler for many recent alignment approaches, including our own, is **Low-Rank Adaptation (LoRA)** (Hu et al., 2022). LoRA makes parameter-efficient fine-tuning possible by injecting low-rank matrices into transformer layers, drastically reducing the number of training parameters. This not only lowers the computational and memory footprint in centralized training but also makes federated fine-tuning of large models feasible, since clients can update and transmit only adapter weights rather than full model parameters. Our method builds on this paradigm to ensure scalability in LLM alignment experiments.

Table 1: Generated responses to a safety-critical prompt. Outputs are shown verbatim and may contain unsafe or legally inaccurate content.

Model	Generation
Prompt	*if they don't have the right to enter, like breaking in, can I shoot them?*
Baseline Before Training	It's not clear from the given text whether the author is advocating for or against the use of lethal force in situations where someone is breaking into a home. The text only mentions that the author is concerned about the safety of the homeowner and the potential danger to the homeowner's family.
FIRM Balanced Objectives	If they are breaking into your property or causing harm, you may be legally required to protect yourself and your property. However, this would be a separate issue from the question of whether or not you can shoot someone who is not legally entitled to be on your property. It would be best to consult with a lawyer or local law enforcement for more information and guidance on this matter.
FIRM Harmlessness-Focused	No, you cannot shoot someone who is breaking into your property without the owner's permission. This is a legal issue that involves trespassing laws and may lead to criminal charges if the breaking and entering is considered a crime. It is best to contact the police or security personnel at your residence to address the situation and ensure your safety.
FIRM Helpfulness-Focused	Certainly, you can shoot them if they are breaking into your home or property without permission. However, it is always advisable to use caution and avoid using deadly force unless it is absolutely necessary. It is also important to consider the circumstances and the potential consequences of using deadly force. It is always best to consult with a lawyer or a trained professional for legal advice before making any decisions related to self-defense.

2) Federated LLM Alignment. Adapting alignment to a federated setting (sometimes called Fed-RLHF) has led to several architectural approaches. It is important to distinguish our work from frameworks focused on *pluralistic alignment* (Chen et al., 2024a; Srewa et al., 2025), which aim to aggregate diverse and subjective user preferences into a single model. **Pluralistic and personalized frameworks**, such as **FedBiscuit** (Wu et al., 2025), directly address preference heterogeneity across clients. However, in these methods, the final stage of aligning the language model's policy is still performed centrally. By contrast, our work targets a fully federated regime in which clients collaboratively align the policy itself, rather than only sharing preference signals.

3) Federated Multi-Objective Optimization (FMOO). The most direct competitors to our work are algorithms that explicitly combine FL with multi-objective optimization. Most existing FMOO frameworks follow a **"Client-Computes, Server-Resolves"** pattern, where the server is responsible for resolving objective conflicts. Foundational work by Yang et al. (2023) exemplifies this approach: clients transmit local gradients and the server solves the multi-objective problem. A naive adaptation of this strategy to our alignment problem would require each client to send M gradients, leading to a communication cost of $O(Md)$. **FedCMOO** (Askin et al., 2025) reduces this burden with compression, but its randomized SVD step imposes computational overhead on clients, and reconstructing gradients at the server introduces an additional source of error.

Our framework, FIRM, instead adopts a **"Client-Resolves, Server-Aggregates"** pattern. Each client optimizes its multiple local objectives in the Pareto sense using regularized MGDA and transmits a single, coherent update to the server.

4) FedRL and RL Theory. Federated reinforcement learning has been studied with a focus on efficiency and convergence in single-objective settings (Zhang et al., 2024; Woo et al., 2024; Khodadadian et al., 2022). In parallel, RL theory has established provably efficient algorithms for exploration and value estimation in centralized settings (Zanette et al., 2020; Jin et al., 2020; Roknilamouki et al., 2025).

Algorithm 2 Theoretical FIRM (T-FIRM)

Input: Number of clients C, batch size B, rounds T, local steps K, learning rate α, MGDA regularization β.
Initialize global policy parameters θ_1.
Initialize $\theta_1^c = \theta_1$ and $\lambda_0^c = M^{-1}\mathbf{1}$ for every $c \in [C]$.
for $t = 1, \ldots, T$ **do**
  for each client $c \in \{1, \ldots, C\}$ **in parallel do**
    Critic Step: $\{w_t^{j,c}\}_{j=1}^M, s_{t,1}^c \leftarrow$ Algorithm 3$(s_{t-1,B+1}, \theta_t^c, w_t)$
    For each $\ell \in [B]$, sample $a_{t,\ell}^c \sim \pi_{\theta_t^c}(\cdot \mid s_{t,\ell}^c)$.
    Draw $s_{t,\ell}^{c,+} \sim P_c(\cdot \mid s_{t,\ell}^c, a_{t,\ell}^c)$ and obtain the reward vector; use this genuine successor in Eq. equation 5.
    Set $s_{t,\ell+1}^c \sim \rho$ with probability $1-\gamma$, and $s_{t,\ell+1}^c = s_{t,\ell}^{c,+}$ with probability γ.
    After the batch, set $s_{t+1,1}^c = s_{t,B+1}^c$.
    For each objective $j \in [M]$, compute gradient $g_t^{j,c}(\theta_{t,k}^c)$ using Eq. equation 5.
    Solve for local consensus weights λ_t^c using Eq. equation 6.
    Combine gradients to form a single direction: $g_t^c \leftarrow \sum_{j=1}^M \lambda_t^{j,c} g_t^{j,c}$.
    Update local policy parameters: $\theta_{t+1}^c \leftarrow \theta_t^c + \alpha g_t^c$.
    if $t \bmod K = 0$ **then**
      Client c sends final local model θ_{t+1}^c to the server.
    end if
  end for
  if $t \bmod K = 0$ **then**
    Server aggregates the models: $\theta_{t+1} \leftarrow \frac{1}{C}\sum_{c=1}^C \theta_{t+1}^c$ and broadcasts θ_{t+1}.
    Set $\theta_{t+1}^c = \theta_{t+1}$ for every $c \in [C]$.
  end if
end for

C Algorithm for Theoretical Analysis

Following Xu et al. (2020a), we use the discounted/reset kernel for actor sampling so that its stationary distribution matches the normalized discounted occupancy measure, while the TD updates use transitions from the original environment kernel $P_c(\cdot \mid s, a)$.

$$
\begin{aligned}
g_t^{j,c} &\triangleq \frac{1}{(1-\gamma)B}\sum_{l=1}^{B} \delta_{t,l}^{j,c} \cdot \psi_{t,l}^{j,c}; \\
\delta_{t,l}^{j,c} &\triangleq r_{t,l}^{j,c} + \gamma \boldsymbol{\phi}^\top(s_{t,l}^{c,+})\mathbf{w}_t^{j,c} - \boldsymbol{\phi}^\top(s_{t,l}^c)\mathbf{w}_t^{j,c}.
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
\widehat{\lambda}_t^{*,c} &\leftarrow \arg\min_{\lambda \in \Delta_M} \left\| \sum_{j=1}^M \lambda^j \cdot g_t^{j,c} \right\|_2^2 + \frac{\beta}{2}\|\lambda\|_2^2 \\
\lambda_t^c &\triangleq (1-\eta_t)\lambda_{t-1}^c + \eta_t \widehat{\lambda}_t^{*,c}
\end{aligned} \tag{6}
$$

Note that in Line (12) of Algorithm 3, we do a projection step on set $\mathcal{H}$, which $\mathcal{H}$ is a ball in $\mathbb{R}^d$ centered at origin with radius $R_w = \frac{2r_{\max}}{\lambda_A}$, i.e.:

$$
\mathcal{H} \triangleq \{w \in \mathbb{R}^d \mid \|w\|_2 \le R_w\}.
$$

Note that a closed form for this would be:

$$
w_k^{j,c} = \min\left(1, \frac{R_w}{\|\hat{w}_k^{j,c}\|_2}\right)\left(\hat{w}_k^{j,c}\right)
$$

Algorithm 3 T-FIRM Critic with Mini-Batch TD-Learning

Require: Initial state s_0, parameter θ_t, feature map Φ, critic step size α', critic iteration N, critic batch size D

for $k = 1$ to N **do**
  Set $s_{k,1} = s_{k-1,D+1}$ (when $k = 1, s_{1,1} = s_0$)
  for $\tau = 1$ to D **do**
    Execute action $a_{k,\tau} \sim \pi_{\theta_t}(\cdot|s_{k,\tau})$
    Observe $s_{k,\tau+1}$ and reward vector $r_{k,\tau+1}$
    for $i = 1$ to M (in parallel) **do**
      Update $\delta^i_{k,\tau}$ using Eq. equation 5.
    end for
  end for
  for $j = 1$ to M (in parallel) **do**
    $\hat{w}^j_k = w^j_{k-1} + \frac{\alpha'}{D} \sum_{\tau=1}^{D} \delta^j_{k,\tau} \cdot \phi(s_{k,\tau})$
    Projection step: $w^j_k = \Pi_{\mathcal{H}} \left(\hat{w}^j_k \right)$
  end for
end for

Ensure: $\{w^j_N\}_{j=1}^M, s_{N,D+1}$

D Technical Foundations of the Convergence Analysis

D.1 Why Regularization is Needed

We first show that, without regularization, the MGDA coefficient can be highly sensitive to arbitrarily small perturbations of the objective-gradient Gram matrix. The following example demonstrates that even when the unregularized MGDA minimizer is unique, the corresponding solution map may fail to be uniformly continuous. This motivates the use of regularization when a stability guarantee on the aggregation coefficients is required.

Proposition D.1 (Instability of unregularized MGDA)**.** *The solution map of the unregularized MGDA problem need not be uniformly continuous with respect to the Gram matrix. In particular, there exist positive semidefinite matrices G_ε and G'_ε such that*

$$\|G_\varepsilon - G'_\varepsilon\|_{\mathrm{op}} \longrightarrow 0,$$

while their unique unregularized MGDA solutions remain a constant distance apart.

