POPULATION-RELATIVE RLHF FOR LLM OUTPUT DIVERSITY: A BI-LEVEL OPTIMIZATION APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the problem of aligning large language models (LLMs) when output diversity is critical alongside output quality. Standard alignment methods (implicitly) optimize a reward model that scores individual responses in isolation, which can lead to mode collapse and overly concentrated policies. We propose *population-relative RLHF*, a framework that conditions pairwise preference feedback on a reference population of responses, enabling the reward model to capture a holistic quality-diversity trade-off. We formulate population-relative RLHF as a bi-level optimization problem, where the upper level learns a population-relative reward model and the lower level solves for the best-response LLM policy induced by that reward. We show that this formulation is an instance of a broader class of bi-level problems whose lower level constitutes a mean-field-type stochastic game. For this general problem class, we develop a first-order Hessian-free algorithm with provable finite-time convergence guarantees. We apply this algorithm to population-relative RLHF, and also introduce a simplified, memory-efficient variant suitable for large-scale implementation. Empirical evaluations demonstrate that our methods outperform prior art on the quality–diversity trade-off.

1 INTRODUCTION

Standard large language model (LLM) alignment methods such as reinforcement learning from human feedback (RLHF) and direct policy optimization (DPO) optimize a (possibly implicit) reward model, to which the aligned LLM policy is the best response (Christiano et al., 2017; Rafailov et al., 2023). The reward model scores the quality of an individual LLM output, and the optimal LLM policy in response to such a reward model can exhibit overly concentrated output distributions (Kirk et al., 2024; Padmakumar & He, 2024; Le Bronnec et al., 2024). This limits the applicability of the existing alignment methods in applications such as creative writing, mathematical reasoning, and synthetic data generation, where the policy should produce high-quality and diverse generations over repeated sampling attempts.

Current approaches to promote LLM output diversity mainly take the form of heuristic modifications to post-training or decoding, which increase diversity at the cost of substantial quality degradation and/or rely on fixed design choices (e.g., diversity metrics and thresholds), rather than adapting to human feedback. For example, the simple decoding-stage approach of increasing the temperature inflates the probability of all low-quality tokens. In the post-training stage, DivPO (Lanchantin et al., 2025) and DDPO (Chung et al., 2025) modify the standard DPO framework by re-designing how preference pairs are constructed and weighted, but the resulting trade-off is set by hand. This motivates a more principled framework that improves the quality-diversity trade-off and can learn directly from data.

In this work, we propose a *population-relative RLHF* framework that balances quality and diversity and allows for a trade-off between them through pairwise preference feedback. In standard RLHF, an annotator is presented with two LLM outputs and is asked for their preference; the pairwise data are then used to train a reward model that returns a scalar measuring the quality of an individual output. We instead consider a new paradigm in which the annotator is presented with two outputs, along with another N outputs which serve as the reference population. We query the annotator for a preference between the two candidates given the population. This allows us to model preference not only based on the absolute quality, but also on the relative diversity of each candidate over the reference population.

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.

Mathematically, this leads to a bi-level optimization problem: we learn the population-relative reward model at the upper level while solving for the induced LLM policy at the lower level. As lower-level policy optimization involves a population-dependent reward function, we recognize that it belongs to a broader class of sequential decision-making problems with mean-field interactions. We develop an algorithmic framework for this broader problem class, and then specialize it to population-relative RLHF. Our algorithm for the general problem class may be of independent interest, as it encapsulates a wider array of applications, including Stackelberg mean-field games (Cui et al., 2024; Zeng et al., 2026b) and inverse mean-field reinforcement learning (Chen et al., 2021; Cao et al., 2026).

We reiterate and further elaborate on our contributions as follows. A detailed review of the related literature is deferred to Appendix A.

Main Contributions

• We propose *population-relative RLHF*, an LLM alignment framework designed to optimize the quality-diversity trade-off through pairwise preference feedback. Each candidate response is evaluated relative to a reference population rather than in isolation. This induces a population-relative reward model. Optimizing the reward model is formulated as a bi-level optimization program.

• We recognize that this formulation can be abstracted by a broader class of bi-level optimization problems whose upper-level objective is a smooth function and whose lower-level problem is policy optimization in a mean-field-type game. For this general problem class, we develop **FD-S-MFG**, a first-order algorithm that estimates the gradient of the main decision variable through finite-differencing. We show that in the tabular setting, FD-S-MFG converges to a first-order stationary point of the bi-level objective, with a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$. To our knowledge, this is the first provably convergent algorithm (and convergence analysis) for this class of bi-level problems.

• We apply FD-S-MFG to LLM alignment, developing two algorithmic variants. The first (full gradient) variant faithfully implements FD-S-MFG, carrying out a finite-differencing step for unbiased gradient estimation. The second is a simplified variant that eliminates finite-differencing. Our experiments show that the full-gradient algorithm achieves a dominating quality-diversity Pareto frontier over prior art, while the simplified algorithm matches (and in some cases, surpasses) the full algorithm's performance with substantially reduced memory and computational overhead.

2 Formulation

We first formalize the population-relative RLHF objective for LLM alignment. Then, we introduce the bi-level optimization problem class with an upper-level smooth, possibly non-convex objective and a lower-level mean-field-type game, which captures population-relative RLHF as a special case.

2.1 Population-Relative RLHF for LLM Alignment

Let $\mathcal{L}$ denote the space of complete natural language sequences. Given a prompt $x \in \mathcal{L}$ drawn from distribution $\mathcal{D}_{\text{prompt}}$, a LLM policy $\pi(\cdot \mid x) \in \Delta_{\mathcal{L}}$ is a conditional distribution over responses $y \in \mathcal{L}$.

We consider a reward model $R_\omega : \mathcal{L} \times \mathcal{L} \times \mathcal{L}^N \to \mathbb{R}$, parameterized by $\omega \in \mathbb{R}^d$, that assigns a scalar score to a response y conditioned on prompt x and a reference population of N responses $\boldsymbol{z}_N = \{z_1, \cdots, z_N\} \in \mathcal{L}^N$. We propose learning the reward model under a special type of human feedback: an annotator is presented with two candidate responses $y_1, y_2 \in \mathcal{L}$ for prompt x, together with $\boldsymbol{z}_N$, and is asked to indicate their preference between y_1 and y_2 considering both absolute response quality and the relative diversity/novelty over $\boldsymbol{z}_N$. We model this preference distribution using an extended Bradley-Terry model: the probability that y_1 is preferred over y_2 given $\boldsymbol{z}_N$ is

$$P_\omega(y_1 \succ y_2 \mid x, \boldsymbol{z}_N) = \frac{\exp(R_\omega(y_1 \mid x, \boldsymbol{z}_N))}{\exp(R_\omega(y_1 \mid x, \boldsymbol{z}_N)) + \exp(R_\omega(y_2 \mid x, \boldsymbol{z}_N))}.$$

Let $D(x, y_1, y_2, \boldsymbol{z}_N)$ denote the distribution of human preference between y_1, y_2 given x and $\boldsymbol{z}_N$. The binary preference label $p \sim D(x, y_1, y_2, \boldsymbol{z}_N)$ equals 1 if y_1 is preferred over y_2, and 0 otherwise. When $y_1, y_2, \boldsymbol{z}_N$ are generated by a fixed policy π, the objective of the reward model is the following cross entropy loss (for brevity we omit the expectation over $x \sim \mathcal{D}_{\text{prompt}}$ and $p \sim D(x, y_1, y_2, \boldsymbol{z}_N)$)

$$\ell(\omega, \pi) \triangleq -\mathbb{E}_{(y_1, y_2, \boldsymbol{z}_N) \sim \pi(\cdot|x)}[\, p \log P_\omega(y_1 \succ y_2 | x, \boldsymbol{z}_N) + (1-p) \log P_\omega(y_1 \prec y_2 | x, \boldsymbol{z}_N)]. \quad (1)$$

Ideally, we would like to optimize ω not simply under an offline LLM, but under an LLM that itself adapts to the reward model. Suppose that the reference population is generated by a policy π'. Given a fixed π' and reward model parameter ω, to optimize the LLM policy π, we consider maximizing R_ω subject to a KL regularization relative to an anchor policy π_{KL} (weighted by a constant $\tau > 0$)

$$J^{\omega,\pi'}_{\mathrm{Population}}(\pi) \triangleq \mathbb{E}_{x\sim D_{\mathrm{prompt}}, y\sim\pi(\cdot|x), \boldsymbol{z}_N \overset{\text{i.i.d.}}{\sim} \pi'(\cdot|x)} \left[R_\omega(y|x, \boldsymbol{z}_N) - \tau \log\left(\frac{\pi(y|x)}{\pi_{\mathrm{KL}}(y|x)}\right)\right]. \quad (2)$$

Combining (1) and (2), we formulate the following bi-level problem, in which we optimize the population-relative reward model while ensuring that the LLM policy remains optimal in response. The goal at the lower level is to find an equilibrium policy $\pi^\star(\omega)$ induced by ω that satisfies a self-consistency equation: the policy $\pi^\star(\omega)$ must maximize $J_{\mathrm{Population}}$ when the reference population $\boldsymbol{z}_N$ itself is generated by $\pi^\star(\omega)$. A sufficiently large regularization weight τ guarantees that $\pi^\star(\omega)$ is unique for any ω, ensuring that the objective in (3) is well-defined.

$$\min_{\omega} \ \ell(\omega, \pi^\star(\omega)) \quad \text{s.t. } \pi^\star(\omega) \in \operatorname*{argmax}_{\pi} J^{\omega,\pi^\star(\omega)}_{\mathrm{Population}}(\pi). \quad (3)$$

Note that in the degenerate case where the reward function does not depend on $\boldsymbol{z}_N$, (3) reduces to the regular bi-level RLHF formulation introduced in Chakraborty et al. (2024).

2.2 Bi-Level Optimization Framework

We formalize the general bi-level framework that abstracts (2) as a special case. The generality arises in two aspects. First, the upper-level objective can be a general smooth, non-convex function of ω and the equilibrium policy, and the reward r_ω may depend on the reference policy arbitrarily rather than necessarily through N samples. Second, the lower-level problem generalizes from a static game to a sequential, stochastic game with mean-field interactions.

Consider an infinite-horizon discounted-reward mean-field-type game with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $\mathcal{P} : \mathcal{S}\times\mathcal{A} \to \Delta_{\mathcal{S}}$, discount factor $\gamma \in (0,1)$, and initial state distribution $\rho \in \Delta_{\mathcal{S}}$. The reward function $r_\omega(s, a \mid \pi') \in [0,1]$ depends on the exogenous control parameter ω, the state-action pair (s,a), and a reference policy π'. Let $d^\pi_\rho \triangleq (1-\gamma)\sum_{k=0}^\infty \gamma^k \mathbb{P}(s_k = s \mid s_0 \sim \rho, \pi) \in \Delta_{\mathcal{S}}$ denote the discounted occupancy measure. Under fixed ω and π', the decision maker's objective is the entropy-regularized cumulative reward

$$J_\omega(\pi, \pi') \triangleq \mathbb{E}_{a_k\sim\pi(\cdot|s_k), s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k)} \left[\sum_{k=0}^\infty \gamma^k r_\omega(s_k, a_k \mid \pi') - \tau \log \pi(a_k \mid s_k) \mid s_0 \sim \rho\right]. \quad (4)$$

Let the upper-level objective be $f(\omega,\pi) \triangleq \mathbb{E}_{\xi\sim\mu}[f(\omega,\pi,\xi)]$ where ξ represents the source of stochasticity. Our overall bi-level problem is to solve

$$\min_{\omega} \ F(\omega) \triangleq f(\omega, \pi^\star(\omega)) \qquad \text{s.t. } \pi^\star(\omega) \in \operatorname*{argmax}_{\pi} J_\omega(\pi, \pi^\star(\omega)). \quad (5)$$

Under continuity conditions that we impose later, the fixed point $\pi^\star(\omega)$ always exists. We further assume that the regularization weight τ is large enough that $\pi^\star(\omega)$ is unique for any ω.

The population-relative RLHF objective in (2) is a single-stage instance of (5), with the following map: the state s collapses to input prompt x and action a to response y; the mean-field reward function absorbs the logarithm anchor policy term, yielding $r_\omega(s, a \mid \pi') = \mathbb{E}_{\boldsymbol{z}_N\sim\pi'(\cdot|x)}[R_\omega(y \mid x, \boldsymbol{z}_N)] + \tau \log \pi_{\mathrm{KL}}(y \mid x)$; the upper-level objective $f(\omega, \pi^\star(\omega))$ corresponds to the cross-entropy loss.

In our mathematical analysis, we parameterize the policy by a parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ through the softmax function $\pi_\theta(a \mid s) = \exp(\theta_{s,a}) / \sum_{a'\in\mathcal{A}} \exp(\theta_{s,a'})$. For unique identifiability, we use $\theta^\star(\omega)$ to denote the softmax parameter of $\pi^\star(\omega)$ satisfying $\sum_a \theta^\star_{s,a}(\omega) = c$ for all $s \in \mathcal{S}$, where c is a constant. We note that tabular softmax parameterization is introduced only for the purpose of theoretical analysis in Section 4. In practice, π_θ can adopt any parameterization, including neural networks.

3 Algorithm Development

In this section, we develop **FD-S-MFG**, a **F**inite-**D**ifference-based algorithm for bi-level optimization with a **S**mooth upper-level objective and lower-level **M**ean-**F**ield-type **G**ame. We first derive the

hypergradient of the bi-level objective and identify its computational bottleneck, which lies in the evaluation and inversion of second-order derivatives. We bypass the challenge via a Hessian-free stochastic approximation scheme, with finite-differencing along adaptive perturbation directions. We specialize FD-S-MFG to the context of population-relative RLHF in Section 3.1 and introduce a simplified, memory-efficient variant of the algorithm in Section 3.2.

For algorithmic development and analysis, we impose a negative-definiteness condition on the second-order operator of the lower-level game. Similar assumptions are frequently considered in bi-level optimization with a non-convex lower-level objective (Huang, 2024; Zeng et al., 2026a).

Assumption 1 *There exists a constant $\underline{\lambda} > 0$ such that for any ω*

$$\lambda_{\min}\Big(- \nabla^2_{1,1} J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}) - \nabla^2_{2,1} J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)})\Big) \geq \underline{\lambda},$$

where we use $\lambda_{\min}(M)$ to denote the smallest eigenvalue of the symmetric part of a matrix M.

Optimizing (5) with gradient descent requires evaluating the hypergradient $\nabla F(\omega)$. The lemma below derives a closed-form expression for $\nabla F(\omega)$, generalizing Pedregosa (2016) where the lower-level problem is an optimization program. We adopt the notations $\nabla^2_{1,1} J_\omega(\pi_\theta, \pi_{\theta'}) = \frac{\partial}{\partial \theta}(\nabla_\theta J_\omega(\pi_\theta, \pi_{\theta'}))$, $\nabla^2_{2,1} J_\omega(\pi_\theta, \pi_{\theta'}) = \frac{\partial}{\partial \theta}(\nabla_{\theta'} J_\omega(\pi_\theta, \pi_{\theta'}))$, and $\nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi_{\theta'}) = \frac{\partial}{\partial \theta}(\nabla_\omega J_\omega(\pi_\theta, \pi_{\theta'}))$.