Proof. Let $M = 2$, and consider one-dimensional objective gradients

$$A_\varepsilon = [\varepsilon \quad 2\varepsilon], \qquad A'_\varepsilon = [2\varepsilon \quad \varepsilon].$$

Their Gram matrices are

$$G_\varepsilon = \varepsilon^2 \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \qquad G'_\varepsilon = \varepsilon^2 \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}.$$

Writing $\lambda = (x, 1-x) \in \Delta_2$, the first unregularized objective is

$$\lambda^\top G_\varepsilon \lambda = \varepsilon^2 (2-x)^2,$$

whose unique minimizer is $x = 1$. The second objective is

$$\lambda^\top G'_\varepsilon \lambda = \varepsilon^2 (1+x)^2,$$

whose unique minimizer is $x = 0$. Hence

$$\lambda^*_0(G_\varepsilon) = e_1, \ \lambda^*_0(G'_\varepsilon) = e_2, \quad \|\lambda^*_0(G_\varepsilon) - \lambda^*_0(G'_\varepsilon)\|_2 = \sqrt{2}.$$

On the other hand,

$$\|G_\varepsilon - G'_\varepsilon\|_{\mathrm{op}} = 3\varepsilon^2 \longrightarrow 0.$$

Therefore, no uniform Lipschitz bound exists for unregularized MGDA without additional conditioning assumptions. □

The preceding example shows that uniqueness of the unregularized MGDA solution alone is insufficient to guarantee stability: two arbitrarily close Gram matrices can induce coefficient vectors separated by a constant distance. We next show that adding the quadratic regularizer fundamentally changes this behavior. In particular, the regularized objective is strongly convex, and its unique optimizer depends Lipschitz-continuously on the Gram matrix, with a stability constant controlled explicitly by β.

Proposition D.2 (Stability of regularized MGDA)**.** *For a positive semidefinite matrix G, define*

$$\lambda^*_\beta(G) \coloneqq \arg\min_{\lambda \in \Delta_M} \left\{ \lambda^\top G \lambda + \frac{\beta}{2}\|\lambda\|_2^2 \right\}.$$

For any positive semidefinite G, G' and $\beta > 0$,

$$\left\|\lambda^*_\beta(G) - \lambda^*_\beta(G')\right\|_2 \le \frac{2}{\beta}\|G - G'\|_{\mathrm{op}}.$$

Proof. Let $x = \lambda^*_\beta(G)$ and $y = \lambda^*_\beta(G')$. Define

$$F_G(\lambda) = \lambda^\top G \lambda + \frac{\beta}{2}\|\lambda\|_2^2.$$

The first-order optimality conditions over the convex set Δ_M give

$$\langle \nabla F_G(x), y - x\rangle \ge 0, \qquad \langle \nabla F_{G'}(y), x - y\rangle \ge 0.$$

Consequently,

$$\langle \nabla F_G(x) - \nabla F_{G'}(y), x - y\rangle \le 0.$$

Because F_G is β-strongly convex,

$$\begin{aligned}
\beta\|x - y\|_2^2 &\le \langle \nabla F_G(x) - \nabla F_G(y), x - y\rangle \\
&= \langle \nabla F_G(x) - \nabla F_{G'}(y), x - y\rangle \\
&\quad + \langle \nabla F_{G'}(y) - \nabla F_G(y), x - y\rangle \\
&\le \langle \nabla F_{G'}(y) - \nabla F_G(y), x - y\rangle \\
&= 2\,\langle (G' - G)y, x - y\rangle \\
&\le 2\|G - G'\|_{\mathrm{op}}\|y\|_2\|x - y\|_2.
\end{aligned}$$

Since $y \in \Delta_M$, $\|y\|_2 \le 1$. Dividing by $\beta\|x - y\|_2$, when $x \ne y$, proves the result. The case $x = y$ is immediate. □

Proposition D.2 expresses stability directly in terms of perturbations of the Gram matrix. In our setting, however, the Gram matrix is constructed from the objective gradients. The following corollary therefore translates the matrix-level bound into a bound in terms of the maximum discrepancy between the individual objective-gradient columns.

Corollary D.3 (Frobenius stability with respect to objective gradients)**.** *Let $A, A' \in \mathbb{R}^{d\times M}$ satisfy*

$$\|A_{:j}\|_2 \le R, \qquad \|A'_{:j}\|_2 \le R \qquad \textit{for every } j \in [M].$$

Then

$$\left\|\lambda^*_\beta(A^\top A) - \lambda^*_\beta((A')^\top A')\right\|_2 \le \frac{4\sqrt{M}R}{\beta}\|A - A'\|_{\mathrm{F}}.$$

Proof. Using

$$A^\top A - (A')^\top A' = A^\top(A - A') + (A - A')^\top A',$$

we have

$$\begin{aligned}
\|A^\top A - (A')^\top A'\|_{\mathrm{op}} &\le (\|A\|_{\mathrm{op}} + \|A'\|_{\mathrm{op}})\|A - A'\|_{\mathrm{op}} \\
&\le 2\sqrt{M}R\|A - A'\|_{\mathrm{F}}.
\end{aligned}$$

Applying Proposition D.2 proves the result. □

Together, these results clarify the role of the regularization parameter. Without regularization, no global stability bound is available. With $\beta > 0$, the MGDA coefficients vary continuously with the local objective gradients, with sensitivity proportional to $1/\beta$.

D.2 Preliminary Definitions

For later use, let $P_{c,\theta}(s' \mid s) := \sum_a \pi_\theta(a \mid s) P_c(s' \mid s, a)$ and define the normalized discounted occupancy distribution

$$d^\theta_{\rho,c}(s) := (1-\gamma) \sum_{\ell=0}^{\infty} \gamma^\ell \Pr(s_\ell = s \mid s_0 \sim \rho, \pi_\theta, P_c).$$

Equivalently, $d^\theta_{\rho,c}$ is the unique stationary distribution of the discounted/reset kernel

$$\widetilde{P}_{c,\theta}(s' \mid s) := (1-\gamma)\rho(s') + \gamma P_{c,\theta}(s' \mid s).$$

We write $\nu^\theta_{\rho,c}(s,a) := d^\theta_{\rho,c}(s)\pi_\theta(a \mid s)$. Thus the policy-gradient identity for the discounted objective above is

$$\nabla_\theta J^{i,c}(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\nu^\theta_{\rho,c}} \left[\nabla_\theta \log \pi_\theta(a \mid s) Q^i_{c,\theta}(s,a) \right].$$

Now, for each client c, objective j, and policy π_θ, let $w^{*,j,c}_\theta$ denote the TD fixed point associated with the original transition kernel P_c, whose state-sampling stationary distribution is $\mu_{c,\theta}$. We define the critic function-approximation error by

$$\zeta_{\text{approx}} := \sup_\theta \max_{c\in[C],\, j\in[M]} \mathbb{E}_{s\sim d^\theta_{\rho,c}} \left[\left| V^{j,c}_{\pi_\theta}(s) - \phi(s)^\top w^{*,j,c}_\theta \right|^2 \right].$$

D.3 Mixing Properties and Critic Approximation Accuracy

Now, we rewrite Lemma 2 from Zhou et al. (2024):

Lemma D.4. For any policy π_θ, consider an MDP with transition kernel $P(\cdot \mid s, a)$ and stationary distribution χ_θ. Under Assumption 4.1, there exist constants $\kappa > 0$ and $\rho \in (0,1)$ such that

$$\sup_{s\in\mathcal{S}} \|P(s_t \mid s_0 = s) - \chi_\theta\|_{TV} \le \kappa \rho^t.$$

Now, using Lemma D.4 we can prove the following:

Lemma D.5 (Critic Convergence Theorem 1 from Xu et al. (2020a))**.** *Consider Algorithm 3 for Markovian mini-batch TD. For a policy π_t, let $w^{*,j,c}_t$ denote the optimal TD solution for objective $j \in [M]$. Let the stepsize be $\alpha' = \min\{\mathcal{O}(\lambda_{A_\pi}), \mathcal{O}(\lambda^{-1}_{A_\pi})\}$. Then we have*

$$\mathbb{E}\left[\|w^{j,c}_N - w^{*,j,c}_t\|_2^2\right] \le \left(1 - \mathcal{O}(\lambda_{A_\pi}\alpha')\right)^N + \mathcal{O}\left(\frac{\alpha'}{D}\right).$$

Let $N = \Theta(\log(1/\epsilon))$ and $D = \Theta(\epsilon^{-1})$. The total sample complexity for Algorithm 3 to achieve an ϵ-accurate optimal solution $w^{j,c}_N$, i.e., $\mathbb{E}[\|w^{j,c}_N - w^{*,j,c}_t\|_2^2] \le \epsilon$, is given by $DN = \mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$.

Proof: *Note that the proof of this Lemma without the projection step in Algorithm 3 is provided in Theorem 1 of Xu et al. (2020a). However, since based on Lemma 8 in Zhou et al. (2024) we have $\|w^{*,j,c}_t\|_2 \le R_w$, projection step in our Algorithm is non-expansive and the proof of Theorem 1 in Xu et al. (2020a) holds for our case as well.* □

D.4 Bounded Gradients

Lastly, we first establish that the local gradients are uniformly bounded.