Lemma 1 (Hypergradient) *Suppose $\pi^\star(\omega)$ is in the interior of the probability simplex. If $\nabla^2_{1,1} J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}) + \nabla^2_{2,1} J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)})$ is invertible, then*

$$\begin{aligned}\nabla F(\omega) &= \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) + \left(\frac{\partial \theta^\star(\omega)}{\partial \omega}\right)^\top \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}) \\ &= \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) - \nabla^2_{\omega,1} J_\omega \cdot (\nabla^2_{1,1} J_\omega + \nabla^2_{2,1} J_\omega)^{-1} \cdot \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}), \end{aligned} \tag{6}$$

where all derivatives of J_ω are evaluated at $(\theta^\star(\omega), \theta^\star(\omega))$.

Formally stated in Algorithm 1, **FD-S-MFG** is a stochastic gradient descent algorithm that iteratively updates the decision variable ω_k in the approximate hypergradient direction $\nabla F(\omega_k)$. However, directly computing $\nabla F(\omega_k)$ according to (6) is costly as the expression involves the second-order Jacobian matrix $\nabla^2_{\omega,1} J_\omega$, as well as the inverse of the Hessian $(\nabla^2_{1,1} J_\omega + \nabla^2_{2,1} J_\omega)^{-1}$. To overcome this computational challenge, our algorithm development proceeds in the two steps below.

Algorithm 1 **F**inite-**D**ifference Algorithm for Bi-Level Optimization with **S**mooth Upper-Level Objective and Lower-Level **M**ean-**F**ield-Type **G**ame (**FD-S-MFG**)

1: **Initialize:** Control variable ω_0, auxiliary variable $u_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.
2: **for** iteration $k = 0, 1, 2, \ldots$ **do**
3: Query a MFG policy optimization oracle for θ_k such that

$$\|\theta_k - \theta^\star(\omega_k)\|^2 \leq \epsilon_{\mathrm{MFG}}. \tag{7}$$

4: Set (perturbed) policies $\pi_k = \pi_{\theta_k}$, $\pi_k^- = \pi_{\theta_k - \varepsilon_k u_k}$, $\pi_k^+ = \pi_{\theta_k + \varepsilon_k u_k}$.
5: Query a policy evaluation oracle for $A_k^-, A_k^+ \in [-B_V, B_V]^{|\mathcal{S}||\mathcal{A}|}$ such that

$$\|A_k^- - A^{\omega_k, \pi_k^-, \pi_k}\|^2 \leq \epsilon_{\mathrm{MFG}}, \quad \|A_k^+ - A^{\omega_k, \pi_k^+, \pi_k}\|^2 \leq \epsilon_{\mathrm{MFG}}, \tag{8}$$

where $A^{\omega, \pi, \pi'} \in [-B_V, B_V]^{|\mathcal{S}||\mathcal{A}|}$, defined in (18), is the advantage function under reward r_ω and policy π with reference population generated by π', and B_V is a constant defined in (19).
6: Sample state $s_k^- \sim d_\rho^{\pi_k^-}$ and take action $a_k^- \sim \pi_k^-(\cdot \mid s_k^-)$. Sample state $s_k^+ \sim d_\rho^{\pi_k^+}$ and take action $a_k^+ \sim \pi_k^+(\cdot \mid s_k^+)$. Sample $\xi_k \sim \mu$.
7: Update control variable ω_k, with $\hat{G}_k$ defined in (14)

$$\omega_{k+1} = \omega_k - \alpha_k(\nabla_\omega f(\omega_k, \theta_k, \xi_k) - \hat{G}_k). \tag{9}$$

8: Update auxiliary variable u_k, with $\hat{H}_k$ defined in (14)

$$u_{k+1} = \mathrm{Proj}_{\|\cdot\| \leq \frac{L_f}{\underline{\lambda}}}\Big(u_k + \eta_k\big(\hat{H}_k - \nabla_\theta f(\omega_k, \theta_k, \xi_k)\big)\Big). \tag{10}$$

9: **end for**

Bypassing Explicit Matrix Inversion. Suppose for a moment that the Hessians $\nabla^2_{1,1} J_\omega$ and $\nabla^2_{2,1} J_\omega$ could be accurately evaluated – then, our sole objective is to avoid explicit matrix inversion. Observe

that the hypergradient ∇_F depends on the Hessian only through the inverse-Hessian-vector product $(\nabla^2_{1,1}J_\omega + \nabla^2_{2,1}J_\omega)^{-1}\nabla_\theta f(\omega, \pi_{\theta^\star(\omega)})$. Following Yang et al. (2023); Huang (2024), we view the product as the unique solution $u^\star(\omega, \theta^\star(\omega))$ to the linear system

$$\left(\nabla^2_{1,1}J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}) + \nabla^2_{2,1}J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)})\right) u = \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}). \tag{11}$$

We can solve (11) via stochastic approximation without matrix inversion. Specifically, under a suitable step size (and Assumption 1), if we maintain a variable u_k and iteratively update it as

$$u_{k+1}=u_k+\eta_k\Big((\nabla^2_{1,1}J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)})+\nabla^2_{2,1}J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}))u_k-\nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}, \xi)\Big), \tag{12}$$

then u_k efficiently converges to $u^\star(\omega, \theta^\star(\omega))$, and we can approximate the hypergradient in (6) as

$$\nabla F(\omega) \approx \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) - \nabla^2_{\omega,1}J_\omega(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)})u_k. \tag{13}$$

In practice, we cannot perform the update (12) exactly, as we do not have access to the Hessian matrices. In Eq. (10) of Algorithm 1, we replace exact evaluation of $(\nabla^2_{1,1}J_\omega(\pi,\pi) + \nabla^2_{2,1}J_\omega(\pi,\pi))u_k$ with stochastic, finite-difference-based estimates computed as follows.

Finite-Difference-Based Matrix-Vector Product Approximation. To execute (12) and to calculate hypergradient according to (13) requires computing matrix-vector products. Our goal is to avoid constructing the matrices explicitly due to the memory cost. To design a strictly first-order algorithm, we approximate these matrix-vector products using finite-difference operators along the directional perturbation u_k. In general, for any differentiable vector-valued function $g(x)$ with Jacobian matrix $M = \nabla_x g(x)$, the matrix-vector product Mu can be expressed via the central finite-difference limit

$$Mu = \lim_{\varepsilon\to 0}\frac{g(x+\varepsilon u) - g(x-\varepsilon u)}{2\varepsilon}.$$

Applying this principle, we design stochastic operators $\hat{H}_{1,k}, \hat{H}_{2,k}, \hat{H}_k, \hat{G}_k$ with perturbation strength $\varepsilon_k > 0$. The operators are defined by advantage function estimates A^+_k, A^-_k, state-action samples $s^+_k, a^+_k, s^-_k, a^-_k$, and policy parameter θ_k which approximates $\theta^\star(\omega_k)$, all formally introduced in Algorithm 1. We plug $\hat{H}_k, \hat{G}_k$ into the algorithm updates in (9) and (10).

$$\hat{H}_{1,k} \triangleq \frac{A^+_k(s^+_k, a^+_k)\nabla_\theta \log \pi_{\theta_k+\epsilon_k u_k}(a^+_k \mid s^+_k) - A^-_k(s^-_k, a^-_k)\nabla_\theta \log \pi_{\theta_k-\epsilon_k u_k}(a^-_k \mid s^-_k)}{2(1-\gamma)\varepsilon_k},$$

$$\hat{H}_{2,k} \triangleq \frac{\nabla_\theta r_{\omega_k}(s^+_k, a^+_k \mid \pi_{\theta_k}) - \nabla_\theta r_{\omega_k}(s^-_k, a^-_k \mid \pi_{\theta_k})}{2(1-\gamma)\varepsilon_k},$$

$$\hat{H}_k \triangleq \hat{H}_{1,k} + \hat{H}_{2,k}, \qquad \hat{G}_k \triangleq \frac{\nabla_\omega r_{\omega_k}(s^+_k, a^+_k \mid \pi_{\theta_k}) - \nabla_\omega r_{\omega_k}(s^-_k, a^-_k \mid \pi_{\theta_k})}{2(1-\gamma)\varepsilon_k}. \tag{14}$$

As $\varepsilon_k \to 0$, $\hat{H}_{1,k}$ serves as an asymptotically unbiased estimator for $\nabla^2_{1,1}J_{\omega_k} \cdot u_k$, while $\hat{H}_{2,k}$ estimates $\nabla^2_{2,1}J_{\omega_k} \cdot u_k$. The Jacobian-vector product $\nabla^2_{\omega,1}J_{\omega_k} \cdot u_k$ is approximated by $\hat{G}_k$.

3.1 Instantiation to Population-Relative RLHF

We now specialize FD-S-MFG to population-relative RLHF, where we make the following practical adaptations. The full algorithm is presented in Algorithm 2.

- **Variance Control:** Recall from Algorithm 1 that $\hat{G}_k$ is a finite-difference-based stochastic estimate of the hypergradient correction term. Due to the $1/\varepsilon_k$ scaling factor, $\hat{G}_k$ may exhibit large variance spikes in practice. To prevent extreme stochastic perturbations from destabilizing the reward model update, we rescale $\hat{G}_k$ whenever its norm exceeds a threshold relative to the direct partial gradient $\|\nabla_\omega L_{\mathrm{RM}}(\omega_{k,t})\|$. This corresponds to Eq. (15) in Algorithm 2.
- **Reference Population Gradient Approximation:** To compute $\hat{H}_{2,k}$ in Algorithm 1, we assume being able to query $\nabla_\theta r_\omega(s, a \mid \pi_\theta)$. In language modeling, the reward depends on the reference policy parameter θ through the sampling distribution of the population $\mathbf{z}_N \sim \pi_\theta(\cdot \mid x)$. By the policy gradient theorem, this derivative evaluates to $\nabla_\theta \mathbb{E}_{\mathbf{z}_N\sim\pi_\theta}[R_\omega(y \mid x, \mathbf{z}_N)] = \mathbb{E}_{\mathbf{z}_N\sim\pi_\theta}[R_\omega(y \mid x, \mathbf{z}_N)\sum_{i=1}^N \nabla_\theta \log \pi_\theta(z_i \mid x)]$. This leads to Eq. (16) of Algorithm 2, in which we approximate $\nabla_\theta r_\omega(s, a \mid \pi_\theta)$ with its unbiased stochastic estimate.

- **Inner-Loop Subroutine:** We replace the oracle query for $\theta^\star(\omega_k)$ (to precision ϵ_{MFG}) with T inner-loop iterations of a policy optimization algorithm of choice (e.g., PPO, GRPO) executed against the fixed reward model R_{ω_k}. The reference population is dynamically sampled from the active policy itself, which allows the policy and population to evolve jointly toward a fixed point.

At each outer iteration k, Algorithm 2 alternates between two steps. First, holding the LLM policy fixed to π_k, we update the reward model over T steps using preference labels for candidate pairs $(y_1, y_2) \sim \pi_k(\cdot \mid x)$ conditioned on the reference population $\boldsymbol{z}_N \sim \pi_k(\cdot \mid x)$. Together with the reward model update, the auxiliary variable u_k is iteratively refined to provide an increasingly accurate estimate of $\hat{G}_k$. Second, we hold the reward model fixed and optimize the LLM policy parameter θ over T steps using advantage functions derived from the reward model R_{ω_k}.

3.2 Memory-Efficient Simplification

While Algorithm 2 is completely sample-based and first-order, it requires maintaining three LLM policy replicas (π_k and its two perturbed versions). This can incur significant memory overhead when fine-tuning multi-billion parameter LLMs. Furthermore, Algorithm 2 introduces perturbation weight ε_k and auxiliary learning rate η_k, which can be difficult to tune in practice. In the context of population-relative RLHF, we observe that $\pi^\star(\omega)$ enters the upper-level objective (i.e., the cross entropy loss in (1)) solely through the sampling distribution. Since sampling distribution shifts have a comparatively small effect on the expected cross entropy loss, it is natural to expect that $\nabla_\theta f$ is dominated by the direct partial gradient $\nabla_\omega f$, implying $\nabla F(\omega) \approx \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)})$. These factors motivate us to introduce a simplified variant of Algorithm 2, in which we drop the hypergradient correction term $\hat{G}_k$. The grayed lines in Algorithm 2 highlight the steps omitted for this simplification. For readability, the simplified pipeline is detailed without the hypergradient components in the appendix (Algorithm 3).

We view both Algorithm 2 and 3 as valuable tools offering a trade-off depending on computational constraints and demands for theoretical rigor. Algorithm 2 provides unbiased hypergradient estimation and stronger convergence guarantees, whereas the simplified variant offers a memory-efficient, easily implementable alternative. Our empirical evaluations later show that the simplified algorithm consistently matches the performance of the full algorithm, and sometimes even performs better, likely because omitting the finite-difference step substantially reduces gradient variance.

4 Finite-Time Analysis

In this section we establish the complexity of Algorithm 1. We start by introducing the assumptions.

Assumption 2 *The function f is differentiable, and we can query stochastic gradient operators $\nabla_\omega f(\omega, \pi_\theta, \xi), \nabla_\theta f(\omega, \pi_\theta, \xi)$ such that for all ω, θ*

$$\mathbb{E}_{\xi\sim\mu}[\nabla_\omega f(\omega, \pi_\theta, \xi)] = \nabla_\omega f(\omega, \pi_\theta), \quad \mathbb{E}_{\xi\sim\mu}[\nabla_\theta f(\omega, \pi_\theta, \xi)] = \nabla_\theta f(\omega, \pi_\theta).$$

In addition, there exists a constant L_f such that $\forall \omega, \omega', \theta, \theta', \xi$

$$\|\nabla_\omega f(\omega, \pi_\theta, \xi)\| \le L_f, \quad \|\nabla_\omega f(\omega, \pi_\theta, \xi) - \nabla_\omega f(\omega', \pi_{\theta'}, \xi)\| \le L_f(\|\omega - \omega'\| + \|\theta - \theta'\|)$$
$$\|\nabla_\theta f(\omega, \pi_\theta, \xi)\| \le L_f, \quad \|\nabla_\theta f(\omega, \pi_\theta, \xi) - \nabla_\theta f(\omega', \pi_{\theta'}, \xi)\| \le L_f(\|\omega - \omega'\| + \|\theta - \theta'\|).$$

Assumption 2 imposes standard regularity conditions common in the literature on bi-level optimization (Kwon et al., 2023; Shen & Chen, 2023; Chen et al., 2024). Note that Assumption 2 is satisfied in the population-relative RLHF setting, given that the reward $R_\omega(y \mid x, \boldsymbol{z}_N)$ is Lipschitz and smooth in ω and the policy π_θ is smoothly parameterized in θ with bounded gradients.