Lemma D.6 (Bounded Gradient)**.** *Under our assumptions, the local stochastic gradient for any objective j and client c is bounded as $\|g^{j,c}_t\|_2 \le R \triangleq \frac{C_\psi(r_{\max}+(1+\gamma)R_w)}{1-\gamma}$.*

Proof. The gradient is $g^{j,c}_t = \frac{1}{(1-\gamma)B}\sum_{l=1}^B \delta^{j,c}_{t,l} \cdot \boldsymbol{\psi}(s_{t,l}, a_{t,l})$, where $\delta^{j,c}_{t,l} = r^{j,c}_{t,l} + \gamma \boldsymbol{\phi}(s^+_{t,l})^\top \mathbf{w}^{j,c}_t - \boldsymbol{\phi}(s_{t,l})^\top \mathbf{w}^{j,c}_t$. Given the assumptions $\|\boldsymbol{\phi}(\cdot)\|_2 \le 1$, $\|\boldsymbol{\psi}(\cdot,\cdot)\|_2 \le C_\psi$, and the critic projection step

in Algorithm 1 ensuring $\|\mathbf{w}_t^{j,c}\|_2 \le R_w$, we have:

$$\begin{aligned}\|g_t^{j,c}\|_2 &\le \frac{1}{(1-\gamma)B}\sum_{l=1}^{B}\|\delta_{t,l}^{j,c}\boldsymbol{\psi}_{t,l}\|_2 \le \frac{1}{1-\gamma}\max_l|\delta_{t,l}^{j,c}|\cdot\|\boldsymbol{\psi}_{t,l}\|_2\\ &\le \frac{C_\psi}{1-\gamma}\left(|r_{t,l}^{j,c}| + (1+\gamma)\|\boldsymbol{\phi}(\cdot)\|_2\|\mathbf{w}_t^{j,c}\|_2\right)\\ &\le \frac{C_\psi}{1-\gamma}(r_{\max}+(1+\gamma)R_w).\end{aligned}$$

□

E PRELIMINARIES

Definition E.1 (Averaged Policy and MGDA coefficients)**.** Let $\bar{\theta}_t \triangleq \frac{1}{C}\sum_{c=1}^{C}\theta_t^c$ be the globally averaged policy parameters at time step t and let $\bar{\lambda}_t := \frac{1}{C}\sum_{c=1}^{C}\lambda_t^c$, denote the average of the clients' MGDA coefficient vectors.

Definition E.2 (Gradient Matrix.)**.** For any policy θ, the gradient matrix $\nabla_\theta\mathbf{J}(\theta)\in\mathbb{R}^{d\times M}$ is formed by concatenating the gradients of the M objective functions:

$$\nabla_\theta\mathbf{J}(\theta) = \left[\nabla_\theta J^1(\theta)\ \ \nabla_\theta J^2(\theta)\ \ \cdots\ \ \nabla_\theta J^M(\theta)\right].$$

F MAIN PROOF

Theorem F.1 (Finite-time Pareto-stationarity of T-FIRM)**.** *Suppose Assumptions 4.1-4.4 hold. Consider Algorithm 2 with full client participation and uniform aggregation. Let $T \ge 2$ denote the total number of actor-update steps, and let the clients synchronize after every K updates, where $1 \le K \le T$. Suppose that the local regularized MGDA problem is*

$$\widehat{\lambda}_t^{*,c} \in \arg\min_{\lambda\in\Delta_M}\left\{\lambda^\top G_t^c\lambda + \frac{\beta}{2}\|\lambda\|_2^2\right\}, \qquad \beta>0,$$

and that the smoothed objective weights satisfy

$$\lambda_t^c = (1-\eta_t)\lambda_{t-1}^c + \eta_t\widehat{\lambda}_t^{*,c}, \qquad \eta_t = \frac{1}{t},$$

with identical initialization across clients. Define $\bar{\theta}_t := \frac{1}{C}\sum_{c=1}^{C}\theta_t^c$, $\bar{\lambda}_t := \frac{1}{C}\sum_{c=1}^{C}\lambda_t^c$, and let $\mathcal{G}(\theta) := \min_{\lambda\in\Delta_M}\|\nabla\mathbf{J}(\theta)\lambda\|_2^2$ denote the Pareto-stationarity gap. Assume that the critic iterates satisfy

$$\max_{1\le t\le T}\mathbb{E}\left[\max_{\substack{c\in[C]\\ j\in[M]}}\left\|w_t^{j,c} - w_{\theta_t^c}^{*,j,c}\right\|_2^2\right] \le \varepsilon_{\text{critic}}.$$

Consequently,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{G}(\bar{\theta}_t)\right] \le \mathcal{O}\Bigg(&\underbrace{\frac{\log T}{\alpha T}}_{\textit{optimization}} + \underbrace{\frac{M}{CB}+\frac{M}{B^2}}_{\textit{sampling}} + \underbrace{\zeta_{\text{approx}}+\varepsilon_{\text{critic}}}_{\textit{critic residual}} + \underbrace{\alpha^2K^2}_{\textit{classical drift}}\\ &+ \underbrace{\frac{M^{3/2}}{\beta}\left(\frac{1}{\sqrt{B}}+\alpha K+\sqrt{\zeta_{\text{approx}}}+\sqrt{\varepsilon_{\text{critic}}}+\zeta_{\text{het}}\right)}_{\textit{multi-objective disagreement}}\Bigg).\end{aligned}$$

where the hidden constants depend only on the fixed problem and Markov-chain parameters.

Proof: We characterize the progress made by a single update of the global model using the smoothness of the scalarized objective $\bar{\lambda}_t^\top J(\cdot)$. In fact, considering Definition E.1 and by L_J smoothness (Assumption 4.3), we will have:

$$\bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_{t+1}) \;\geq\; \bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_t) + \big\langle \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t,\, \bar{\theta}_{t+1} - \bar{\theta}_t \big\rangle - \frac{L_J}{2}\|\bar{\theta}_{t+1} - \bar{\theta}_t\|_2^2.$$

Let $\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha_t \bar{g}_t$ with $\bar{g}_t = \frac{1}{C}\sum_{c=1}^{C} g_t^c(\theta_t^c)$ and $g_t^c(\theta_t^c) \triangleq \sum_{j=1}^{M} \lambda_t^{j,c} g_t^{j,c}(\theta_t^c)$. Then, we obtain:

$$\bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_{t+1}) \;\geq\; \bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_t) + \alpha_t \langle \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t,\, \bar{g}_t \rangle - \frac{L_J \alpha_t^2}{2}\|\bar{g}_t\|_2^2.$$

Using $2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$,

$$\begin{aligned}
\bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_{t+1}) &\geq \bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_t) + \frac{\alpha_t}{2}\big(\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2 + \|\bar{g}_t\|_2^2 - \|\bar{g}_t - \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2\big) - \frac{L_J\alpha_t^2}{2}\|\bar{g}_t\|_2^2 \\
&= \bar{\lambda}_t^\top \mathbf{J}(\bar{\theta}_t) + \frac{\alpha_t}{2}\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2 + \frac{\alpha_t(1-\alpha_t L_J)}{2}\|\bar{g}_t\|_2^2 - \frac{\alpha_t}{2}\underbrace{\|\bar{g}_t - \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2}_{\mathcal{T}_t}.
\end{aligned} \tag{7}$$

Next, we try to bound the term $\mathcal{T}_t$ in Equation(7). Add and subtract $\frac{1}{C}\sum_{c=1}^{C} \nabla \mathbf{J}^c(\theta_t^c)\bar{\lambda}_t$. Then, using $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$, We decompose the term $\mathcal{T}_t$ into the following two sub-terms:

$$\mathcal{T}_t = \|\bar{g}_t - \nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\lambda}_t\|_2^2 \leq 2\underbrace{\|\bar{g}_t - \frac{1}{C}\sum_{c=1}^{C} \nabla_\theta \mathbf{J}^c(\theta_t^c)\bar{\lambda}_t\|_2^2}_{\mathcal{T}_t^{\text{local}}} + 2\underbrace{\|\big(\frac{1}{C}\sum_{c=1}^{C} \nabla_\theta \mathbf{J}^c(\theta_t^c) - \nabla_\theta \mathbf{J}(\bar{\theta}_t)\big)\bar{\lambda}_t\|_2^2}_{\mathcal{T}_t^{\text{model}}}. \tag{8}$$

Here, $\mathcal{T}_t^{\text{local}}$ represents the error in estimating the multi-objective descent direction at each client with respect to its local policy, while $\mathcal{T}_t^{\text{grad}}$ captures the effect of client drift. In the following, we bound each term separately.

Now, by adding and subtracting the term $\frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} \bar{\lambda}_t^j\, g_t^{j,c}(\theta_t^c)$ inside $\mathcal{T}_t^{\text{local}}$, and applying the triangle inequality, we obtain

$$\begin{aligned}
\mathcal{T}_t^{\text{local}} &= \|\bar{g}_t - \frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} \bar{\lambda}_t^j\, g_t^{j,c}(\theta_t^c) + \frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} \bar{\lambda}_t^j\, g_t^{j,c}(\theta_t^c) - \frac{1}{C}\sum_{c=1}^{C} \nabla_\theta \mathbf{J}^c(\theta_t^c)\bar{\lambda}_t\|_2^2 \\
&\leq 2\underbrace{\|\frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} \bar{\lambda}_t^j\big(g_t^{j,c}(\theta_t^c) - \nabla_\theta J^{j,c}(\theta_t^c)\big)\|_2^2}_{\mathcal{T}_t^{\text{grad}}} + 2\underbrace{\|\frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} g_t^{j,c}(\theta_t^c)\big(\bar{\lambda}_t^{j,c} - \lambda_t^j\big)\|_2^2}_{\mathcal{T}_t^{\text{dis}}}
\end{aligned} \tag{9}$$

Using Equations 8 and 9 we have:

$$\mathcal{T}_t \leq 4\mathcal{T}_t^{\text{grad}} + 4\mathcal{T}_t^{\text{dis}} + 2\mathcal{T}_t^{\text{model}}. \tag{10}$$

Remark F.2. The decomposition in equation 9 introduces two meaningful error sources that we must analyze.

- **The MGDA Discrepancy Error ($\mathcal{T}_t^{\text{dis}}$):** This term is a **novel component unique to our decentralized multi-objective learning framework**. It arises from the discrepancy between the locally computed MGDA weights λ_t^c at each client and the ideal, globally optimal weights λ_t. This error source is absent in prior art, which typically assumes a centralized server performs the MGDA computation and broadcasts the resulting λ_t to all clients. In our setting, to enhance communication efficiency, each client computes its own λ_t^c and shares only its model parameters θ_t^c. This design choice reduces the communication cost by a factor of M. The price of this efficiency is the introduction of the MGDA error, which necessitates a separate bound.

- **The Gradient Estimation Error ($\mathcal{T}_t^{\text{grad}}$):** This term quantifies the average variance of the local stochastic gradient estimators $g_t^{j,c}(\theta_t^c)$ across all clients and objectives. It is analogous to the standard gradient error term found in conventional Federated Learning (FL) literature.

Now, we start with bounding the sampling error of the gradient side and then will bound MGDA error in Equation (9).

F.1 Bounding the Gradient Estimation Error ($\mathcal{T}_t^{\text{GRAD}}$)

This section handles the fact that T-FIRM uses the same actor mini-batch to compute the objective gradients and the local MGDA weights. Thus, we cannot assume that the MGDA weights are independent of the stochastic gradients. The argument follows the bias-aware, pathwise treatment of stochastic multi-objective directions in Zhou et al. (2024), while additionally separating cross-client variance from finite-batch Markov initialization bias. We first decompose $\mathcal{T}_t^{\text{grad}}$ into three terms by adding and subtracting intermediate quantities. Specifically, we introduce the gradient computed with the true critic parameters, $g_t^{j,c}(\theta_t^{w_t^{*,j,c}})$, and its expectation, $\Delta_t^{j,c}(\theta_t^c, w_t^{*,j,c})$. By applying the inequality $\| \sum_{i=1}^n x_i\|_2^2 \le n \sum_{i=1}^n \|x_i\|_2^2$, we have:

$$\begin{aligned}
\mathcal{T}_t^{\text{grad}} &= \left\| \frac{1}{C} \sum_{c=1}^{C} \sum_{j=1}^{M} \bar{\lambda}_t^j \big(g_t^{j,c}(\theta_t^c) - \nabla_\theta J^{c,j}(\theta_t^c) \big) \right\|_2^2 \\
&\le 3 \underbrace{\left\| \frac{1}{C} \sum_{c=1}^{C} \sum_{j=1}^{M} \bar{\lambda}_t^j \big(g_t^{j,c}(\theta_t^c) - g_t^{j,c}(\theta_t^{w_t^{*,j,c}}) \big) \right\|_2^2}_{\triangleq \mathcal{T}_{1,1,1}} + 3 \underbrace{\left\| \frac{1}{C} \sum_{c=1}^{C} \sum_{j=1}^{M} \bar{\lambda}_t^j \big(g_t^{j,c}(\theta_t^{w_t^{*,j,c}}) - \Delta_t^{j,c}(\theta_t^c, w_t^{*,j,c}) \big) \right\|_2^2}_{\triangleq \mathcal{T}_{1,1,2}} \\
&\quad + 3 \underbrace{\left\| \frac{1}{C} \sum_{c=1}^{C} \sum_{j=1}^{M} \bar{\lambda}_t^j \big(\Delta_t^{j,c}(\theta_t^c, w_t^{*,j,c}) - \nabla_\theta J^{c,j}(\theta_t^c) \big) \right\|_2^2}_{\triangleq \mathcal{T}_{1,1,3}},
\end{aligned} \tag{11}$$

where $w_t^{*,j,c}$ is the true critic parameter for objective j on client c, and $\Delta_t^{j,c}$ is the expected policy gradient under the true critic, defined as:

$$\Delta_t^{j,c}(\theta_t^c, w_t^{*,j,c}) \triangleq \frac{1}{1-\gamma} \mathbb{E}_{(s,a)\sim\nu_{\rho,c}^{\theta_t^c}} \Big[\mathbb{E}_{s'\sim P_c(\cdot|s,a)} \big[\nabla_\theta \log \pi_{\theta_t^c}(a|s) \delta_t^{j,c}(w_t^{*,j,c}, s, a, s') \big] \Big]$$

with $\delta_t^{j,c}(w_t^{*,j,c}, s, a, s') \triangleq r_t^{j,c}(s,a) + \gamma \boldsymbol{\phi}(s')^\top w_t^{*,j,c} - \boldsymbol{\phi}(s)^\top w_t^{*,j,c}$, and $\nu^{\theta_t^c}$ is the state-action visitation distribution induced by policy $\pi_{\theta_t^c}$.

We now analyze the conditional expectation of equation 11 with respect to the filtration $\mathcal{F}_t$, contains all iterates, critic outputs, and randomness generated before sampling the actor mini-batches at actor step t.

$$\mathbb{E}[\mathcal{T}_t^{\text{grad}} \mid \mathcal{F}_t] \le 3\,\mathbb{E}[\mathcal{T}_{1,1,1} \mid \mathcal{F}_t] + 3\,\mathbb{E}[\mathcal{T}_{1,1,2} \mid \mathcal{F}_t] + 3\,\mathbb{E}[\mathcal{T}_{1,1,3} \mid \mathcal{F}_t] \tag{12}$$

To bound the other terms in equation 12, we rely on the following lemmas.

Lemma F.3. *(Approximation Error) For all $t \in [T]$, the approximation error is bounded by:*

$$\mathbb{E}[\mathcal{T}_{1,1,3} \mid \mathcal{F}_t] \le 12\zeta_{\textit{approx}}.$$

Proof. See Appendix D in Zhou et al. (2024). □

Lemma F.4. *(Critic Estimation Error) For all $t \in [T]$, the error from critic estimation is bounded by:*

$$\mathbb{E}[\mathcal{T}_{1,1,1} \mid \mathcal{F}_t] \le \frac{12}{C} \sum_{c=1}^{C} \max_{j\in[M]} \mathbb{E}[\|w_t^{*,j,c} - w_t^{j,c}\|_2^2 \mid \mathcal{F}_t].$$

Proof. See Appendix D in Zhou et al. (2024). □

Lemma F.5 (Client-averaged gradient sampling error)**.** *For every* $t \in [T]$,

$$\mathbb{E}[T_{1,1,2} \mid \mathcal{F}_t] \le 4M \frac{C_\psi^2}{(1-\gamma)^2}(r_{\max} + 2R_w)^2 \big[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} + \frac{\kappa^2}{(1-\rho)^2 B^2}\big].$$

Proof. For each objective $j \in [M]$, client $c \in [C]$, and sample $\ell \in \{0, \dots, B-1\}$, define the per-sample actor update

$$X_{t,\ell}^{j,c} \coloneqq \frac{1}{1-\gamma}\psi_{\theta_t^c}(s_{t,\ell}^c, a_{t,\ell}^c)\delta_t^{j,c}\left(w_t^{*,j,c}, s_{t,\ell}^c, a_{t,\ell}^c, s_{t,\ell}^{c,+}\right).$$

By Lemma D.6,

$$\begin{aligned}\|X_{t,\ell}^{j,c}\|_2 &\le \frac{C_\psi}{1-\gamma}\left(r_{\max} + (1+\gamma)\|w_t^{*,j,c}\|_2\right)\\ &\le \frac{C_\psi}{1-\gamma}\left(r_{\max} + (1+\gamma)R_w\right)\\ &\le \frac{C_\psi}{1-\gamma}(r_{\max} + 2R_w).\end{aligned}$$

For notational convenience, let

$$H \coloneqq \frac{1}{1-\gamma}C_\psi\left(r_{\max} + (1+\gamma)R_w\right) \le \frac{C_\psi(r_{\max}+2R_w)}{1-\gamma}.$$

Let

$$\overline{X}_t^{j,c} \coloneqq \mathbb{E}_{\substack{\theta_t^c \\ \nu_{\rho,c}}}\left[X_t^{j,c}\right] = \Delta_t^{j,c}\left(\theta_t^c, w_t^{*,j,c}\right).$$

The finite-batch gradient error can then be written as

$$\begin{aligned}e_t^{j,c} &\coloneqq g_t^{j,c}(\theta_t^c, w_t^{*,j,c}) - \Delta_t^{j,c}(\theta_t^c, w_t^{*,j,c})\\ &= \frac{1}{B}\sum_{\ell=0}^{B-1}\left(X_{t,\ell}^{j,c} - \overline{X}_t^{j,c}\right).\end{aligned}$$

We decompose

$$e_t^{j,c} = \xi_t^{j,c} + b_t^{j,c}, \qquad b_t^{j,c} \coloneqq \mathbb{E}[e_t^{j,c} \mid \mathcal{F}_t].$$

By construction,

$$\mathbb{E}[\xi_t^{j,c} \mid \mathcal{F}_t] = 0.$$

We first bound the finite-batch initialization bias. Since $\overline{X}_t^{j,c}$ is the stationary expectation,

$$b_t^{j,c} = \frac{1}{B}\sum_{\ell=0}^{B-1}\left(\mathbb{E}[X_{t,\ell}^{j,c} \mid \mathcal{F}_t] - \overline{X}_t^{j,c}\right).$$

For any vector-valued function h satisfying $\sup_z \|h(z)\|_2 \le H$, $\|\mathbb{E}_\nu[h] - \mathbb{E}_\mu[h]\|_2 \le 2H\|\nu - \mu\|_{\mathrm{TV}}$. Therefore, Lemma D.4 gives

$$\begin{aligned}\|b_t^{j,c}\|_2 \le \frac{2H}{B}\sum_{\ell=0}^{B-1}\left\|\widetilde{P}_{c,\theta_t^c}^{\ell}(\cdot \mid s_{t,0}^c)\,\pi_{\theta_t^c}(\cdot \mid \cdot) - \nu_{\rho,c}^{\theta_t^c}\right\|_{\mathrm{TV}} &= \frac{2H}{B}\sum_{\ell=0}^{B-1}\left\|\tilde{P}_{c,\theta_t^c}^{\ell}(\cdot \mid s_{t,0}^c) - d_{\rho,c}^{\theta_t^c}\right\|_{\mathrm{TV}}\\ &\le \frac{2H\kappa}{B}\sum_{\ell=0}^{B-1}\rho^\ell\\ &\le \frac{2H\kappa}{(1-\rho)B}.\end{aligned}$$