Assumption 3 *There exists a bounded constant L_r such that for all $s, a, \omega_1, \omega_2, \pi_1', \pi_2'$*

$$|r_{\omega_1}(s, a \mid \pi_1') - r_{\omega_2}(s, a \mid \pi_2')| \le L_r(\|\omega_1 - \omega_2\| + \|\pi_1' - \pi_2'\|),$$
$$\|\nabla_\omega r_{\omega_1}(s, a \mid \pi_1') - \nabla_\omega r_{\omega_2}(s, a \mid \pi_2')\| \le L_r(\|\omega_1 - \omega_2\| + \|\pi_1' - \pi_2'\|),$$
$$\|\nabla_\pi r_{\omega_1}(s, a \mid \pi_1') - \nabla_\pi r_{\omega_2}(s, a \mid \pi_2')\| \le L_r(\|\omega_1 - \omega_2\| + \|\pi_1' - \pi_2'\|).$$

Assumption 3 requires the reward function and its derivatives to be Lipschitz in the control parameter and reference policy, which are again common in the literature on (hierarchical) mean-field games (Cui et al., 2024; Zhang et al., 2025; Zeng et al., 2026b).

Algorithm 2 Population-Relative Reinforcement Learning with Human Feedback

Require: Prompt distribution $\mathcal{D}_{\text{prompt}}$, population size N, inner-loop iterations T, step sizes $\alpha_k, \beta_k, \eta_k$, perturbation scale $\varepsilon_k > 0$, KL coefficient $\tau > 0$, anchor policy π_{KL}.
Initialize: Reward model parameter ω_0, LLM parameter θ_0 for policy π_0, auxiliary variable u_0.
for iteration $k = 0, 1, 2, \ldots$ **do**

Set $\omega_{k,0} = \omega_k$ and $u_{k,0} = u_k$. **Reward Model & Auxiliary Variable Update**
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (batch) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample population $\boldsymbol{z}_N = \{z_1, \ldots, z_N\}$ and candidates (y_1, y_2) i.i.d. from $\pi_k(\cdot \mid x)$.
  Sample perturbed candidates $y^+ \sim \pi_{\theta_k + \varepsilon_k u_k}(\cdot|x)$ and $y^- \sim \pi_{\theta_k - \varepsilon_k u_k}(\cdot|x)$. **[Full Grad]**
  Query human preference

$$p = \begin{cases} 1, & \text{if } y_1 \text{ is preferred over } y_2 \text{ given } \boldsymbol{z}_N, \\ 0, & \text{otherwise.} \end{cases}$$

  Compute direct reward model loss under Bradley-Terry model

$$L_{\text{RM}}(\omega_{k,t}) = \text{BCE}\big(\text{sigmoid}(R_{\omega_{k,t}}(y_1 \mid x, \boldsymbol{z}_N) - R_{\omega_{k,t}}(y_2 \mid x, \boldsymbol{z}_N)),\, p\big).$$

  Compute finite-difference Jacobian-vector estimate $\hat{G}_k$ **[Full Grad]**

$$\hat{G}_k = \frac{1}{2\varepsilon_k}\Big(\nabla_\omega R_{\omega_{k,t}}(y^+ \mid x, \boldsymbol{z}_N) - \nabla_\omega R_{\omega_{k,t}}(y^- \mid x, \boldsymbol{z}_N)\Big).$$

  Rescale high-variance correction term $\hat{G}_k$ **[Full Grad]**

$$\hat{G}_k = \begin{cases} \dfrac{\kappa \|\nabla_\omega L_{\text{RM}}(\omega_{k,t})\|}{\|\hat{G}_k\|}\, \hat{G}_k, & \text{if } \|\hat{G}_k\| > \kappa \|\nabla_\omega L_{\text{RM}}(\omega_{k,t})\|, \\ \hat{G}_k, & \text{otherwise.} \end{cases} \tag{15}$$

  Compute finite-difference Hessian-vector estimate $\hat{H}_k = \hat{H}_{1,k} + \hat{H}_{2,k}$ **[Full Grad]**

$$\hat{H}_{1,k} = \frac{R_{\omega_k}(y^+|x,\boldsymbol{z}_N)\nabla_\theta \log \pi_{\theta_k+\varepsilon_k u_k}(y^+|x) - R_{\omega_k}(y^-|x,\boldsymbol{z}_N)\nabla_\theta \log \pi_{\theta_k-\varepsilon_k u_k}(y^-|x)}{2\varepsilon_k}$$

$$\hat{H}_{2,k} = \frac{R_{\omega_k}(y^+ \mid x, \boldsymbol{z}_N) - R_{\omega_k}(y^- \mid x, \boldsymbol{z}_N)}{2\varepsilon_k} \sum_{i=1}^N \nabla_\theta \log \pi_{\theta_k}(z_i \mid x). \tag{16}$$

  Update reward parameter ω (and auxiliary variable u)

$$\omega_{k,t+1} = \omega_{k,t} - \alpha_k\big(\nabla_\omega L_{\text{RM}}(\omega_{k,t}) - \hat{G}_k\big),$$

$$u_{k,t+1} = u_{k,t} + \eta_k\big(\hat{H}_k - \nabla_\theta L_{\text{RM}}(\omega_{k,t}, \theta_k)\big). \quad \textbf{[Full Grad]}$$

end for
Set $\omega_{k+1} = \omega_{k,T}$ and $u_{k+1} = u_{k,T}$.

Set $\theta_{k,0} = \theta_k$ and $\pi_{k,0} = \pi_k$. **Language Model Update**
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (a batch of) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample population $\boldsymbol{z}_N = \{z_1, \ldots, z_N\}$ and candidate y i.i.d. from $\pi_{k,t}(\cdot \mid x)$.
  Score candidate with reward model to obtain $R_{\omega_k}(y \mid x, \boldsymbol{z}_N)$.
  Estimate advantage function A from $R_{\omega_k}(y \mid x, \boldsymbol{z}_N)$.
  Compute (token-level) KL divergence to anchor policy

$$\text{KL} = \log \pi_{k,t}(y \mid x) - \log \pi_{\text{KL}}(y \mid x).$$

  Update LLM parameter

$$\theta_{k,t+1} = \theta_{k,t} + \beta_k\big(A - \tau \text{KL}\big)\nabla_\theta \log \pi_\theta(y \mid x).$$

end for
Set $\theta_{k+1} = \theta_{k,T}$ and $\pi_{k+1} = \pi_{k,T}$.

end for

Theorem 1 *Under Assumptions 1-3 and properly selected parameters $\alpha_k, \eta_k, \varepsilon_k$ (their exact choices are specified in Appendix F), the iterates of Algorithm 1 satisfy for all k*

$$\min_{k\le K} \mathbb{E}[\|\nabla F(\omega_k)\|^2] \le \mathcal{O}\big(\epsilon_{\mathrm{MFG}}(K+1)^{1/5}\big) + \tilde{\mathcal{O}}\big((K+1)^{-2/5}\big).$$

Theorem 1 establishes the convergence of the best iterate of Algorithm 1 to a stationary point of the bi-level objective. To achieve $\min_{k\le K} \mathbb{E}[\|\nabla F(\omega_k)\|^2] \le \epsilon$ for a desired precision ϵ, we need $K = \mathcal{O}(\epsilon^{-5/2})$, $\epsilon_{\mathrm{MFG}} = \mathcal{O}(\epsilon^{3/2})$. In the tabular case, under proper assumptions (i.e., sufficiently large regularization τ), the state-of-the-art policy optimization method for MFGs reaches the precision ϵ_{MFG} with $\widetilde{\mathcal{O}}(\epsilon_{\mathrm{MFG}}^{-1})$ samples (Yardim et al., 2023). This implies that Algorithm 1 can achieve an overall sample complexity of $\mathcal{O}\left(K\epsilon_{\mathrm{MFG}}^{-1}\right) = \widetilde{\mathcal{O}}(\epsilon^{-4})$ (measured by the squared gradient norm).

Theorem 1 establishes the first finite-time and finite-sample complexity bound for a bi-level optimization algorithm targeting a lower-level regularized mean-field-type game. The design of our algorithm draws inspiration from Huang (2024), but we make three distinct contributions. First, while Huang (2024) considers a static lower-level optimization problem, we consider a sequential mean-field-type game where the policy must reach a self-consistent equilibrium. Second, Huang (2024) assumes deterministic gradient access, allowing the perturbation radius ε_k to be driven arbitrarily small. In contrast, our algorithm operates in the stochastic setting, which necessitates a careful choice of ε_k against the step sizes via a multi-timescale analysis to balance finite-difference approximation bias with variance. Finally, the algorithm in Huang (2024) carries out a practically infeasible projection step, which we completely eliminate. A detailed discussion on the distinction is in Appendix D.

5 Experimental Results

We empirically evaluate the proposed algorithm on two tasks. First, we conduct a controlled experiment on happy tweet generation, where the language model is small enough for us to implement both the full hypergradient and simplified variants of population-relative RLHF and to analyze their quality-diversity trade-offs across varying population sizes N. Subsequently, we scale population-relative RLHF (simplified) to a larger language model on a benchmark for creative writing.

Baselines. We compare against state-of-the-art diversity-promoting post-training methods, DDPO (Chung et al., 2025), DivPO (Lanchantin et al., 2025), and CD-RLHF (Sun et al., 2025). DDPO and DivPO build on the standard DPO framework. The former re-weights the DPO loss by a deviation score, whereas the latter constructs preference pairs differently through reward thresholding. CD-RLHF instead builds on the standard RLHF framework, augmenting the conventional reward with a learned intrinsic reward derived from state-prediction error to encourage exploration of less familiar generations. Note that DDPO was initially proposed for the off-policy setting, where pairwise responses are sampled from a fixed dataset. For fair and consistent comparison, we adapt DDPO to the on-policy setting and generate pairwise responses from the policy being optimized.

We also compare against standard RLHF and DPO (implemented in an online, alternating fashion for direct comparison with Algorithm 2). These methods do not account for diversity but theoretically set an upper bound on the quality score.

Finally, we introduce *Set RLHF*, a novel algorithm that we design by extending the recently proposed Set RL method (Hamid et al., 2026) to the RLHF setting. Set RLHF relies on preference feedback between two candidate sets of responses and optimizes a set-level reward model. We present the algorithm details and discuss its potential drawback relative to population-relative RLHF in Appendix B.

We defer details of the experimental setup and additional results to Appendix C. A Python implementation of both experiments is provided as part of the supplementary material.

5.1 Happy Tweet Generation

Our aim in the experiment is to train a tiny language model to generate tweets with a positive sentiment, while maintaining output diversity across repeated samples, under pairwise "human preference feedback". The preference feedback is simulated via the "ground-truth" reward function

$$R_{\mathrm{truth}}(y \mid \boldsymbol{z}_N) = (1-\delta)\cdot R_{\mathrm{quality}}(y) + \delta \cdot R_{\mathrm{div}}(\{y\} \cup \boldsymbol{z}_N). \tag{17}$$

The parameter $\delta \in [0,1]$ controls the trade-off between response quality and population diversity. Here, $R_{\mathrm{quality}}(y) \in [-1,1]$ is a sentiment score given by a rule-based sentiment analyzer (Hutto & Gilbert, 2014), and $R_{\mathrm{div}}(\boldsymbol{z})$ measures the diversity of a set $\boldsymbol{z}$ via word-level Jaccard distance. Preference between two responses y_1, y_2 is determined by whether $R_{\mathrm{truth}}(y_1 \mid \boldsymbol{z}_N)$ exceeds

$R_{\text{truth}}(y_2 \mid \boldsymbol{z}_N)$. Note that we never expose the ground-truth rewards to the learning algorithms; instead, they are used solely to generate preferences $p \in \{0, 1\}$, which are then passed to the algorithms.

In Figure 1, we visualize the quality-diversity trade-off under population sizes $N = 2, 5, 8$. Across all N, population-relative RLHF consistently outperforms all other diversification methods, establishing a dominating Pareto frontier. The simplified population-relative RLHF algorithm closely tracks the performance of the full algorithm, achieving an even better Pareto frontier in the case of $N = 8$.

To numerically compare the performance of population-relative RLHF and set RLHF across N, we show in Table 1 the areas under the curves (AUCs)[1]. Population-relative RLHF's AUCs expand as N increases from 2 to 8, indicating that a larger reference population leads to an improved trade-off. However, even with $N = 2$, population-relative RLHF already shows a highly competitive AUC. This observation suggests that in real-world scenarios where collecting human annotations over a large population is expensive, setting $N = 2$ provides a cost-effective choice without incurring heavy annotation overhead. In contrast, the sentiment score achieved by Set RLHF unsurprisingly suffers substantial degradation as N increases – since Set RLHF assigns a single scalar reward to an entire set of responses, it becomes increasingly difficult to attribute quality improvements to specific responses within the set. In Appendix C.1, we include additional figures showing the learning dynamics.

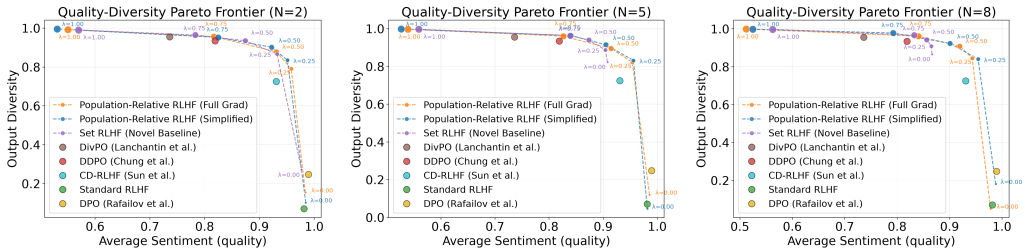

Figure 1: Happy tweet: quality-diversity trade-off, averaged across 5 seeds (Left: $N = 2$. Middle: $N = 5$. Right: $N = 8$). Averaged performance with confidence intervals tabulated in Appendix C.1.

5.2 LitBench

We scale our framework to fine-tune a Llama 8B model (via LoRA (Hu et al., 2021)) to generate creative narratives. We again simulate "ground-truth human feedback" through reward R_{truth} as decomposed in (17). Here R_{quality} is the writing quality score, given by an LLM evaluator provided in Fein et al. (2026). Diversity R_{div} is measured by the marginal Vendi score contribution of candidate y to the reference population $\boldsymbol{z}_N$ in sentence-embedding space. Motivated by the strong performance of simplified population-relative RLHF and constrained by computational resources, we evaluate the simplified population-relative RLHF algorithm only in this larger-scale experiment.

Figure 2 shows that population-relative RLHF (simplified) again establishes a superior Pareto frontier compared to standard RLHF and DPO, which collapse output diversity, and DivPO and CD-RLHF, which preserves diversity at a higher cost of quality degradation. DDPO falls outside the Pareto frontier of population-relative RLHF but substantially underperforms in terms of diversity.

Table 1: Happy tweet: Pareto frontier AUC across set sizes N (5 seeds, 95% confidence interval).

	$N = 2$	$N = 5$	$N = 8$
Population-Relative RLHF (Full Grad)	0.400 ± 0.049	0.412 ± 0.053	0.435 ± 0.015
Population-Relative RLHF (Simplified)	0.418 ± 0.036	0.424 ± 0.024	0.432 ± 0.016
Set RLHF	0.373 ± 0.021	0.345 ± 0.014	0.302 ± 0.016

$$\text{AUC} = \int \text{diversity}\, d(\text{sentiment}).$$

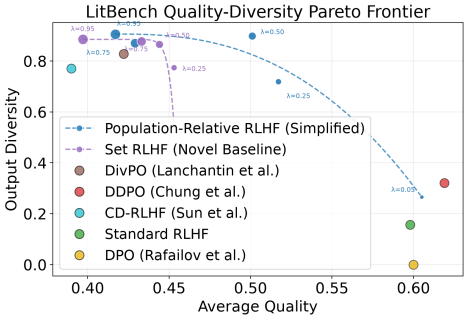

Figure 2: LitBench: Pareto frontier.

[1]However, note that a larger AUC does not exactly correspond to a better Pareto frontier, as AUC rewards extra width on the x-axis. In Figure 1 ($N = 8$), visually the simplified population RLHF algorithm achieves the most desirable Pareto frontier, but AUC for the full hypergradient algorithm is larger.

AI USE STATEMENT

We used generative AI tools in algorithm implementation. Specifically, AI tools assisted in the implementation of the full hypergradient population-relative RLHF algorithm in happy tweet experiments, and in the implementation of all algorithms in the LitBench experiments. All LLM-generated code was reviewed and verified by the authors, and the authors take responsibility for the content of this work, including text and implementation produced with the aid of AI tools. We also used AI to refine the writing of the main paper, appendix, and proofs of technical lemma. However, AI tools did not create the initial content; they only revised what the authors had written.

In this work, we did not use generative AI tools for other purposes, including retrieval and discovery, synthetic data generation, or the writing of entire sections and mathematical results beyond minor edits.

REPRODUCIBILITY STATEMENT

For reproducibility of our mathematical and experimental results, we present the full proofs of all mathematical results in the appendix and provide an implementation of our experiments in the supplementary material.

REFERENCES

Ying Cao, Xun Li, and Bing-Chang Wang. Inverse reinforcement learning for indefinite mean-field social optimization with multiplicative noise. *arXiv preprint arXiv:2608.07252*, 2026.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. Parl: A unified framework for policy alignment in reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 24410–24449, 2024.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024.

Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654*, 2021.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*, 2025.

Kai Cui, Gökçe Dayanıklı, Mathieu Laurière, Matthieu Geist, Olivier Pietquin, and Heinz Koeppl. Learning discrete-time major-minor mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9616–9625, 2024.

Shehzaad Dhuliawala, Ilia Kulikov, Ping Yu, Asli Celikyilmaz, Jason Weston, Sainbayar Sukhbaatar, and Jack Lanchantin. Adaptive decoding via latent preference optimization. *arXiv preprint arXiv:2411.09661*, 2024.

Li Ding, Jenny Zhang, Jeff Clune, Lee Spector, and Joel Lehman. Quality diversity through human feedback: Towards open-ended diversity-driven optimization. *arXiv preprint arXiv:2310.12103*, 2023.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, 2018.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7740–7755, 2026.

Mudit Gaur, Utsav Singh, Amrit Singh Bedi, Raghu Pasupathy, and Vaneet Aggarwal. On the sample complexity bounds of bilevel reinforcement learning. *Advances in Neural Information Processing Systems*, 38:67538–67570, 2026.

Jubayer Hamid, Ifdita Orney, Ellen Xu, Chelsea Finn, and Dorsa Sadigh. Polychromic objectives for reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 37356–37381, 2026.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Feihu Huang. Optimal hessian/jacobian-free nonconvex-pl bilevel optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19598–19621, 2024.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pp. 216–225, 2014.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. In *International Conference on Learning Representations*, volume 2024, pp. 20620–20653, 2024.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, and Hanbaek Lyu. On the complexity of first-order methods in stochastic bilevel optimization. *arXiv preprint arXiv:2402.07101*, 2024a.

Oh Joon Kwon, Daiki E Matsunaga, and Kee-Eung Kim. Gdpo: Learning to directly align language models with diversity using gflownets. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17120–17139, 2024b.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Florian Le Bronnec, Alexandre Verine, Benjamin Negrevergne, Yann Chevaleyre, and Alexandre Allauzen. Exploring precision and recall to assess the quality and diversity of llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11418–11441, 2024.

Maggie, Phil Culliton, and Wei Chen. Tweet sentiment extraction. `https://kaggle.com/competitions/tweet-sentiment-extraction`, 2020. Kaggle.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11241–11258, 2025.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs. In *International Conference on Learning Representations*, volume 2025, pp. 70333–70366, 2025.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*, volume 2024, pp. 642–669, 2024.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pp. 737–746. PMLR, 2016.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International conference on machine learning*, pp. 30992–31015. PMLR, 2023.

Haoran Sun, Yekun Chai, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Curiosity-driven reinforcement learning from human feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23517–23534, 2025.

Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:39491–39503, 2023.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pp. 39722–39754. PMLR, 2023.

Sihan Zeng, Sujay Bhatt, Sumitra Ganesh, and Alec Koppel. A hessian-free actor-critic algorithm for bi-level reinforcement learning with applications to llm fine-tuning. *arXiv preprint arXiv:2601.16399*, 2026a.

Sihan Zeng, Benjamin Patrick Evans, Sujay Bhatt, Leo Ardon, Sumitra Ganesh, and Alec Koppel. Learning in stackelberg mean field games: A non-asymptotic analysis. *Advances in Neural Information Processing Systems*, 38:170504–170554, 2026b.

Chenyu Zhang, Xu Chen, and Xuan Di. Stochastic semi-gradient descent for learning mean field games with population-aware function approximation. In *International Conference on Learning Representations*, volume 2025, pp. 42202–42249, 2025.

Shimao Zhang, Yu Bao, and Shujian Huang. Edt: Improving large language models' generation by entropy-based dynamic temperature sampling. *arXiv preprint arXiv:2403.14541*, 2024.

Appendix

CONTENTS

1 Introduction **1**

2 Formulation **2**
2.1 Population-Relative RLHF for LLM Alignment 2
2.2 Bi-Level Optimization Framework 3

3 Algorithm Development **3**
3.1 Instantiation to Population-Relative RLHF 5
3.2 Memory-Efficient Simplification 6

4 Finite-Time Analysis **6**

5 Experimental Results **8**
5.1 Happy Tweet Generation 8
5.2 LitBench 9

A Related Work **14**

B Set RLHF Algorithm **15**

C Experimental Setup and Additional Results **17**
C.1 Happy Tweet Generation 18
C.2 LitBench 18

D Projection Operator in Huang (2024) **20**

E Frequently Used Notations, Inequalities, and Lemmas **21**
E.1 Supporting Lemmas 23

F Proof of Theorem 1 **25**

G Proof of Propositions **27**
G.1 Proof of Proposition 1 27
G.2 Proof of Proposition 2 31

H Proof of Lemmas **32**
H.1 Proof of Lemma 1 32

H.2 Proof of Lemma 2 33
H.3 Proof of Lemma 3 33
H.4 Proof of Lemma 4 34
H.5 Proof of Lemma 5 36
H.6 Proof of Lemma 6 37
H.7 Proof of Lemma 7 37
H.8 Proof of Lemma 8 38
H.9 Proof of Lemma 9 39
H.10 Proof of Lemma 10 40

A Related Work

Our paper is closely related to the literature on diversity-aware LLM post-training methods and bi-level optimization. We discuss the most relevant prior work below to give context to our contributions.

LLM Output Diversity. Many studies have shown that post-training with human feedback can reduce the diversity of LLM outputs. In particular, RLHF has been found to improve some aspects of generalization while significantly reducing output diversity relative to supervised fine-tuning across multiple measures and tasks (Kirk et al., 2024). Similar effects have been observed particularly in the context of creative writing, where feedback-tuned LLMs produce more homogeneous content and less lexical/content diversity (Padmakumar & He, 2024). More recent analyses further suggest that aligned models can exhibit reduced conceptual diversity compared to instruction fine-tuned counterparts (Murthy et al., 2025). These observations motivate training objectives that preserve or explicitly model diversity in applications where variation or exploration is important, such as creative writing and open-ended content generation.

One line of work seeks to increase diversity in the decoding stage, through enhanced temperature scaling or top-p sampling (Dhuliawala et al., 2024; Zhang et al., 2024; Nguyen et al., 2025). However, these methods operate only through the sampling rule and do not train the LLM to internalize a diversity-aware behavior. In addition, temperature-based decoding can overly inflate the probability of lower-quality or nonsensical tokens, which degrades output quality and coherence (Lanchantin et al., 2025).

Another relevant line of work develops diversity-promoting preference optimization methods in the post-training stage, and our work also falls within this line. Earlier approaches include GDPO (Kwon et al., 2024b), which uses a GFlowNet objective to preserve multiple high-reward modes, and QDHF (Ding et al., 2023), which progressively learns a diversity representation from human similarity judgments and uses it as the behavioral descriptor in quality-diversity optimization.

More recent methods in the post-training stage include Curiosity-Driven RLHF (Sun et al., 2025), DivPO (Lanchantin et al., 2025), and DDPO (Chung et al., 2025), all of which we implement as baselines. Curiosity-Driven RLHF (CD-RLHF) augments the conventional reward in RLHF with a learned intrinsic reward based on state-prediction error, encouraging the policy to explore less familiar generations. DivPO and DDPO are built on top of the standard DPO framework. The former differs from DPO in the construction of the preference pairs: instead of using the preference pairs as given, it first samples a pool of responses, scores them with a reward model, applies reward thresholds, and then selects the chosen and rejected examples using a diversity criterion. DDPO instead reweights the standard preference loss by a deviation score, so that rare but high-quality responses receive greater emphasis. While both methods promote diversity through preference optimization, they still rely on fixed design choices such as deviation measures, thresholds, and pair-selection rules. Furthermore, DivPO relies on quality scores given by a reward model to construct preference pairs, partially reintroducing the reward-model dependency that DPO is designed to avoid. In contrast, our proposed method does not require such additional information.

Bi-Level Optimization. Bi-level optimization studies structured optimization problems where the upper-level objective depends on the solution of a lower-level optimization problem. The challenge

in bi-level optimization stems mainly from differentiating through the lower-level best response, which introduces Hessian- and Jacobian-related terms expensive to compute and estimate in practice, especially when the lower-level objective is stochastic and/or non-convex.

There are two common ways to design algorithms without requiring a second-order oracle or explicitly estimating second-order information: penalty-based method (Kwon et al., 2023; 2024a; Shen & Chen, 2023; Chen et al., 2024; Zeng et al., 2026a; Gaur et al., 2026), which reformulates the problem through a penalty term that couples the two levels, and finite-difference-based methods, which approximate the needed Hessian-vector- and Jacobian-vector-products through stochastic approximation (Yang et al., 2023; Huang, 2024). We take the second approach, and our algorithm design and analysis are inspired by Huang (2024). However, we make three important generalizations: 1) our lower-level problem is not a static optimization problem but rather to find the equilibrium in a mean-field-type game; 2) Huang (2024) considers the deterministic gradient setting where the finite-difference precision ε can be chosen arbitrarily, whereas our work handles stochastic gradients and noisy finite-difference estimates, which requires careful selection of ε to balance bias and variance and to interact properly with the step sizes through a multi-time-scale analysis; 3) our algorithm eliminates a practically infeasible operation required by Huang (2024) that projects a matrix-vector product onto the column space of a transformed version of the matrix.

B Set RLHF Algorithm

We design the Set RLHF algorithm based on the same core observation that inspires population-relative RLHF: standard RLHF learns a reward model that scores individual responses in isolation, making it incapable of optimizing distribution-level properties such as output diversity.

A natural alternative to address this limitation is to collect preference feedback directly over sets of responses. Specifically, given an input prompt $x \sim \mathcal{D}_{\text{prompt}}$, the LLM policy π_θ generates two independent candidate sets of N responses:

$$\boldsymbol{y}_1 = \{y_{1,1}, \ldots, y_{1,N}\} \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid x), \quad \boldsymbol{y}_2 = \{y_{2,1}, \ldots, y_{2,N}\} \overset{\text{i.i.d.}}{\sim} \pi_\theta(\cdot \mid x).$$

A human annotator is presented with both sets and asked to provide a preference label $p \in \{0, 1\}$ indicating which set as a whole exhibits a superior balance of individual response quality and intra-set diversity. In the happy tweet generation task, we simulate the human preference distribution by the Bradley-Terry model, where the probability that set $\boldsymbol{y}_1$ is preferred over set $\boldsymbol{y}_2$ is given by

$$P_\omega(\boldsymbol{y}_1 \succ \boldsymbol{y}_2 \mid x) = \text{sigmoid}\Big(R_\omega(\boldsymbol{y}_1 \mid x) - R_\omega(\boldsymbol{y}_2 \mid x)\Big),$$

parameterized by a set-level reward model $R_\omega(\boldsymbol{y} \mid x)$. To generate synthetic preference labels in our experiment, the ground-truth set utility is computed as

$$R_{\text{truth}}(\boldsymbol{y} \mid x) = (1 - \delta) \cdot \frac{1}{N} \sum_{n=1}^{N} R_{\text{quality}}(y_n) + \delta \cdot D_{\text{Jaccard}}(\boldsymbol{y}),$$

where R_{quality} and R_{div} are defined the same as in Section 5. A similar construction of R_{truth} is used in the LitBench experiments, where R_{quality} is given by a judge LLM provided in Fein et al. (2026) and $R_{\text{div}}(\boldsymbol{z})$ calculates the normalized Vendi score (in the sentence embedding space) averaged over pairs of responses in $\boldsymbol{z}$.

Reward Model Learning. Using the preference labels $p \in \{0, 1\}$ generated from R_{truth}, we fit the parameterized set-level reward model $R_\omega(\boldsymbol{y} \mid x)$ by minimizing the binary cross-entropy loss over candidate set pairs.

$$\mathcal{L}_{\text{RM}}(\omega) = \text{BCE}\Big(\text{sigmoid}\big(R_\omega(\boldsymbol{y}_1 \mid x) - R_\omega(\boldsymbol{y}_2 \mid x)\big),\, p\Big).$$

Note that $R_\omega(\boldsymbol{y} \mid x)$ learns a parametric mapping from an input set of N responses to a scalar directly from preference data without direct access to R_{truth}.