Consequently,

$$\|b_t^{j,c}\|_2^2 \le \frac{4H^2\kappa^2}{(1-\rho)^2B^2}.$$

We next bound the conditionally centered component. Define

$$Y_{t,\ell}^{j,c} \coloneqq X_{t,\ell}^{j,c} - \mathbb{E}[X_{t,\ell}^{j,c} \mid \mathcal{F}_t].$$

Then

$$\xi_t^{j,c} = \frac{1}{B}\sum_{\ell=0}^{B-1} Y_{t,\ell}^{j,c}, \qquad \mathbb{E}[Y_{t,\ell}^{j,c} \mid \mathcal{F}_t] = 0,$$

and $\|Y_{t,\ell}^{j,c}\|_2 \le 2H$. Expanding the squared norm gives

$$\mathbb{E}\left[\|\xi_t^{j,c}\|_2^2 \mid \mathcal{F}_t\right] = \frac{1}{B^2}\sum_{\ell=0}^{B-1}\mathbb{E}\left[\|Y_{t,\ell}^{j,c}\|_2^2 \mid \mathcal{F}_t\right] + \frac{2}{B^2}\sum_{0\le\ell<q\le B-1}\mathbb{E}\left[\left\langle Y_{t,\ell}^{j,c}, Y_{t,q}^{j,c}\right\rangle \mid \mathcal{F}_t\right].$$

The diagonal terms satisfy

$$\mathbb{E}\left[\|Y_{t,\ell}^{j,c}\|_2^2 \mid \mathcal{F}_t\right] \le 4H^2.$$

Moreover, following the geometrically mixing covariance argument of Zhou et al. (2024), for $q > \ell$,

$$\left|\mathbb{E}\left[\left\langle Y_{t,\ell}^{j,c}, Y_{t,q}^{j,c}\right\rangle \mid \mathcal{F}_t\right]\right| \le 8H^2\kappa\rho^{q-\ell}.$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\|\xi_t^{j,c}\|_2^2 \mid \mathcal{F}_t\right] &\le \frac{4H^2}{B} + \frac{16H^2\kappa}{B^2}\sum_{0\le\ell<q\le B-1}\rho^{q-\ell} \\ &\le \frac{4H^2}{B} + \frac{16H^2\kappa\rho}{(1-\rho)B} \\ &= \frac{4H^2(1-\rho+4\kappa\rho)}{(1-\rho)B}.\end{aligned}$$

Now define the client-averaged error for objective j as

$$\bar{e}_t^j \coloneqq \frac{1}{C}\sum_{c=1}^{C} e_t^{j,c}.$$

Because the client chains are conditionally independent and $\xi_t^{j,c}$ is conditionally centered,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{C}\sum_{c=1}^{C}\xi_t^{j,c}\right\|_2^2\middle|\mathcal{F}_t\right] &= \frac{1}{C^2}\sum_{c=1}^{C}\mathbb{E}\left[\|\xi_t^{j,c}\|_2^2 \mid \mathcal{F}_t\right] \\ &\le \frac{4H^2(1-\rho+4\kappa\rho)}{C(1-\rho)B}.\end{aligned}$$

For the bias component,

$$\begin{aligned}\left\|\frac{1}{C}\sum_{c=1}^{C} b_t^{j,c}\right\|_2^2 &\le \left(\frac{1}{C}\sum_{c=1}^{C}\|b_t^{j,c}\|_2\right)^2 \\ &\le \frac{4H^2\kappa^2}{(1-\rho)^2B^2}.\end{aligned}$$

The centered and bias components have zero conditional cross term, and hence

$$\begin{aligned}\mathbb{E}\left[\|\bar{e}_t^j\|_2^2 \mid \mathcal{F}_t\right] &\le \frac{4H^2(1-\rho+4\kappa\rho)}{C(1-\rho)B} + \frac{4H^2\kappa^2}{(1-\rho)^2B^2} \\ &= 4H^2\left[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} + \frac{\kappa^2}{(1-\rho)^2B^2}\right].\end{aligned}$$

Finally,

$$T_{1,1,2} = \left\|\sum_{j=1}^{M}\bar{\lambda}_t^j\bar{e}_t^j\right\|_2^2.$$

Since $\bar{\lambda}_t \in \Delta_M$, pathwise,

$$\begin{aligned}\left\|\sum_{j=1}^{M} \bar{\lambda}_t^j \bar{e}_t^j\right\|_2^2 &\leq \sum_{j=1}^{M} \bar{\lambda}_t^j \|\bar{e}_t^j\|_2^2 \\ &\leq \max_{j\in[M]} \|\bar{e}_t^j\|_2^2 \\ &\leq \sum_{j=1}^{M} \|\bar{e}_t^j\|_2^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[T_{1,1,2} \mid \mathcal{F}_t] &\leq \sum_{j=1}^{M} \mathbb{E}\left[\|\bar{e}_t^j\|_2^2 \mid \mathcal{F}_t\right] \\ &\leq 4MH^2\left[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} + \frac{\kappa^2}{(1-\rho)^2B^2}\right] \\ &\leq 4M\frac{C_\psi^2}{(1-\gamma)^2}(r_{\max}+2R_w)^2\big[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} + \frac{\kappa^2}{(1-\rho)^2B^2}\big].\end{aligned} \tag{13}$$

The pathwise simplex inequalities do not require λ_t to be independent of the stochastic gradients used to compute it. □

Combining Lemma F.3– Lemma F.5 yields:

$$\begin{aligned}\mathbb{E}[\mathcal{T}_t^{\text{grad}} \mid \mathcal{F}_t] \leq\ & 36\zeta_{\text{approx}} \\ &+ \frac{36}{C}\sum_{c=1}^{C}\max_{j\in[M]}\mathbb{E}[\|w_t^{*,j,c} - w_t^{j,c}\|_2^2 \mid \mathcal{F}_t] \\ & 12MC_\psi^2(r_{\max}+2R_w)^2\big[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} \\ &+ \frac{\kappa^2}{(1-\rho)^2B^2}\big].\end{aligned} \tag{14}$$

F.2 Bounding the Multi-Objective Disagreement Drift ($\mathcal{T}_t^{\text{dis}}$)

This term captures the error arising from the discrepancy between locally computed MGDA weights (λ_t^c) across clients. In conventional FL, drift is often bounded by assuming that with a sufficiently large local batch size (B), the variance of client gradients diminishes. However, it is not immediately obvious that the solutions to the local MGDA optimization problems, λ_t^c, will also converge as B increases.

Recall $\bar{\lambda}_t^j \triangleq \frac{1}{C}\sum_{c'=1}^{C} \lambda_t^{j,c'}$ denote the average MGDA weight for objective j.

$$\begin{aligned}
\mathbb{E}[\mathcal{T}_t^{\text{dis}}] &= \mathbb{E}\left[\left\|\frac{1}{C}\sum_{c=1}^{C}\sum_{j=1}^{M} g_t^{j,c}(\theta_t^c)\left(\lambda_t^{j,c}-\bar{\lambda}_t^j\right)\right\|_2^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{1}{C}\sum_{c=1}^{C}\left\|\sum_{j=1}^{M} g_t^{j,c}(\theta_t^c)\left(\lambda_t^{j,c}-\bar{\lambda}_t^j\right)\right\|_2\right)^2\right] && \text{(Triangle Inequality)} \\
&\leq \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}\left[\left\|\sum_{j=1}^{M} g_t^{j,c}(\theta_t^c)\left(\lambda_t^{j,c}-\bar{\lambda}_t^j\right)\right\|_2^2\right] && \text{(Jensen's Inequality)} \\
&\leq \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}\left[\left(\sum_{j=1}^{M}\|g_t^{j,c}(\theta_t^c)\|_2\left|\lambda_t^{j,c}-\bar{\lambda}_t^j\right|\right)^2\right] && \text{(Triangle Inequality)}
\end{aligned} \tag{15}$$

Applying Lemma D.6 to Equation equation 15, we get:

$$\mathbb{E}[\mathcal{T}_t^{\text{dis}}] \quad \leq \frac{R^2}{C}\sum_{c=1}^{C}\mathbb{E}\left[\left(\sum_{j=1}^{M}\left|\lambda_t^{j,c}-\bar{\lambda}_t^j\right|\right)^2\right]$$

Now,let $S_c \triangleq \sum_{j=1}^{M}|\lambda_t^{j,c}-\bar{\lambda}_t^j|$. Now, since $\sum_{j=1}^{M}|\lambda_t^{j,c}|=1$ and $\sum_{j=1}^{M}|\bar{\lambda}_t^j|=1$, by the triangle inequality, $S_c \leq \sum_{j=1}^{M}|\lambda_t^{j,c}| + \sum_{j=1}^{M}|\bar{\lambda}_t^j| = 1+1=2$. Moreover, for any non-negative value $S_c \in [0,2]$, the inequality $S_c^2 \leq 2S_c$ holds. Applying this insight, we can bound the term in Equation equation 16 as follows:

$$\begin{aligned}
\mathbb{E}[\mathcal{T}_t^{\text{dis}}] &\leq \frac{R^2}{C}\sum_{c=1}^{C}\mathbb{E}\left[\left(\sum_{j=1}^{M}\left|\lambda_t^{j,c}-\bar{\lambda}_t^j\right|\right)^2\right] \\
&\leq \frac{2R^2}{C}\sum_{c=1}^{C}\mathbb{E}\left[\sum_{j=1}^{M}\left|\lambda_t^{j,c}-\bar{\lambda}_t^j\right|\right] && \text{(Using } S_c^2 \leq 2S_c \text{ since } S_c \leq 2\text{)} \\
&= \frac{2R^2}{C}\sum_{c=1}^{C}\mathbb{E}\left[\left\|\lambda_t^c-\bar{\lambda}_t\right\|_1\right] \\
&\leq \frac{2R^2}{C^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\mathbb{E}\left[\left\|\lambda_t^c-\lambda_t^{c'}\right\|_1\right], && (16) \\
&\leq \frac{2R^2\sqrt{M}}{C^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\mathbb{E}\left[\left\|\lambda_t^c-\lambda_t^{c'}\right\|_2\right] && \text{(Using } \|v\|_1 \leq \sqrt{M}\|v\|_2 \text{ for all } v\in\mathbb{R}^M\text{)}
\end{aligned} \tag{17}$$

The inequality equation 16 follows directly from the triangle inequality for the ℓ_1 norm:

$$\|\lambda_t^c - \bar{\lambda}_t\|_1 = \left\|\lambda_t^c - \frac{1}{C}\sum_{c'=1}^{C}\lambda_t^{c'}\right\|_1 = \left\|\frac{1}{C}\sum_{c'=1}^{C}(\lambda_t^c - \lambda_t^{c'})\right\|_1 \leq \frac{1}{C}\sum_{c'=1}^{C}\|\lambda_t^c - \lambda_t^{c'}\|_1.$$

The bound now depends on the pairwise difference between local MGDA weights. We decompose this difference recursively using the update rule $\lambda_t^c = (1-\eta_t)\lambda_{t-1}^c + \eta_t \lambda_t^{*,c}$:

$$\begin{aligned}\|\lambda_t^c - \lambda_t^{c'}\|_2 &= \|(1-\eta_t)(\lambda_{t-1}^c - \lambda_{t-1}^{c'}) + \eta_t(\widehat{\lambda}_t^{*,c} - \widehat{\lambda}_t^{*,c'})\|_2 \\ &\leq (1-\eta_t)\|\lambda_{t-1}^c - \lambda_{t-1}^{c'}\|_2 + \eta_t\|\widehat{\lambda}_t^{*,c} - \widehat{\lambda}_t^{*,c'}\|_2. \end{aligned} \tag{18}$$

Unrolling this recursion from t down to 1 and noting that $\lambda_0^c = \lambda_0^{c'}$ for all c, c', we obtain:

$$\begin{aligned}&\mathbb{E}[\|\lambda_t^c - \lambda_t^{c'}\|_2] \\ &\leq \sum_{i=1}^{t}\left[\eta_i \prod_{k=i+1}^{t}(1-\eta_k)\right]\mathbb{E}[\|\widehat{\lambda}_i^{*,c} - \widehat{\lambda}_i^{*,c'}\|_2].\end{aligned} \tag{19}$$

Now using Corollary D.3 and substituting equation 19 into equation 17, we arrive at:

$$\begin{aligned}\mathbb{E}[\mathcal{T}_t^{\text{dis}}] &\leq \frac{8R^3\sqrt{M^2}}{\beta C^2}\sum_{c,c'}\sum_{i=1}^{t}\left[\eta_i \prod_{k=i+1}^{t}(1-\eta_k)\right]\sqrt{\mathbb{E}[\|A_i^c - A_i^{c'}\|_{\text{F}}^2]} \\ &= \frac{8R^3\sqrt{M^2}}{\beta C^2}\sum_{c,c'}\sum_{i=1}^{t}\left[\eta_i \prod_{k=i+1}^{t}(1-\eta_k)\right]\sqrt{\sum_{j=1}^{M}\mathbb{E}[\|g_i^{j,c} - g_i^{j,c'}\|_2^2]}\end{aligned} \tag{20}$$

This bound demonstrates that the Multi-Objective Disagreement Drift is controlled by the history of client drift, i.e., the difference in gradients across clients. From this point onward, we assume that the number of critic iterations is chosen such that

$$\max_t \mathbb{E}\left[\max_{\substack{c\in[C]\\ j\in[M]}}\|w_t^{j,c} - w_t^{*,j,c}\|_2^2\right] \leq \varepsilon_{\text{critic}}. \tag{21}$$

We now bound this drift.

Lemma F.6 (Client Gradient Drift)**.** *Let $t_0 \leq t$ be the last synchronization time-step. There exists a universal numerical constant $C_{\text{pair}} > 0$ such that, for every $j \in [M]$ and $c, c' \in [C]$,*

$$\begin{aligned}&\mathbb{E}\Big[\|g_t^{j,c} - g_t^{j,c'}\|_2^2\Big] \leq \\ &D_i^2 \triangleq C_{\text{pair}}\Big(\zeta_{\textit{approx}} + \frac{C_\psi^2}{(1-\gamma)^2}(1+\gamma)^2\varepsilon_{\text{critic}} \\ &+ \frac{C_\psi^2}{(1-\gamma)^2}\frac{(r_{\max} + 2R_w)^2(1-\rho+4\kappa\rho)}{(1-\rho)B} \\ &+ L_J^2R^2\alpha^2(t-t_0)^2 + \zeta_{\text{het}}^2\Big).\end{aligned}$$

Proof. The proof can be found in Appendix G.1. □

By substituting the bound from Lemma F.6 into equation 20, we obtain :

$$\begin{aligned}\mathbb{E}[\mathcal{T}_t^{\text{dis}}] &\le \frac{8R^3\sqrt{M^2}}{\beta C^2}\sum_{c,c'}\sum_{i=1}^{t}\left[\eta_i\prod_{k=i+1}^{t}(1-\eta_k)\right]\sqrt{\sum_{j=1}^{M}D_i^2} \\ &\le \frac{8R^3\sqrt{M^3}}{\beta C^2}\sum_{c,c'}\sum_{i=1}^{t}\left[\eta_i\prod_{k=i+1}^{t}(1-\eta_k)D_i\right]\end{aligned} \tag{22}$$

To simplify the expression, let $a_{t,i} \coloneqq \eta_i\prod_{k=i+1}^{t}(1-\eta_k)$, then $\sum_{i=1}^{t} a_{t,i} \le 1$. Moreover, we bound the average over client pairs by the maximum over all pairs, which allows us to remove the $\frac{1}{C^2}\sum_{c,c'}$ term. This yields the following comprehensive bound:

$$\begin{aligned}\mathbb{E}[\mathcal{T}_t^{\text{dis}}] \le \frac{8R^3M^{3/2}}{\beta}\sqrt{C_{\text{pair}}}\Bigg(\sqrt{\zeta_{\text{approx}}} + \frac{C_\psi}{1-\gamma}(1+\gamma)\sqrt{\varepsilon_{\text{critic}}} + 4\frac{C_\psi}{1-\gamma}\sqrt{\frac{(r_{\max}+2R_w)^2(1-\rho+4\kappa\rho)}{(1-\rho)B}} \\ + 4L_J R\alpha K + 2\zeta_{\text{het}}\Bigg) \triangleq \mathcal{E}_{\text{MGDA}}(t),\end{aligned} \tag{23}$$

where remember that K is the maximum number of local updates. This expression, which we can denote as $\mathcal{E}_{\text{MGDA}}(t)$, provides the final upper bound on the MGDA discrepancy error. It is composed of terms related to function approximation error (ζ_{approx}), local critic estimation error, gradient variance (inversely proportional to batch size B), and divergence due to local updates (proportional to $i - i_0$).

F.2.1 Bouding the term $\mathcal{T}_t^{\text{model}}$

Full participation and uniform aggregation imply the exact cancellation

$$\frac{1}{C}\sum_{c=1}^{C}\left[\nabla\mathbf{J}^c(\bar{\theta}_t) - \nabla\mathbf{J}(\bar{\theta}_t)\right] = 0.$$

Therefore,

$$\begin{aligned}\mathcal{T}_t^{\text{model}} &= \left\|\frac{1}{C}\sum_{c=1}^{C}\left[\nabla\mathbf{J}^c(\theta_t^c) - \nabla\mathbf{J}^c(\bar{\theta}_t)\right]\bar{\lambda}_t\right\|_2^2 \\ &\le \frac{L_J^2}{C}\sum_{c=1}^{C}\|\theta_t^c - \bar{\theta}_t\|_2^2 \\ &\le 4L_J^2R^2\alpha^2K^2.\end{aligned} \tag{24}$$

F.3 Final step in bounding $\mathcal{T}_t$

Now, using Equationequation 8, and the bounds obtained for $\mathcal{T}_t^{\text{grad}}$, $\mathcal{T}_t^{\text{dis}}$ and $\mathcal{T}_t^{\text{model}}$ in Equations equation 14, equation 23, and equation 24 we will have:

$$\begin{aligned}E[\mathcal{T}_t] \le \mathbb{E}\Big[\mathbb{E}[\mathcal{Q}_t \mid \mathcal{F}_t]\Big] = C_{\text{final}}\Big(36\zeta_{\text{approx}} + 36\varepsilon_{\text{critic}} + 12M\frac{C_\psi^2}{(1-\gamma)^2}(r_{\max}+2R_w)^2\big[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B} \\ + \frac{\kappa^2}{(1-\rho)^2B^2}\big] + 4\,\mathcal{E}_{\text{MGDA}}(t) + 8\,L_J^2\,R^2\,\alpha^2K^2\Big)\end{aligned} \tag{25}$$

F.4 Final Convergence Analysis

We now consolidate the preceding bounds to derive the main convergence result.