Algorithm 3 Population-Relative Reinforcement Learning with Human Feedback (Simplified)

Require: Prompt distribution $\mathcal{D}_{\text{prompt}}$, population size N, number of inner-loop iterations for reward and language model updates T, KL coefficient $\tau > 0$, and anchor policy π_{KL}.
Initialize: Reward model parameter ω_0, LLM model parameter θ_0 encoding the policy π_0.
for iteration $k = 0, 1, 2, \ldots$ **do**

Reward Model Update

Set $\omega_{k,0} = \omega_k$.
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (batch) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample population $\boldsymbol{z}_N = \{z_1, \ldots, z_N\}$ and candidates (y_1, y_2) i.i.d. from $\pi_k(\cdot \mid x)$.
  Query human preference

$$p = \begin{cases} 1, & \text{if } y_1 \text{ is preferred over } y_2 \text{ given } \boldsymbol{z}_N, \\ 0, & \text{otherwise.} \end{cases}$$

  Compute reward model loss under Bradley-Terry model

$$L_{\text{RM}}(\omega_{k,t}) = \text{BCE}\big(\text{sigmoid}(R_{\omega_{k,t}}(y_1 \mid x, \boldsymbol{z}_N) - R_{\omega_{k,t}}(y_2 \mid x, \boldsymbol{z}_N)),\, p\big).$$

  Update reward model parameter

$$\omega_{k,t+1} = \omega_{k,t} - \alpha_k \nabla_\omega L_{\text{RM}}(\omega_{k,t}).$$

end for
Set $\omega_{k+1} = \omega_{k,T}$.

Language Model Update

Set $\theta_{k,0} = \theta_k$ and $\pi_{k,0} = \pi_k$.
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (a batch of) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample population $\boldsymbol{z}_N = \{z_1, \ldots, z_N\}$ and candidate y i.i.d. from $\pi_{k,t}(\cdot \mid x)$.
  Score candidate with reward model to obtain $R_{\omega_k}(y \mid x, \boldsymbol{z}_N)$.
  Estimate advantage function A from $R_{\omega_k}(y \mid x, \boldsymbol{z}_N)$.
  Compute (token-level) KL divergence to anchor policy

$$\text{KL} = \log \pi_{k,t}(y \mid x) - \log \pi_{\text{KL}}(y \mid x).$$

  Update LLM parameter

$$\theta_{k,t+1} = \theta_{k,t} + \beta_k \big(A - \tau \text{KL}\big) \nabla_\theta \log \pi_\theta(y \mid x).$$

end for
Set $\theta_{k+1} = \theta_{k,T}$ and $\pi_{k+1} = \pi_{k,T}$.

end for

Policy Optimization via Set RL. Given the iteratively updated set-level reward model $R_\omega(\boldsymbol{y} \mid x)$, the LLM policy is trained to maximize the expected set reward subject to a token-level KL-divergence regularization relative to the anchor policy π_{KL}. The complete procedure for reward model and LLM updates is detailed in Algorithm 4.

While Set RLHF provides a direct formulation for optimizing set-level properties, it suffers from two major practical and algorithmic bottlenecks.

- **Annotation Difficulty.** The human annotator must read, evaluate, and compare two full candidate sets of N responses simultaneously – requiring the mental processing of $2N$ complete text generations per prompt. Synthesizing quality and diversity across $2N$ outputs places an extraordinarily high burden on the annotator. In contrast, population-relative RLHF (Algorithm 2) retains standard pairwise comparisons between individual candidate outputs (y_1 vs. y_2), using the reference population $\boldsymbol{z}_N$ solely as a contextual anchor for diversity evaluation. This significantly reduces human annotation complexity.
- **Credit Assignment.** Because the set reward model $R_\omega(\boldsymbol{y} \mid x)$ returns a single scalar score for an entire set of N responses, attributing credit to a specific output $y_n \in \boldsymbol{y}$ is inherently difficult. As set size N increases, this severe credit assignment problem leads to significantly slower empirical convergence and inferior plateau performance compared to population-relative RLHF.

Algorithm 4 Set Reinforcement Learning with Human Feedback

Require: Prompt distribution $\mathcal{D}_{\text{prompt}}$, set size N, number of inner-loop iterations for reward and language model updates T, KL coefficient $\tau > 0$, and anchor policy π_{KL}.
Initialize: Reward model parameter ω_0, LLM model parameter θ_0 encoding the policy π_0.
for iteration $k = 0, 1, 2, \ldots$ **do**

Reward Model Update

Set $\omega_{k,0} = \omega_k$.
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (batch) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample two candidate sets of responses

$$\boldsymbol{y}_1 = \{y_{1,1}, \ldots, y_{1,N}\} \overset{\text{i.i.d.}}{\sim} \pi_k(\cdot \mid x),$$
$$\boldsymbol{y}_2 = \{y_{2,1}, \ldots, y_{2,N}\} \overset{\text{i.i.d.}}{\sim} \pi_k(\cdot \mid x).$$

  Query human preference

$$p = \begin{cases} 1, & \text{if } \boldsymbol{y}_1 \text{ is preferred over } \boldsymbol{y}_2 \text{ (considering both quality and diversity),} \\ 0, & \text{otherwise.} \end{cases}$$

  Compute reward model loss under Bradley-Terry model

$$L_{\text{RM}}(\omega_{k,t}) = \text{BCE}\Big(\text{sigmoid}(R_{\omega_{k,t}}(\boldsymbol{y}_1 \mid x) - R_{\omega_{k,t}}(\boldsymbol{y}_2 \mid x)),\, p\Big).$$

  Update reward model parameter:

$$\omega_{k,t+1} = \omega_{k,t} - \alpha_k \nabla_\omega L_{\text{RM}}(\omega_{k,t}).$$

end for
Set $\omega_{k+1} = \omega_{k,T}$.

Language Model Update

Set $\theta_{k,0} = \theta_k$ and $\pi_{k,0} = \pi_k$.
for inner loop iteration $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample (a batch of) prompt $x \sim \mathcal{D}_{\text{prompt}}$.
  Sample set of responses $\boldsymbol{y} = \{y_1, \ldots, y_N\} \overset{\text{i.i.d.}}{\sim} \pi_{k,t}(\cdot \mid x)$.
  Score candidate with reward model to obtain $R_{\omega_k}(\boldsymbol{y} \mid x)$.
  Estimate (generalized) advantage function A from $R_{\omega_k}(\boldsymbol{y} \mid x)$, following Hamid et al. (2026).
  Compute (token-level) KL divergence to anchor policy

$$\text{Kn}_i = \log \pi_{k,t}(y_n \mid x) - \log \pi_{\text{KL}}(y_n \mid x).$$

  Update LLM parameter

$$\theta_{k,t+1} = \theta_{k,t} + \beta_k \sum_{n=1}^{N} \big(A - \tau \text{KL}_n\big) \nabla_\theta \log \pi_\theta(y_n \mid x).$$

end for
Set $\theta_{k+1} = \theta_{k,T}$ and $\pi_{k+1} = \pi_{k,T}$.

end for

C Experimental Setup and Additional Results

In both happy tweet and LitBench experiments, we keep all hyper-parameters and configurations strictly identical across algorithms whenever possible, including LLM architecture, learning rate, KL regularization weight, and decoding strategy. The only exception is that in LitBench, DivPO is optimized under the initial learning rate 1×10^{-6}, whereas all other methods use 1×10^{-5}. The learning rate for DivPO is selected differently as we find that the algorithm under 1×10^{-5} converges to a collapsed policy that solely optimizes the quality score.

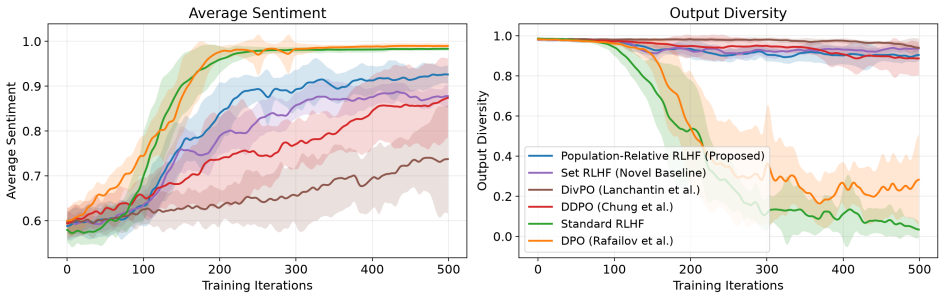

Figure 3: Happy tweet – quality and diversity learning curves ($N = 2$, $\delta = 0.5$).

C.1 Happy Tweet Generation

We simulate human preferences according to (17), we use a rule-based sentiment analyzer and a fixed diversity metric. Specifically, $R_{\text{quality}}(y) \in [-1, 1]$ is a sentiment score given by VADER, which assigns predefined scores to words and other lexical features and adjusts them using rules that account for context, such as negation, capitalization, and punctuation (Hutto & Gilbert, 2014). Meanwhile, $R_{\text{div}}(\boldsymbol{z})$ measures the diversity of a set $\boldsymbol{z}$ (of cardinality M) via the following average pairwise word-level Jaccard distance

$$R_{\text{div}}(\boldsymbol{z}) = \frac{2}{M(M-1)} \sum_{1 \le i < j \le M} \left(1 - \frac{|\text{unique words}(z_i) \cap \text{unique words}(z_j)|}{|\text{unique words}(z_i) \cup \text{unique words}(z_j)|}\right).$$

In this experiment, the language model has a tiny architecture – it has 4 transformer layers, 4 attention heads, and a hidden dimension of 128. The model is first pretrained on the Kaggle Tweet Sentiment Extraction dataset (Maggie et al., 2020) and then supervised fine-tuned using only the positive-sentiment examples. The resulting SFT model is used both as the initial policy π_0 and as the anchor policy π_{KL} for KL regularization. All post-training methods start from the same SFT checkpoint.

For pretraining, we use a sequence length of 32, a batch size of 64, AdamW with learning rate 1×10^{-3}, weight decay 0.1, and gradient-norm clipping at 1.0. The model is pretrained for 10 epochs. We then fine-tune the same model for 6 epochs on positive-sentiment tweets using learning rate 3×10^{-4}, with the other optimizer settings unchanged.

All methods are trained for 500 iterations with batch size 32. The policy learning rate is 2×10^{-4}, and for RLHF-family methods, which learn an explicit reward model, the reward model learning rate is 5×10^{-4}. The KL regularization coefficient is $\tau = 0.2$.

In Tables 2-4, we report the confidence intervals behind the Pareto frontier plots in Figure 1. Note that DivPO, DDPO, CD-RLHF, DPO, and standard RLHF does not involve the population size N. The scores of these methods across Tables 2-4 are from the same 5 runs and therefore identical.

In Figures 3-5, we plot the learning curves of population-relative RLHF and the baseline methods, which highlight that population-relative RLHF prevents mode collapse while optimizing response quality. In comparison, standard RLHF and DPO rapidly destroys response diversity within the first 200 iterations to maximize the sentiment alone.

C.2 LitBench

For LitBench, our language model is Llama-3.1-8B-Instruct. The model is first supervised fine-tuned on the WritingPrompts dataset (Fan et al., 2018). We use a maximum prompt length of 256 tokens and a maximum story length of 1024 tokens. The SFT model is trained for one epoch with learning rate 1×10^{-4} and batch size 2. We use a 100-step linear warm-up followed by cosine decay to 10% of the peak learning rate. The SFT model is subsequently used as the initialization policy π_0 and anchor policy π_{KL} for all post-training methods.

Table 2: Average sentiment (quality) and output diversity after training for $N = 2$ (5 seeds, 95% CI).

Method	Sentiment	Diversity
Population-Relative RLHF ($\delta = 0$)	0.985 ± 0.006	0.103 ± 0.164
Population-Relative RLHF ($\delta = 0.25$)	0.951 ± 0.009	0.835 ± 0.031
Population-Relative RLHF ($\delta = 0.5$)	0.922 ± 0.016	0.902 ± 0.020
Population-Relative RLHF ($\delta = 0.75$)	0.825 ± 0.046	0.952 ± 0.027
Population-Relative RLHF ($\delta = 1$)	0.531 ± 0.036	0.995 ± 0.004
Set RLHF ($\delta = 0$)	0.977 ± 0.006	0.271 ± 0.346
Set RLHF ($\delta = 0.25$)	0.932 ± 0.017	0.867 ± 0.028
Set RLHF ($\delta = 0.5$)	0.874 ± 0.009	0.935 ± 0.004
Set RLHF ($\delta = 0.75$)	0.783 ± 0.023	0.964 ± 0.006
Set RLHF ($\delta = 1$)	0.570 ± 0.021	0.991 ± 0.005
DivPO ($\rho = 0$)	0.859 ± 0.086	0.869 ± 0.065
DivPO ($\rho = 0.25$)	0.736 ± 0.077	0.955 ± 0.019
DivPO ($\rho = 0.5$)	0.659 ± 0.052	0.978 ± 0.013
DDPO	0.861 ± 0.086	0.891 ± 0.080
CD-RLHF ($\eta = 0.01$)	0.986 ± 0.004	0.317 ± 0.135
CD-RLHF ($\eta = 0.1$)	0.931 ± 0.101	0.725 ± 0.198
CD-RLHF ($\eta = 0.2$)	0.595 ± 0.208	0.972 ± 0.061
Standard RLHF	0.983 ± 0.003	0.060 ± 0.026
DPO	0.990 ± 0.006	0.248 ± 0.116

Table 3: Average sentiment (quality) and output diversity after training for $N = 5$ (5 seeds, 95% CI).

Method	Sentiment	Diversity
Population-Relative RLHF ($\delta = 0$)	0.982 ± 0.004	0.046 ± 0.043
Population-Relative RLHF ($\delta = 0.25$)	0.956 ± 0.012	0.832 ± 0.030
Population-Relative RLHF ($\delta = 0.5$)	0.905 ± 0.010	0.915 ± 0.023
Population-Relative RLHF ($\delta = 0.75$)	0.840 ± 0.027	0.964 ± 0.005
Population-Relative RLHF ($\delta = 1$)	0.527 ± 0.019	0.998 ± 0.002
Set RLHF ($\delta = 0$)	0.908 ± 0.017	0.824 ± 0.053
Set RLHF ($\delta = 0.25$)	0.904 ± 0.016	0.888 ± 0.029
Set RLHF ($\delta = 0.5$)	0.874 ± 0.007	0.940 ± 0.004
Set RLHF ($\delta = 0.75$)	0.839 ± 0.019	0.963 ± 0.006
Set RLHF ($\delta = 1$)	0.559 ± 0.010	0.997 ± 0.001
DivPO ($\rho = 0$)	0.859 ± 0.086	0.869 ± 0.065
DivPO ($\rho = 0.25$)	0.736 ± 0.077	0.955 ± 0.019
DivPO ($\rho = 0.5$)	0.659 ± 0.052	0.978 ± 0.013
DDPO	0.861 ± 0.086	0.891 ± 0.080
CD-RLHF ($\eta = 0.01$)	0.986 ± 0.004	0.317 ± 0.135
CD-RLHF ($\eta = 0.1$)	0.931 ± 0.101	0.725 ± 0.198
CD-RLHF ($\eta = 0.2$)	0.595 ± 0.208	0.972 ± 0.061
Standard RLHF	0.983 ± 0.003	0.060 ± 0.026
DPO	0.990 ± 0.006	0.248 ± 0.116

All LitBench post-training methods use LoRA rather than full-parameter fine-tuning. We use rank $r = 32$, scaling factor $\alpha = 64$, and dropout 0.05. LoRA adapters are inserted into the query, key, value, and output projection matrices as well as the gated MLP projections.

Table 4: Average sentiment (quality) and output diversity after training for $N = 8$ (5 seeds, 95% CI).