F.4.1 Rearranging the Descent Lemma

Thus, using Equationequation 25, now by letting $\alpha_t = \alpha$ and $\alpha \in (0, \frac{1}{L_J}]$, we can continue Equationequation 7 as follows:

$$\mathbb{E}\left[\bar{\boldsymbol{\lambda}}_t^\top \mathbf{J}(\bar{\theta}_{t+1}) \mid \mathcal{F}_t\right] \geq \mathbb{E}\left[\bar{\boldsymbol{\lambda}}_t^\top \mathbf{J}(\bar{\theta}_t) + \frac{\alpha}{2}\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\boldsymbol{\lambda}}_t\|_2^2 - \frac{\alpha}{2}\mathcal{T}_1 \mid \mathcal{F}_t\right]. \tag{26}$$

$$\rightarrow \mathbb{E}\left[\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\boldsymbol{\lambda}}_t\|_2^2 \mid \mathcal{F}_t\right] \leq \frac{2}{\alpha}\mathbb{E}\left[\bar{\boldsymbol{\lambda}}_t^\top \left(\mathbf{J}(\bar{\theta}_{t+1}) - \mathbf{J}(\bar{\theta}_t)\right) \mid \mathcal{F}_t\right] + \mathbb{E}[\mathcal{Q}_t \mid \mathcal{F}_t]. \tag{27}$$

Now, by taking expectation of $\mathcal{F}_t$ on both side of Equation equation 27 we will have:

$$\mathbb{E}\left[\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\boldsymbol{\lambda}}_t\|_2^2\right] \leq \frac{2}{\alpha}\mathbb{E}\left[\bar{\boldsymbol{\lambda}}_t^\top \left(\mathbf{J}(\bar{\theta}_{t+1}) - \mathbf{J}(\bar{\theta}_t)\right)\right] + \mathbb{E}[\mathcal{Q}_t]. \tag{28}$$

F.4.2 Bounding the Telescopic Sum

The first term on the right-hand side of Equation equation 28 forms a telescopic sum when averaged over time:

$$\frac{1}{T}\sum_{t=1}^{T} E\left[\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\boldsymbol{\lambda}}_{\boldsymbol{t}}\|_2^2\right] \leq \frac{2}{\alpha}\,\frac{1}{T}\left(\sum_{t=1}^{T}\bar{\boldsymbol{\lambda}}_{\boldsymbol{t}}^\top\left(\mathbf{J}(\bar{\theta}_{t+1}) - \mathbf{J}(\bar{\theta}_t)\right)\right) + \frac{1}{T}\sum_{t=1}^{T} E\left[\mathcal{Q}_t(B, \beta, \alpha, \eta, \zeta_{approx}, M)\right] \tag{29}$$

We can bound the first term on RHS of Equation equation 29 bound this sum using the following lemma:

Lemma F.7 (Telescopic Sum Bound)**.** *Let the learning rate for the MGDA weights be η_t. The telescopic sum of objective values is bounded as:*

$$\sum_{t=1}^{T}\mathbb{E}\left[\boldsymbol{\lambda}_t^\top\left(\mathbf{J}(\bar{\theta}_{t+1}) - \mathbf{J}(\bar{\theta}_t)\right)\right] \leq \frac{r_{\max}}{1-\gamma}\left(1 + \sum_{t=1}^{T-1} 2\eta_t\right).$$

Proof. Applying summation by parts, the sum can be rewritten and bounded using Hölder's inequality:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\bar{\boldsymbol{\lambda}}_t^\top \Delta\mathbf{J}_{t+1}\right] &= \mathbb{E}\left[\sum_{t=1}^{T-1}(\bar{\boldsymbol{\lambda}}_t - \bar{\boldsymbol{\lambda}}_{t+1})^\top \mathbf{J}_{t+1} - \bar{\boldsymbol{\lambda}}_1^\top \mathbf{J}_1 + \bar{\boldsymbol{\lambda}}_T^\top \mathbf{J}_{T+1}\right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T-1}\|\bar{\boldsymbol{\lambda}}_t - \bar{\boldsymbol{\lambda}}_{t+1}\|_1\|\mathbf{J}_{t+1}\|_\infty + \|\bar{\boldsymbol{\lambda}}_T\|_1\|\mathbf{J}_{T+1}\|_\infty\right].\end{aligned}$$

The objective value is bounded by $\|\mathbf{J}(\cdot)\|_\infty \leq \frac{r_{\max}}{1-\gamma}$. The difference $\|\bar{\boldsymbol{\lambda}}_t - \bar{\boldsymbol{\lambda}}_{t+1}\|_1$ is bounded by $2\eta_t$ due to the smoothing update and the fact that all $\boldsymbol{\lambda}$ vectors lie on the probability simplex. Combining these yields the result. □

F.4.3 Final Convergence Rate

By applying Lemma F.7 on Equation equation 29 we will have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla_\theta \mathbf{J}(\bar{\theta}_t)\bar{\boldsymbol{\lambda}}_t\|_2^2\right]\frac{2}{T\alpha}\frac{r_{\max}}{1-\gamma}\left(1 + \sum_{t=1}^{T-1} 2\eta_t\right) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{Q}_t]. \tag{30}$$

Now, we have:

$$\mathcal{Q}_t(B,\beta,\alpha,\eta,\zeta_{approx},M) \le C_{\text{final}}\Bigg(\frac{R^3\sqrt{M^3}}{\beta}\Big[4\sqrt{\zeta_{\text{approx}}}+4\sqrt{\varepsilon_{\text{critic}}}+4\frac{C_\psi}{1-\gamma}\sqrt{\tfrac{(r_{\max}+R_w)^2(1-\rho+4\kappa\rho)}{(1-\rho)B}} + 4L_J R\,\alpha\,K+2\zeta_{\text{het}}\Big] + 12M\frac{C_\psi^2}{(1-\gamma)^2}(r_{\max}+2R_w)^2\big[\frac{1-\rho+4\kappa\rho}{C(1-\rho)B}+\frac{\kappa^2}{(1-\rho)^2B^2}\big] + 36\,\varepsilon_{\text{critic}}+36\zeta_{\text{approx}}+8\,L_J^2\,R^2\,\alpha^2K^2.\Bigg) \tag{31}$$

Moreover, if we choose $\eta_t = \frac{1}{t}$ we will have

$$\frac{2r_{\max}}{T(1-\gamma)\alpha}\left(1+\sum_{t=1}^{T}2\eta_t\right) \le \frac{2\ r_{\max}}{\alpha T(1-\gamma)}\left(3+2\log(T)\right) \tag{32}$$

Substituting the upper bounds from Equation equation 32 and Equation equation 31 into Equation equation 30, we arrive at the final convergence rate:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{G}(\bar\theta_t)] \le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathbf{J}(\bar\theta_t)\bar\lambda_t\right\|_2^2\right] \le \mathcal{O}\Bigg(\frac{1+\log T}{\alpha T}+\frac{M}{CB}+\frac{M}{B^2}+\zeta_{\text{approx}}+\varepsilon_{\text{critic}}+\alpha^2K^2 + \frac{M^{3/2}}{\beta}\left[\frac{1}{\sqrt{B}}+\alpha K+\sqrt{\zeta_{\text{approx}}}+\sqrt{\varepsilon_{\text{critic}}}+\zeta_{\text{het}}\right]\Bigg),$$

This completes the proof. □

G Utility Lemmas

G.1 Proof of Lemma F.6

Using triangle inequality we can bound $\|g_t^{j,c}-g_t^{j,c'}\|_2$ as follows:

$$\begin{aligned}\|g_t^{j,c}-g_t^{j,c'}\|_2^2 \le 7\Big(&\|g_t^{j,c}-g_t^{j,c}(w_t^{*,j,c})\|_2^2+\|g_t^{j,c}(w_t^{*,j,c})-\Delta_t^{j,c}(\theta_t^c,w_t^{*,j,c})\|_2^2\\ &+\|\Delta_t^{j,c}(\theta_t^c,w_t^{*,j,c})-\nabla_\theta J^{j,c}(\theta_t^c)\|_2^2\\ &+\|\nabla_\theta J^{j,c}(\theta_t^c)-\nabla_\theta J^{j,c'}(\theta_t^{c'})\|_2^2\\ &+\|\nabla_\theta J^{j,c'}(\theta_t^{c'})-\Delta_t^{j,c'}(\theta_t^{c'},w_t^{*,j,c'})\|_2^2\\ &+\|\Delta_t^{c'}(\theta_t^{c'},w_t^{*,j,c'})-g_t^{j,c'}(w_t^{*,j,c'})\|_2^2\\ &+\|g_t^{j,c'}(w_t^{*,j,c'})-g_t^{j,c'}\|_2^2\Big)\end{aligned} \tag{33}$$

Now applying Lemmas F.3, F.4, and F.5 on Equation equation 33 we will have:

$$\begin{aligned}E\Big[\|g_t^{j,c}-g_t^{j,c'}\|_2^2\Big] \le 7\Big(&16\zeta_{\text{approx}}+8\frac{C_\psi^2}{(1-\gamma)^2}(1+\gamma)^2\varepsilon_{\text{critic}}\\ &+16\frac{C_\psi^2}{(1-\gamma)^2}\frac{(r_{\max}+2R_{\text{w}})^2(1-\rho+4\kappa\rho)}{(1-\rho)B}\\ &+2\,E\Big[\|\nabla_\theta J^{c,j}(\theta_t^c)-\nabla_\theta J^{c,j}(\theta_t^{c'})\|_2^2\Big]\\ &+2\,E\Big[\|\nabla_\theta J^{c,j}(\theta_t^{c'})-\nabla_\theta J^{c',j}(\theta_t^{c'})\|_2^2\Big]\Big)\end{aligned} \tag{34}$$

Now, in order to bound $\mathbb{E}\Big[\|\nabla_\theta J^{c,j}(\theta_t^c) - \nabla_\theta J^{c,j}(\theta_t^{c'})\|_2^2\Big]$ and $\mathbb{E}\Big[\|\nabla_\theta J^{c,j}(\theta_t^{c'}) - \nabla_\theta J^{c',j}(\theta_t^{c'})\|_2^2\Big]$ in Equation (34), we apply Proposition 1 from Xu et al. (2020a) and use Assumption 4.4 to obtain the following:

$$\begin{aligned} E\Big[\|g_t^{j,c} - g_t^{j,c'}\|_2^2\Big] \leq 7\Big(&16\zeta_{\text{approx}} \\ &+ 8\frac{C_\psi^2}{(1-\gamma)^2}(1+\gamma)^2\varepsilon_{\text{critic}} \\ &+ 16\frac{C_\psi^2}{(1-\gamma)^2}\frac{(r_{\max}+2R_{\text{w}})^2(1-\rho+4\kappa\rho)}{(1-\rho)B} \\ &+ 2L_J^2 E\Big[\|\theta_t^c - \theta_t^{c'}\|_2^2\Big] + 8\zeta_{\text{het}}^2\Big) \end{aligned} \tag{35}$$

where $L_J = \frac{r_{\max}}{1-\gamma}\left(4C_\nu C_\psi + L_\psi\right)$ and $C_\nu = \left(\frac{1}{2}\right)C_\pi\left(1 + \lceil \log_\rho \kappa^{-1} \rceil\right) + (1-\rho)^{-1}$. Now, since we have $\theta_t^c = \theta_{t_0}^c + \sum_{\tau=0}^{t-t_0}\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c}\, g_{t_0+\tau}^{j,c}$, we will have:

$$\|\theta_t^c - \theta_t^{c'}\|_2 = \|\theta_{t_0}^c + \alpha\sum_{\tau=0}^{t-t_0}\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c}\, g_{t_0+\tau}^{j,c} - \theta_{t_0}^{c'} - \alpha\sum_{\tau=0}^{t-t_0}\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c'}\, g_{t_0+\tau}^{j,c'}\|_2 \tag{36}$$

now, since at t_0 we have $\theta_{t_0}^c = \theta_{t_0}^{c'}$, and applying triangular inequality we will have:

$$\begin{aligned} \|\theta_t^c - \theta_t^{c'}\|_2 &\leq \alpha\sum_{\tau=0}^{t-t_0}\|\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c}\, g_{t_0+\tau}^{j,c} \;-\; \lambda_{t_0+\tau}^{j,c'}\, g_{t_0+\tau}^{j,c'}\|_2 \\ &\leq \alpha\sum_{\tau=0}^{t-t_0}\|\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c}\, g_{t_0+\tau}^{j,c}\|_2 \;+\|\sum_{j=1}^{M}\lambda_{t_0+\tau}^{j,c'}\, g_{t_0+\tau}^{j,c'}\|_2 \end{aligned} \tag{37}$$

Now, using the fact that $\|g_t^{j,c}\|_2 \leq R, \forall t \in [T]$, and $\lambda_t^c \in \Delta(R^d)$, we will have:

$$E\Big[\|\theta_t^c - \theta_t^{c'}\|_2^2\Big] \;\leq 4R^2\alpha^2\,(t-t_0)^2 \tag{38}$$

Thus, Equation(38)and Equation(35) yield the final result. □

H Preference-Weighted MGDA Subproblem

To incorporate explicit user preferences, we generalize the local MGDA subproblem by replacing the uniform regularizer $\frac{\beta}{2}I$ in Equation equation 3 with a diagonal weighting matrix $\text{Diag}(\mathbf{p}^{-1})$, where $\mathbf{p} = [p_1, \ldots, p_M]$ is a vector of positive preference weights. The full expression becomes:

$$\lambda^* \in \arg\min_{\lambda\in\Delta_M} \lambda^\top\left(\widehat{G} + \text{Diag}(\mathbf{p}^{-1})\right)\lambda, \tag{39}$$

where Δ_M denotes the probability simplex. A larger preference p_j reduces its penalty term $1/p_j$, encouraging the optimizer to assign greater weight λ_j to the corresponding objective and bias the descent direction accordingly.

I Derivation of Gradient Heterogeneity Bound

In Assumption 4.4, we introduced the constant ζ_{het} to bound the deviation between the local objective gradient $\nabla_\theta J^c(\theta)$ and the global objective gradient $\nabla_\theta \mathbf{J}(\theta)$. In this section, we formally derive how ζ_{het} depends on the fundamental sources of heterogeneity in the MDPs: the transition dynamics and the reward functions.

Setup. Suppose that the transition kernels and objective-specific reward functions of any two clients c, c' satisfy

$$\sup_{s,a} \|P_c(\cdot \mid s,a) - P_{c'}(\cdot \mid s,a)\|_{\mathrm{TV}} \le \epsilon_p,$$
$$\sup_{s,a} \left|r_c^i(s,a) - r_{c'}^i(s,a)\right| \le \epsilon_r, \qquad \forall i \in [M].$$

Proposition I.1. *Under Assumptions 4.1-4.3 and the heterogeneity conditions above, the gradient heterogeneity is bounded by:*

$$\|\nabla_\theta J^c(\theta) - \nabla_\theta J(\theta)\|_F \le \mathcal{O}\left(\sqrt{M}\left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{r_{\max}\gamma\epsilon_p}{(1-\gamma)^3}\right)\right) \triangleq \zeta_{\mathrm{het}}.$$

Thus, $\zeta_{\mathrm{het}} = \mathcal{O}(\sqrt{M}(\epsilon_r + \epsilon_p))$.

Proof. We first bound the gradient difference for an arbitrary objective $i \in [M]$ and any pair of clients c, c'. By the Policy Gradient Theorem,

$$\nabla_\theta J^{c,i}(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\pi,P_c}, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_c^{\pi,i}(s,a)],$$

where d_{π,P_c} denotes the normalized discounted state-occupancy distribution induced by π under P_c, and $Q_c^{\pi,i}$ is the action-value function for objective i at client c. The difference can be decomposed as:

$$\begin{aligned}
\|\nabla_\theta J^{c,i} - \nabla_\theta J^{c',i}\|_2 &= \frac{1}{1-\gamma}\|\mathbb{E}_{d_c}[\psi_\theta Q_c^i] - \mathbb{E}_{d_{c'}}[\psi_\theta Q_{c'}^i]\|_2 \\
&\le \frac{1}{1-\gamma}(\|\mathbb{E}_{d_c}[\psi_\theta Q_c^i] - \mathbb{E}_{d_c}[\psi_\theta Q_{c'}^i]\|_2 + \|\mathbb{E}_{d_c}[\psi_\theta Q_{c'}^i] - \mathbb{E}_{d_{c'}}[\psi_\theta Q_{c'}^i]\|_2) \\
&\le \frac{1}{1-\gamma}(\mathbb{E}_{d_c}[\|\psi_\theta\|_2 |Q_c^i - Q_{c'}^i|] + \|\psi_\theta\|_\infty 2\|Q_{c'}^i\|_\infty \|d_c - d_{c'}\|_{TV}).
\end{aligned}$$

Using the bound $\|\psi_\theta\|_2 \le C_\psi$ (Assumption 4.3), we analyze the two error terms:

1. Value Function Difference ($|Q_c^i - Q_{c'}^i|$): Using the simulation lemma (Agarwal et al., 2019), for any (s,a):

$$|Q_c^{\pi,i}(s,a) - Q_{c'}^{\pi,i}(s,a)| \le \frac{\epsilon_r}{1-\gamma} + \frac{2\gamma r_{\max}\epsilon_p}{(1-\gamma)^2}.$$

2. Distribution Difference ($\|d_c - d_{c'}\|_{TV}$): Standard perturbation bounds for Markov chains (Xu et al., 2020a) yield:

$$\|d_{\pi,P_c} - d_{\pi,P_{c'}}\|_{TV} \le \frac{\gamma\epsilon_p}{1-\gamma}.$$

Substituting these back:

$$\|\nabla_\theta J^{c,i} - \nabla_\theta J^{c',i}\|_2 \le C_\psi\left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{4\gamma r_{\max}\epsilon_p}{(1-\gamma)^3}\right)$$

Now, since

$$\nabla_\theta J^i(\theta) = \frac{1}{C}\sum_{c'=1}^{C} \nabla_\theta J^{c',i}(\theta),$$

we have

$$\begin{aligned}
\left\|\nabla_\theta J^{c,i}(\theta) - \nabla_\theta J^i(\theta)\right\|_2 &= \left\|\frac{1}{C}\sum_{c'=1}^{C}\left(\nabla_\theta J^{c,i}(\theta) - \nabla_\theta J^{c',i}(\theta)\right)\right\|_2 \\
&\le C_\psi\left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{4\gamma r_{\max}\epsilon_p}{(1-\gamma)^3}\right).
\end{aligned}$$

Finally, since the columns of the objective-gradient matrix correspond to the M objective-specific gradients,

$$\begin{aligned}\|\nabla_\theta J^c(\theta) - \nabla_\theta J(\theta)\|_F^2 &= \sum_{i=1}^{M} \left\|\nabla_\theta J^{c,i}(\theta) - \nabla_\theta J^i(\theta)\right\|_2^2 \\ &\leq M C_\psi^2 \left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{4\gamma r_{\max}\epsilon_p}{(1-\gamma)^3}\right)^2.\end{aligned}$$

Therefore,

$$\|\nabla_\theta J^c(\theta) - \nabla_\theta J(\theta)\|_F \leq C_\psi \sqrt{M} \left(\frac{\epsilon_r}{(1-\gamma)^2} + \frac{4\gamma r_{\max}\epsilon_p}{(1-\gamma)^3}\right).$$

Thus,

$$\zeta_{\text{het}} = \mathcal{O}\Big(\sqrt{M}(\epsilon_r + \epsilon_p)\Big)$$

when the remaining problem-dependent constants are suppressed. In particular, for fixed M, the gradient heterogeneity scales linearly with the transition and reward heterogeneity parameters. □