Method	Sentiment	Diversity
Population-Relative RLHF ($\delta = 0$)	0.988 ± 0.004	0.181 ± 0.257
Population-Relative RLHF ($\delta = 0.25$)	0.954 ± 0.010	0.840 ± 0.017
Population-Relative RLHF ($\delta = 0.5$)	0.901 ± 0.033	0.924 ± 0.030
Population-Relative RLHF ($\delta = 0.75$)	0.792 ± 0.026	0.978 ± 0.004
Population-Relative RLHF ($\delta = 1$)	0.525 ± 0.017	0.998 ± 0.003
Set RLHF ($\delta = 0$)	0.866 ± 0.024	0.867 ± 0.041
Set RLHF ($\delta = 0.25$)	0.865 ± 0.008	0.909 ± 0.009
Set RLHF ($\delta = 0.5$)	0.856 ± 0.014	0.942 ± 0.003
Set RLHF ($\delta = 0.75$)	0.833 ± 0.020	0.967 ± 0.003
Set RLHF ($\delta = 1$)	0.563 ± 0.013	0.997 ± 0.001
DivPO ($\rho = 0$)	0.859 ± 0.086	0.869 ± 0.065
DivPO ($\rho = 0.25$)	0.736 ± 0.077	0.955 ± 0.019
DivPO ($\rho = 0.5$)	0.659 ± 0.052	0.978 ± 0.013
DDPO	0.861 ± 0.086	0.891 ± 0.080
CD-RLHF ($\eta = 0.01$)	0.986 ± 0.004	0.317 ± 0.135
CD-RLHF ($\eta = 0.1$)	0.931 ± 0.101	0.725 ± 0.198
CD-RLHF ($\eta = 0.2$)	0.595 ± 0.208	0.972 ± 0.061
Standard RLHF	0.983 ± 0.003	0.060 ± 0.026
DPO	0.990 ± 0.006	0.248 ± 0.116

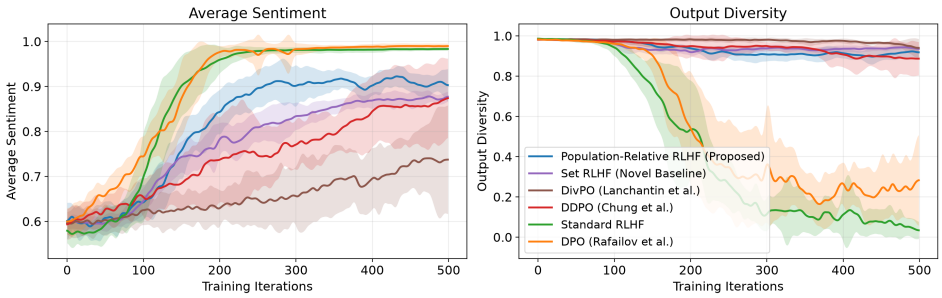

Figure 4: Happy tweet – quality and diversity learning curves ($N = 5$, $\delta = 0.5$).

The main training runs use 250 optimization iterations, a policy batch size of 4, KL coefficient $\tau = 0.1$. For RLHF-family methods, the learning rate is 1×10^{-5} for both the language model and reward model. For DPO and DDPO, the language model learning rate is also 1×10^{-5}. For DivPO, the language model learning rate is also 1×10^{-6}.

Representative stories generated by models trained with DPO, DivPO, and population-relative RLHF ($\delta = 0.75$) are shown in Table 5. DPO exhibits pronounced reward hacking: because the underlying reward model is exploitable and assigns a high score to "Damn", the DPO-trained model collapses entirely to generating this single word. Meanwhile, as shown in Figure 2, population-relative RLHF ($\delta = 0.75$) achieves comparable quality to DivPO with slightly higher output diversity, though the difference in diversity remains subtle to the human eye.

D Projection Operator in Huang (2024)

The following discussion uses the notation from Huang (2024). The algorithm in Huang (2024) relies on a projection operator $\mathcal{M}_{r_h}(H, v) \triangleq \mathcal{S}_{[\mu, L_g]}[H]\mathcal{P}_{r_v}(v)$, where $\mathcal{S}_{[\mu, L_g]}[\cdot]$ projects a $p \times p$ matrix onto the set of matrices with eigenvalues in $[\mu, L_g]$ via singular value decomposition

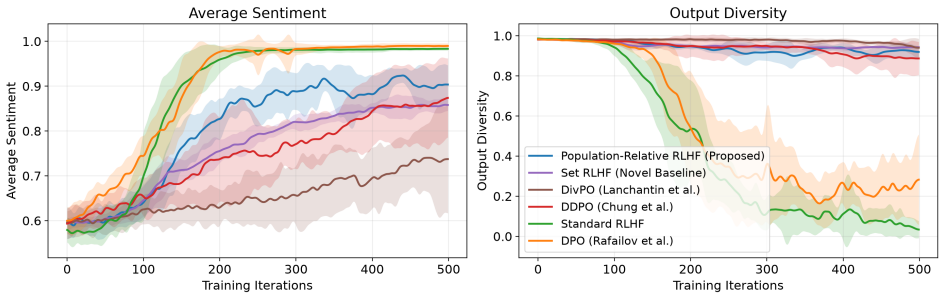

Figure 5: Happy tweet – quality and diversity learning curves ($N = 8$, $\delta = 0.5$).

Writing Prompt: *The last human on Earth sits alone in a room. There is a knock on the door.*		
DPO	DivPO	Population Relative RLHF ($\delta = 0.75$)
Damn	"You!" The last man on earth shouted excitedly, he was thrilled to finally have someone to talk to! "I don't get it, how'd you get in here?" He asked, he was sure he had locked this room ...	It wasn't the knock that startled me, but the fact that I heard a knock at all. I was still listening to make sure I was just imagining things when I heard it again. The next time it was louder ...
Damn	"Hey, Jim, where you at? You said you'd meet me 10 minutes ago." The other end was a loud static. "Hello? Jim, please come to your senses." After a few tries, Jim hung up the receiver ...	I am alone now. The world has ended. I am the last one remaining. I sit in my room, quietly crying, my tears falling into my glass of wine ...
Damn	I sighed. How do you greet the last thing left to greet you. "Hello?" I called out. Silence. ...	"Hello, my name is Alex." I said. "Do you want to help?" The other person said. "Help what?" "The apocalypse, my friend."

Table 5: LitBench: examples of generated stories.

(SVD). This introduces a fundamental contradiction: either one must explicitly instantiate the $p \times p$ lower-level Hessian matrix and perform SVD at every iteration (incurring $\mathcal{O}(p^3)$ time and $\mathcal{O}(p^2)$ memory complexity, which defeats the purpose of designing a Hessian-free algorithm), or if $\mathcal{M}_{r_h}$ is applied directly to a finite-difference vector $\tilde{H} \in \mathbb{R}^p$, the matrix SVD operator $\mathcal{S}_{[\mu, L_g]}$ becomes mathematically ill-defined.

In contrast, we eliminate the projection in our algorithm. It is worth noting that our analysis employs a similar projection operator to define $u^\star(\omega, \theta)$ in (35) (recall that $u^\star(\omega_k, \theta^\star(\omega_k))$ is a learning target for u_k). However, this projection serves strictly as an analytical construction to ensure that the learning target $u^\star(\omega, \theta)$ is well-defined and uniformly bounded. Crucially, the projection is never executed within Algorithm 1.

E Frequently Used Notations, Inequalities, and Lemmas

We introduce a few notations frequently used in our proofs and analysis.

- Given control variable ω, decision policy π, and reference policy π', we define the soft value functions

$$V^{\omega,\pi,\pi'}(s) \triangleq \mathbb{E}_{a_k\sim\pi(\cdot|s_k),s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k)}\left[\sum_{k=0}^{\infty}\gamma^k\big(r_\omega(s_k,a_k\mid\pi')-\tau\log\pi(a_k\mid s_k)\big)\,|\,s_0=s\right],$$

$$A^{\omega,\pi,\pi'}(s,a) \triangleq r_\omega(s,a\mid\pi')-\tau\log\pi(a\mid s)+\gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[V^{\omega,\pi,\pi'}(s')]-V^{\omega,\pi,\pi'}(s). \quad (18)$$

Define

$$B_V \triangleq \frac{1+\tau\log|\mathcal{A}|}{1-\gamma}. \quad (19)$$

Since $r_\omega(s,a\mid\pi')\in[0,1]$, the value functions are uniformly bounded element-wise by B_V, i.e., for all s,a

$$0\le V^{\omega,\pi,\pi'}(s)\le B_V,\quad |A^{\omega,\pi,\pi'}(s,a)|\le B_V.$$

- We define the filtration (i.e. randomness information up to iteration k)

$$\mathcal{F}_k \triangleq \{\xi_0,\cdots,\xi_k,s_0^-,\cdots,s_k^-,a_0^-,\cdots,a_k^-,s_0^+,\cdots,s_k^+,a_0^+,\cdots,a_k^+\}.$$

- We introduce the following operators that define the updates of ω_k and u_k in Algorithm 1.

$$\hat{H}_1(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+)$$
$$\triangleq \frac{A^+(s^+,a^+)\nabla_\theta\log\pi_{\theta+\epsilon u}(a^+\mid s^+)-A^-(s^-,a^-)\nabla_\theta\log\pi_{\theta-\epsilon u}(a^-\mid s^-)}{2(1-\gamma)\varepsilon}, \quad (20)$$

$$\hat{H}_2(\omega,\theta',\varepsilon,s^-,a^-,s^+,a^+) \triangleq \frac{\nabla_{\theta'}r_\omega(s^+,a^+\mid\pi_{\theta'})-\nabla_{\theta'}r_\omega(s^-,a^-\mid\pi_{\theta'})}{2(1-\gamma)\varepsilon}, \quad (21)$$

$$\hat{H}(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+)$$
$$\triangleq \hat{H}_1(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+)+\hat{H}_2(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+), \quad (22)$$

$$\hat{G}(\omega,\theta',\varepsilon,s^-,a^-,s^+,a^+) \triangleq \frac{\nabla_\omega r_\omega(s^+,a^+\mid\pi_{\theta'})-\nabla_\omega r_\omega(s^-,a^-\mid\pi_{\theta'})}{2(1-\gamma)\varepsilon}. \quad (23)$$

Recall the definition of $\hat{H}_k,\hat{G}_k$ in (14). It can be seen

$$\hat{H}_k \triangleq \hat{H}(\omega_k,\theta_k,u_k,\varepsilon_k,A_k^-,A_k^+,s_k^-,a_k^-,s_k^+,a_k^+),$$
$$\hat{G}_k \triangleq \hat{G}(\omega_k,\theta_k,\varepsilon_k,s_k^-,a_k^-,s_k^+,a_k^+).$$

- We define

$$\bar{H}_1(\omega,\theta,u,\varepsilon)=\frac{\nabla_1 J_\omega(\pi_{\theta+\varepsilon u},\pi_\theta)-\nabla_1 J_\omega(\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}, \quad (24)$$

$$\bar{H}_2(\omega,\theta,u,\varepsilon)=\frac{\nabla_2 J_\omega(\pi_{\theta+\varepsilon u},\pi_\theta)-\nabla_2 J_\omega(\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}, \quad (25)$$

$$\bar{H}(\omega,\theta,u,\varepsilon)=\bar{H}_1(\omega,\theta,u,\varepsilon)+\bar{H}_2(\omega,\theta,u,\varepsilon), \quad (26)$$

$$\bar{G}(\omega,\theta,u,\varepsilon)=\frac{\nabla_\omega J_\omega(\omega,\pi_{\theta+\varepsilon u},\pi_\theta)-\nabla_\omega J_\omega(\omega,\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}, \quad (27)$$

$$\tilde{H}(\omega,\theta,u)=\Big(\nabla^2_{1,1}J_\omega(\omega,\pi_\theta,\pi_\theta)+\nabla^2_{2,1}J_\omega(\omega,\pi_\theta,\pi_\theta)\Big)u, \quad (28)$$

$$\tilde{G}(\omega,\theta,u)=\nabla^2_{\omega,1}J_\omega(\omega,\pi_\theta,\pi_\theta)u. \quad (29)$$

Note that $\bar{H},\bar{G}$ are unbiased estimators of $\hat{H},\hat{G}$ under the correct value functions

$$\bar{H}(\omega,\theta,u,\varepsilon)$$
$$=\mathbb{E}_{\substack{s^-\sim d_\rho^{\pi_{\theta-\varepsilon u}},a^-\sim\pi_{\theta-\varepsilon u}(\cdot|s^-)\\ s^+\sim d_\rho^{\pi_{\theta+\varepsilon u}},a^+\sim\pi_{\theta+\varepsilon u}(\cdot|s^+)}}[\hat{H}(\omega,\theta,u,\varepsilon,A^{\omega,\pi_{\theta-\varepsilon u},\pi},A^{\omega,\pi_{\theta+\varepsilon u},\pi},s^-,a^-,s^+,a^+)], \quad (30)$$

$$\bar{G}(\omega,\theta,u,\varepsilon)=\mathbb{E}_{s^-\sim d_\rho^{\pi_{\theta-\varepsilon u}},a\sim\pi_{\theta-\varepsilon u},s^+\sim d_\rho^{\pi_{\theta+\varepsilon u}},a\sim\pi_{\theta+\varepsilon u}}[\hat{G}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+)]. \quad (31)$$

- We define the following operators that abstract the update direction of ω_k in Algorithm 1. For $\omega, \theta, \varepsilon, s^-, a^-, s^+, a^+, \xi$,

$$\hat{D}(\omega, \theta, \varepsilon, s^-, a^-, s^+, a^+, \xi) = \nabla_\omega f(\omega, \pi_\theta, \xi) - \hat{G}(\omega, \theta, \varepsilon, s^-, a^-, s^+, a^+). \tag{32}$$

  The update of ω_k in Algorithm 1 can be alternatively expressed as follows

$$\omega_{k+1} = \omega_k - \alpha_k \hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k).$$

- We introduce the operator $\bar{D}$, which takes the expectation of operators $\hat{D}$ defined above.

$$\bar{D}(\omega, \theta, u, \varepsilon) \triangleq \mathbb{E}_{\substack{s^- \sim d_\rho^{\pi_\theta - \varepsilon u}, a^- \sim \pi_{\theta - \varepsilon u}(\cdot|s^-) \\ s^+ \sim d_\rho^{\pi_\theta + \varepsilon u}, a^+ \sim \pi_{\theta + \varepsilon u}(\cdot|s^+),\ \xi \sim \mu}} [\hat{D}(\omega, \theta, \varepsilon, s^-, a^-, s^+, a^+, \xi)]. \tag{33}$$

  It follows from (31) that

$$\bar{D}(\omega, \theta, u, \varepsilon) = \nabla_\omega f(\omega, \pi_\theta) - \bar{G}(\omega, \theta, u, \varepsilon). \tag{34}$$

- Define

$$u^\star(\omega, \theta) = \mathrm{P}_{\underline{\lambda}}\Big(\nabla_{1,1}^2 J_\omega(\pi_\theta, \pi_\theta) + \nabla_{2,1}^2 J_\omega(\pi_\theta, \pi_\theta)\Big)^{-1} \nabla_\theta f(\omega, \pi_\theta), \tag{35}$$

  where we use $\mathrm{P}_{\underline{\lambda}}(M)$ to denote the projection of a matrix M onto the set of matrices whose symmetric part has eigenvalues upper bounded by $-\underline{\lambda}$. The projection can be performed by eigendecomposing the symmetric part $(M + M^\top)/2$, clamping its eigenvalues above $-\lambda$, and retaining the skew-symmetric part.

- Due to Assumption 1, the norm of $u^\star(\omega, \theta)$ is upper-bounded by $B_u \triangleq \frac{L_f}{\underline{\lambda}}$. Note that $u^\star(\omega_k, \theta^\star(\omega_k))$ is the learning target for u_k. For algorithm stability, we enforce $\|u_k\| \le B_u$ through the projection in (10).

- Finally, we define the following residuals that measure the sub-optimality in the policy parameter θ_k, auxiliary variable u_k, and advantage function estimates A_k^+, A_k^-.

$$\begin{aligned} \mathrm{res}_k^\theta &= \|\theta_k - \theta^\star(\omega_k)\|^2, \quad \mathrm{res}_k^u = \|u_k - u^\star(\omega_k, \theta^\star(\omega_k))\|^2, \\ \mathrm{res}_k^{A,-} &= \|A_k^- - A^{\omega_k, \pi_k^-, \pi_k}\|^2, \quad \mathrm{res}_k^{A,+} = \|A_k^+ - A^{\omega_k, \pi_k^+, \pi_k}\|^2. \end{aligned} \tag{36}$$

E.1 Supporting Lemmas

A few important technical lemmas establish intermediate results that are later used in the proof of Theorem 1. We introduce the lemmas in this section and defer their proofs to Appendix H.

The following lemma establishes the Lipschitz continuity of the (advantage) value function and its higher order derivatives.

Lemma 2 *There exists a bounded constant L_V such that for any control variables ω_1, ω_2 and policy parameters $\theta_1, \theta_1', \theta_2, \theta_2'$, we have*

$$\begin{aligned}
\|A^{\omega_1, \pi_{\theta_1}, \pi_{\theta_1'}} - A^{\omega_2, \pi_{\theta_2}, \pi_{\theta_2'}}\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
|J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_\omega J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_\omega J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_1 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_1 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_2 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_2 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_{\omega,1}^2 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_{\omega,1}^2 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_{1,1}^2 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_{1,1}^2 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_{2,1}^2 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_{2,1}^2 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_{\omega,1,1}^3 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_{\omega,1,1}^3 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right), \\
\|\nabla_{1,1,1}^3 J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1'}) - \nabla_{1,1,1}^3 J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2'})\| &\le L_V \left(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\| + \|\theta_1' - \theta_2'\|\right).
\end{aligned}$$

Lemma 3 shows that the operators $\hat{H}, \hat{G}, \hat{D}$ defined in (22), (23), (32) are uniformly upper bounded.

Lemma 3 *There exist bounded constants B_H, B_G, B_D such that for all $\varepsilon \in (0,1]$ and u satisfying $\|u\| \le B_u$*

$$\big\|\hat{H}(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+)\big\| \le \frac{B_H}{\varepsilon},$$
$$\big\|\hat{G}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+)\big\| \le \frac{B_G}{\varepsilon},$$
$$\big\|\hat{D}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+,\xi)\big\| \le \frac{B_D}{\varepsilon},$$

and

$$\|\bar{H}(\omega,\theta,u,\varepsilon)\| \le B_H, \quad \|\bar{G}(\omega,\theta,u,\varepsilon)\| \le B_G, \quad \|\bar{D}(\omega,\theta,u,\varepsilon)\| \le B_D.$$

In the later analysis, we employ the Lipschitz continuity of $\hat{H}, \bar{H}, \bar{D}$ with respect to a subset of their arguments. We establish the conditions in the following lemma.

Lemma 4 *There exist bounded constants L_H, L_D such that for all $\varepsilon \in (0,1]$ and u satisfying $\|u\| \le B_u$*

$$\begin{aligned}
&\|\hat{H}(\omega,\theta,u,\varepsilon,A_1^-,A_1^+,s^-,a^-,s^+,a^+) - \hat{H}(\omega,\theta,u,\varepsilon,A_2^-,A_2^+,s^-,a^-,s^+,a^+)\| \\
&\qquad\qquad \le \frac{L_H}{\varepsilon}\big(\|A_1^- - A_2^-\| + \|A_1^+ - A_2^+\|\big), \\
&\|\bar{H}(\omega,\theta_1,u_1,\varepsilon) - \bar{H}(\omega,\theta_2,u_2,\varepsilon)\| \le L_H\big(\|\theta_1-\theta_2\| + \|u_1-u_2\|\big), \\
&\|\bar{D}(\omega,\theta_1,u_1,\varepsilon) - \bar{D}(\omega,\theta_2,u_2,\varepsilon)\| \le L_D\big(\|\theta_1-\theta_2\| + \|u_1-u_2\|\big).
\end{aligned}$$

The following lemma shows that $\bar{D}$ provides an approximation of the hypergradient ∇F, up to an error proportional to ε^2.

Lemma 5 *For any ω, ε,*

$$\|\nabla F(\omega) - \bar{D}(\omega,\theta^\star(\omega),u^\star(\omega,\theta^\star(\omega)),\varepsilon)\| \le \frac{L_V\varepsilon^2}{3}\|u^\star(\omega,\theta^\star(\omega))\|^3.$$

Similarly, Lemma 6 shows that $\hat{H}$ estimates $\hat{H}$ up to an error proportional to ε^2.

Lemma 6 *For any $\omega, \theta, u, \varepsilon$,*

$$\|\bar{H}(\omega,\theta,u,\varepsilon) - \tilde{H}(\omega,\theta,u)\| \le \frac{2L_V\varepsilon^2}{3}\|u\|^3.$$

Lemma 7 shows that the lower-level learning target $\theta^\star$ is Lipschitz in the upper-level control variable.

Lemma 7 *For any ω_1, ω_2, we have*

$$\|\theta^\star(\omega_1) - \theta^\star(\omega_2)\| \le \frac{L_V}{\underline{\lambda}}\|\omega_1 - \omega_2\|.$$

Recall the definition of the bi-level objective F in (5). We show that F is Lipschitz below.

Lemma 8 *Define $L_F = \frac{L_f(\underline{\lambda}+L_V)}{\underline{\lambda}} + \frac{L_f L_V(2\underline{\lambda}+3L_V)}{\underline{\lambda}^2} + \frac{2L_f L_V^2(\underline{\lambda}+2L_V)}{\underline{\lambda}^3}$. Under Assumptions 2 and 1, we have for any ω_1, ω_2*

$$\|\nabla F(\omega_1) - \nabla F(\omega_2)\| \le L_F\|\omega_1 - \omega_2\|.$$

Finally, Lemma 9 establishes the Lipschitz continuity of $u^\star$ and its Jacobian.

Lemma 9 *Define the constant* $L_u \triangleq \frac{4L_f L_V^2}{\underline{\lambda}^3} + \frac{5L_f L_V}{\underline{\lambda}^2} + \frac{L_f}{\underline{\lambda}}$. *For any* $\omega_1, \omega_2, \theta_1, \theta_2$, *we have*

$$\|u^\star(\omega_1, \theta_1) - u^\star(\omega_2, \theta_2)\| \le L_u(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\|),$$
$$\|u^\star(\omega_1, \theta^\star(\omega_1)) - u^\star(\omega_2, \theta^\star(\omega_2))\| \le L_u\|\omega_1 - \omega_2\|.$$

In addition, the Jacobian $\nabla u^\star$ *is* L_u*-Lipschitz with respect to* (ω, θ) *jointly for all* (ω, θ) *that satisfies*

$$\lambda_{\min}(-\nabla^2_{1,1} J_\omega(\pi_\theta, \pi_\theta) - \nabla^2_{2,1} J_\omega(\pi_\theta, \pi_\theta)) \ge \underline{\lambda}. \tag{37}$$

Specifically, for all $\omega_1, \omega_2, \theta_1, \theta_2$ *such that*

$$\lambda_{\min}(-\nabla^2_{1,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1}) - \nabla^2_{2,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1})) \ge \underline{\lambda},$$
$$\lambda_{\min}(-\nabla^2_{1,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2}) - \nabla^2_{2,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2})) \ge \underline{\lambda},$$

we have

$$\|\nabla u^\star(\omega_1, \pi_{\theta_1}) - \nabla u^\star(\omega_2, \pi_{\theta_2})\| \le L_u(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\|).$$

F Proof of Theorem 1

The exact requirements on the step sizes are

$$\alpha_k = \frac{c_\alpha}{(k+1)^{3/5}}, \quad \eta_k = \frac{c_\eta}{(k+1)^{3/5}}, \quad \varepsilon_k = \frac{c_\varepsilon}{(k+1)^{1/10}},$$

with $c_\alpha, c_\eta, c_\varepsilon$ selected such that the following conditions are satisfied

$$\alpha_k \le \eta_k \le \varepsilon_k \le 1, \quad \eta_k \le \frac{4}{\underline{\lambda}},$$

$$\frac{\alpha_k}{\eta_k} \le \min\left\{ \frac{\underline{\lambda}}{8B_D^2}, \frac{\underline{\lambda}}{96L_u^2}, \frac{8(L_f^2 + L_H^2)}{B_D^2 \underline{\lambda}}, \frac{5\underline{\lambda}}{192L_u^2}, \sqrt{\frac{60(B_H^2 + L_f^2)}{B_D^2(6L_u^2 + L_F)}}, \frac{\underline{\lambda}}{8\sqrt{3}L_D L_u}, \frac{\sqrt{L_f^2 + L_H^2}}{2\sqrt{3}L_D L_u} \right\}.$$

We prove Theorem 1 by first characterizing the iteration-wise convergence of u_k and ω_k, in the following propositions. The proofs of Propositions 1 and 2 are deferred to Appendix G.

Proposition 1 *Under the conditions of Theorem 1, the iterates of Algorithm 1 satisfy for all* $k \ge 0$

$$\begin{aligned}\mathbb{E}[\mathrm{res}_{k+1}^u] \le{}& (1 - \frac{\underline{\lambda}\eta_k}{2})\mathbb{E}[\mathrm{res}_k^u] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{16(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] \\ &+ \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \frac{16B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} + \frac{20(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2} \\ &+ \frac{8B_u^6 L_V^2 L_u^2 \alpha_k^2 \varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k} + 3B_D^2 L_u^2 \alpha_k^2.\end{aligned}$$

Proposition 2 *Under the conditions of Theorem 1, the iterates of Algorithm 1 satisfy for all* $k \ge 0$

$$\begin{aligned}&\mathbb{E}[F(\omega_{k+1}) - F(\omega_k)] \\ &\le -\frac{\alpha_k}{2}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + 2B_D^2\alpha_k\mathbb{E}[\mathrm{res}_k^\theta + \mathrm{res}_k^u] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \frac{B_u^6 L_V^2 \varepsilon_k^4 \alpha_k}{9}.\end{aligned}$$

Combining the bounds in Propositions 1 and 2,

$$\begin{aligned}&\mathbb{E}\left[\mathrm{res}_{k+1}^u + \Big(F(\omega_{k+1}) - F(\omega_k)\Big)\right] \\ &\le (1 - \frac{\underline{\lambda}\eta_k}{2})\mathbb{E}[\mathrm{res}_k^u] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{16(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] \\ &\quad + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \frac{16B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} + \frac{20(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2}\end{aligned}$$

$$
\begin{aligned}
&+ \frac{8B_u^6 L_V^2 L_u^2 \alpha_k^2 \varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k} + 3B_D^2 L_u^2 \alpha_k^2 \\
&- \frac{\alpha_k}{2}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + 2B_D^2 \alpha_k \mathbb{E}[\mathrm{res}_k^\theta + \mathrm{res}_k^u] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \frac{B_u^6 L_V^2 \varepsilon_k^4 \alpha_k}{9} \\
\leq\ & (1 - \frac{\underline{\lambda}\eta_k}{4})\mathbb{E}[\mathrm{res}_k^u] - \frac{\alpha_k}{4}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{32(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] \\
&+ \frac{8L_H^2 \eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \frac{16B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} + \frac{20(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2} \\
&+ \frac{32B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k} + 3B_D^2 L_u^2 \alpha_k^2 + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \frac{5B_u^6 L_V^2 \varepsilon_k^4 \alpha_k}{36} \\
\leq\ & (1 - \frac{\underline{\lambda}\eta_k}{4})\mathbb{E}[\mathrm{res}_k^u] - \frac{\alpha_k}{4}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{32(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] \\
&+ \frac{8L_H^2 \eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \frac{32B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} + \frac{40(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2} + \frac{32B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k},
\end{aligned}
$$

where the second and third inequalities are due to the step size conditions $\frac{\alpha_k}{\eta_k} \leq \min\left\{\frac{\underline{\lambda}}{8B_D^2}, \frac{\underline{\lambda}}{96L_u^2}, \frac{8(L_f^2+L_H^2)}{B_D^2\underline{\lambda}}, \frac{5\underline{\lambda}}{192L_u^2}, \sqrt{\frac{60(B_H^2+L_f^2)}{B_D^2(6L_u^2+L_F)}}\right\}$ and $\varepsilon_k \leq 1$.

Re-arranging the terms and taking the telescoping sum,

$$
\begin{aligned}
&\sum_{k=0}^K \alpha_k \mathbb{E}[\|\nabla F(\omega_k)\|^2] \\
&\leq \sum_{k=0}^K 4\mathbb{E}[\mathrm{res}_k^u - \mathrm{res}_{k+1}^u] + \sum_{k=0}^K \frac{128(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] \\
&\quad + \sum_{k=0}^K \frac{32L_H^2 \eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \sum_{k=0}^K \frac{128B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} \\
&\quad + \sum_{k=0}^K \frac{160(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2} + \sum_{k=0}^K \frac{32B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k} \\
&\leq 4\mathbb{E}[\mathrm{res}_0^u] + \epsilon_{\mathrm{MFG}} \sum_{k=0}^K \frac{128(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}} + \epsilon_{\mathrm{MFG}} \sum_{k=0}^K \frac{64L_H^2 \eta_k}{\underline{\lambda}\varepsilon_k^2} \\
&\quad + \sum_{k=0}^K \frac{128B_u^6 L_V^2 L_u^2 \varepsilon_k^4 \eta_k}{3\underline{\lambda}} + \sum_{k=0}^K \frac{160(B_H^2 + L_f^2)\eta_k^2}{\varepsilon_k^2} + \sum_{k=0}^K \frac{32B_D^4 L_u^2 \alpha_k^4 \eta_k}{\underline{\lambda}\varepsilon_k^4},
\end{aligned} \tag{38}
$$

where the second inequality plugs in the oracles in (7) and (8).

We choose $\alpha_k = \frac{c_\alpha}{(k+1)^{3/5}}$, $\eta_k = \frac{c_\eta}{(k+1)^{3/5}}$, $\varepsilon_k = \frac{c_\varepsilon}{(k+1)^{1/10}}$. It is a standard (easily verifiable) result that for all K

$$
\frac{(K+1)^{2/5}}{2} \leq \sum_{k=0}^K \frac{1}{(k+1)^{3/5}} \leq \frac{5(K+1)^{2/5}}{2},
$$

$$
\sum_{k=0}^K \frac{1}{(k+1)^{2/5}} \leq 2(K+1)^{3/5}, \quad \sum_{k=0}^K \frac{1}{k+1} \leq 4\log(K+2).
$$

Then, (38) implies

$$
\min_{k \leq K} \mathbb{E}[\|\nabla F(\omega_k)\|^2] \leq \frac{1}{\sum_{k'=0}^K \alpha_{k'}} \sum_{k=0}^K \alpha_k \mathbb{E}[\|\nabla F(\omega_k)\|^2]
$$

$$\leq \mathcal{O}\Big(\frac{\mathrm{res}_0^u}{(K+1)^{2/5}} + \frac{\log(K+2)}{(K+1)^{2/5}}\Big) + \mathcal{O}\Big(\frac{\epsilon_{\mathrm{MFG}}(K+1)^{3/5}}{(K+1)^{2/5}}\Big)$$
$$\leq \mathcal{O}(\epsilon_{\mathrm{MFG}}(K+1)^{1/5}) + \tilde{\mathcal{O}}\Big((K+1)^{-2/5}\Big).$$

■

G Proof of Propositions

G.1 Proof of Proposition 1

The proof of Proposition 1 leverages the following intermediate result that bounds an important cross term. We defer the proof of the lemma to Appendix H.10.

Lemma 10 *The iterates of Algorithm 1 satisfy for all $k \geq 0$*

$$\begin{aligned}
&\mathbb{E}[\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \theta_k)\Big), \\
&\qquad\qquad u^\star(\omega_k, \theta^\star(\omega_k)) - u^\star(\omega_{k+1}, \theta^\star(\omega_{k+1}))\Big\rangle] \\
&\leq \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big)\|^2] \\
&\quad + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^u] \\
&\quad + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k}.
\end{aligned}$$

By the definition of res_k^u in (36) and the update rule of u_k,

$$\begin{aligned}
&\mathrm{res}_{k+1}^u \\
&\triangleq \|u_{k+1} - u^\star(\omega_{k+1}, \theta^\star(\omega_{k+1}))\|^2 \\
&= \Big\|\mathrm{Proj}_{\|\cdot\|\leq B_u}\Big(u_k + \eta_k\big(\hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+) - \nabla_\theta f(\omega_k, \pi_{\theta_k}, \xi_k)\big)\Big) \\
&\qquad\qquad - u^\star(\omega_{k+1}, \theta^\star(\omega_{k+1}))\Big\|^2 \\
&\leq \Big\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \Big(u^\star(\omega_k, \theta^\star(\omega_k)) - u^\star(\omega_{k+1}, \theta^\star(\omega_{k+1}))\Big) \\
&\quad + \eta_k\Big(\hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+) - \nabla_\theta f(\omega_k, \pi_{\theta_k}, \xi_k)\Big)\Big\|^2 \\
&= \Big\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big) \\
&\quad - \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+)\Big) \\
&\quad + \eta_k\Big(\nabla_\theta f(\omega_k, \pi_{\theta_k}) - \nabla_\theta f(\omega_k, \pi_{\theta_k}, \xi_k)\Big) + \Big(u^\star(\omega_k, \theta^\star(\omega_k)) - u^\star(\omega_{k+1}, \theta^\star(\omega_{k+1}))\Big)\Big\|^2 \\
&\leq \Big\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big)\Big\|^2 \quad ① \\
&\quad + 2\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big), \\
&\qquad\qquad \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+)\Big)\Big\rangle \quad ② \\
&\quad + 2\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big), \\
&\qquad\qquad \eta_k\Big(\nabla_\theta f(\omega_k, \pi_{\theta_k}) - \nabla_\theta f(\omega_k, \pi_{\theta_k}, \xi_k)\Big)\Big\rangle \quad ③ \\
&\quad + 2\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big),
\end{aligned}$$

$$u^\star(\omega_k,\theta^\star(\omega_k)) - u^\star(\omega_{k+1},\theta^\star(\omega_{k+1}))\Big\rangle \quad ④$$
$$+ 3\eta_k^2\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \hat{H}(\omega_k,\theta_k,u_k,\varepsilon_k,A_k^-,A_k^+,s_k^-,a_k^-,s_k^+,a_k^+)\|^2 \quad ⑤$$
$$+ 3\eta_k^2\|\nabla_\theta f(\omega_k,\pi_{\theta_k}) - \nabla_\theta f(\omega_k,\pi_{\theta_k},\xi_k)\|^2 \quad ⑥$$
$$+ 3\|u^\star(\omega_k,\theta^\star(\omega_k)) - u^\star(\omega_{k+1},\theta^\star(\omega_{k+1}))\|^2, \quad ⑦ \tag{39}$$

where the first inequality follows from the fact that $\|u^\star(\omega,\theta)\| \le B_u$ for any ω,θ and that the operator $\mathrm{Proj}_{\|\cdot\|\le B_u}$ is non-expansive.

Bounding ①. We bound each term of (39) individually. First, from the definition of $u^\star$ in (35) and the assumption that $\nabla^2_{1,1}J_\omega(\pi_{\theta^\star(\omega)},\pi_{\theta^\star(\omega)}) + \nabla^2_{2,1}J_\omega(\pi_{\theta^\star(\omega)},\pi_{\theta^\star(\omega)})$ is invertible, we have

$$\Big(\nabla^2_{1,1}J_{\omega_k}(\pi_{\theta^\star(\omega_k)},\pi_{\theta^\star(\omega_k)}) + \nabla^2_{2,1}J_{\omega_k}(\pi_{\theta^\star(\omega_k)},\pi_{\theta^\star(\omega_k)})\Big)u^\star(\omega_k,\theta^\star(\omega_k)) = \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)}).$$

In light of the definition of $\tilde{H}$ in (28), this implies

$$\tilde{H}\Big(\omega_k,\theta^\star(\omega_k),u^\star(\omega_k,\theta^\star(\omega_k))\Big) - \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)}) = 0. \tag{40}$$

By Assumption 1, $-\nabla^2_{1,1}J_\omega(\pi_{\theta^\star(\omega)},\pi_{\theta^\star(\omega)}) - \nabla^2_{2,1}J_\omega(\pi_{\theta^\star(\omega)},\pi_{\theta^\star(\omega)})$ has a symmetric part whose eigenvalues are uniformly lower bounded. Therefore, we have

$$\begin{aligned}
&\Big\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \tilde{H}\Big(\omega_k,\theta^\star(\omega_k),u_k\Big) - \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)})\Big\rangle\\
&= \Big\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \tilde{H}\Big(\omega_k,\theta^\star(\omega_k),u_k\Big) - \tilde{H}\Big(\omega_k,\theta^\star(\omega_k),u^\star(\omega_k,\theta^\star(\omega_k))\Big)\Big\rangle\\
&= \Big(u_k - u^\star(\omega_k,\theta^\star(\omega_k))\Big)^\top\\
&\qquad\cdot\Big(\nabla^2_{1,1}J_{\omega_k}(\pi_{\theta^\star(\omega_k)},\pi_{\theta^\star(\omega_k)}) + \nabla^2_{2,1}J_{\omega_k}(\pi_{\theta^\star(\omega_k)},\pi_{\theta^\star(\omega_k)})\Big)\Big(u_k - u^\star(\omega_k,\theta^\star(\omega_k))\Big)\\
&\le -\underline{\lambda}\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2.
\end{aligned} \tag{41}$$

Then,

$$\begin{aligned}
&①\\
&\le \|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2 + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\rangle\\
&\quad + \eta_k^2\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\|^2\\
&= \|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2 + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k)\rangle\\
&\quad + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k) - \tilde{H}(\omega_k,\theta^\star(\omega_k),u_k)\rangle\\
&\quad + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \tilde{H}(\omega_k,\theta^\star(\omega_k),u_k) - \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)})\rangle\\
&\quad + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)}) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\rangle\\
&\quad + \eta_k^2\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\|^2\\
&\le \|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2 + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k)\rangle\\
&\quad + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k) - \tilde{H}(\omega_k,\theta^\star(\omega_k),u_k)\rangle\\
&\quad - 2\underline{\lambda}\eta_k\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2\\
&\quad + 2\eta_k\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)), \nabla_\theta f(\omega_k,\pi_{\theta^\star(\omega_k)}) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\rangle\\
&\quad + \eta_k^2\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\|^2\\
&\le \|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2\\
&\quad + \frac{\underline{\lambda}\eta_k}{4}\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2 + \frac{4\eta_k}{\underline{\lambda}}\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k)\|^2\\
&\quad + \frac{\underline{\lambda}\eta_k}{4}\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2 + \frac{4\eta_k}{\underline{\lambda}}\|\bar{H}(\omega_k,\theta^\star(\omega_k),u_k,\varepsilon_k) - \tilde{H}(\omega_k,\theta^\star(\omega_k),u_k)\|^2\\
&\quad - 2\underline{\lambda}\eta_k\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{\underline{\lambda}\eta_k}{4}\|u_k - u^\star(\omega_k, \theta^\star(\omega_k))\|^2 + \frac{4\eta_k}{\underline{\lambda}}\|\nabla_\theta f(\omega_k, \pi_{\theta^\star(\omega_k)}) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\|^2 \\
&+ \eta_k^2\|\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\|^2 \\
\leq\ & (1 - \frac{5\underline{\lambda}\eta_k}{4})\|u_k - u^\star(\omega_k, \theta^\star(\omega_k))\|^2 + \frac{4\eta_k}{\underline{\lambda}}\|\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \bar{H}(\omega_k, \theta^\star(\omega_k), u_k, \varepsilon_k)\|^2 \\
&+ \frac{4\eta_k}{\underline{\lambda}}\|\nabla_\theta f(\omega_k, \theta^\star(\omega_k)) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\|^2 \\
&+ \frac{4\eta_k}{\underline{\lambda}}\|\bar{H}(\omega_k, \theta^\star(\omega_k), u_k, \varepsilon_k) - \tilde{H}(\omega_k, \theta^\star(\omega_k), u_k)\|^2 + 2(B_H^2 + L_f^2)\eta_k^2, \qquad (42)
\end{aligned}
$$

where the second inequality follows from (41), and the last inequality is due to the boundedness of $\bar{H}$ (from Lemma 3) and $\nabla_\theta f$ (by Assumption 2).

Due to the L_H-Lipschitz continuity of $\bar{H}$ with respect to θ from Lemma 4,

$$
\|\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \bar{H}(\omega_k, \theta^\star(\omega_k), u_k, \varepsilon_k)\|^2 \leq L_H^2\|\theta_k - \theta^\star(\omega_k)\|^2. \qquad (43)
$$

Similarly, the L_f Lipschitz continuity of $\nabla_\theta f$ implies

$$
\|\nabla_\theta f(\omega_k, \theta^\star(\omega_k)) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\|^2 \leq L_f^2\|\theta_k - \theta^\star(\omega_k)\|^2. \qquad (44)
$$

From Lemma 6,

$$
\|\bar{H}(\omega_k, \theta^\star(\omega_k), u_k, \varepsilon_k) - \tilde{H}(\omega_k, \theta^\star(\omega_k), u_k)\|^2 \leq \frac{4L_V^2\varepsilon_k^4}{9}\|u_k\|^6 \leq \frac{4B_u^6 L_V^2\varepsilon_k^4}{9}. \qquad (45)
$$

Combining (42)-(45),

$$
① \leq (1 - \frac{5\underline{\lambda}\eta_k}{4})\mathrm{res}_k^u + \frac{4(L_f^2 + L_H^2)\eta_k}{\underline{\lambda}}\|\theta_k - \theta^\star(\omega_k)\|^2 + \frac{16B_u^6 L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2 + L_f^2)\eta_k^2. \qquad (46)
$$

Bounding ②. According to (30), we have in expectation

$$
\begin{aligned}
\mathbb{E}[②] &= 2\mathbb{E}\Big[\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big), \\
&\qquad \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+)\Big)\Big\rangle\Big] \\
&= 2\eta_k\mathbb{E}\Big[\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big), \\
&\qquad \mathbb{E}[\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+) \mid \mathcal{F}_{k-1}]\Big\rangle\Big] \\
&= 2\eta_k\mathbb{E}\Big[\Big\langle u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big), \\
&\quad \mathbb{E}[\hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A^{\omega_k, \pi_k^-, \pi_k}, A^{\omega_k, \pi_k^+, \pi_k}, s_k^-, a_k^-, s_k^+, a_k^+) \\
&\qquad - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+) \mid \mathcal{F}_{k-1}]\Big\rangle\Big] \\
&\leq \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}\Big[\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big)\|^2\Big] \\
&\quad + \frac{4\eta_k}{\underline{\lambda}}\mathbb{E}[\|\hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A^{\omega_k, \pi_k^-, \pi_k}, A^{\omega_k, \pi_k^+, \pi_k}, s_k^-, a_k^-, s_k^+, a_k^+) \\
&\qquad - \hat{H}(\omega_k, \theta_k, u_k, \varepsilon_k, A_k^-, A_k^+, s_k^-, a_k^-, s_k^+, a_k^+)\|^2].
\end{aligned}
$$

Due to the L_H/ε-Lipschitz continuity of $\hat{H}$ (with respect to the advantage functions) established in Lemma 4,

$$
\mathbb{E}[②] \leq \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}\Big[\|u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k}))\|^2\Big]
$$

$$
\begin{aligned}
&\quad + \frac{4\eta_k}{\underline{\lambda}}\mathbb{E}\Big[\frac{2L_H^2}{\varepsilon_k^2}(\|A_k^- - A^{\omega_k,\pi_k^-,\pi_k}\|^2 + \|A_k^+ - A^{\omega_k,\pi_k^+,\pi_k}\|^2)\Big] \\
&= \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\textcircled{1}] + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] \\
&\le \frac{\underline{\lambda}\eta_k}{4}(1 - \frac{5\underline{\lambda}\eta_k}{4})\mathbb{E}[\|u_k - u^\star(\omega_k,\theta^\star(\omega_k))\|^2] + \frac{\underline{\lambda}\eta_k}{4}\cdot\frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\|\theta_k - \theta^\star(\omega_k)\|^2] \\
&\quad + \frac{\underline{\lambda}\eta_k}{4}\cdot\frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + \frac{\underline{\lambda}\eta_k}{4}\cdot 2(B_H^2+L_f^2)\eta_k^2 + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] \\
&\le \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\mathrm{res}_k^u] + \frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\|\theta_k - \theta^\star(\omega_k)\|^2] + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] \\
&\quad + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2,
\end{aligned}
$$

where the second inequality plugs in the bound on ① from (46), and the last inequality follows from the step size condition $\eta_k \le \frac{4}{\underline{\lambda}}$.

Bounding ③. Note that $u_k - u^\star(\omega_k,\theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\Big)$ only depends on the randomness up to time $k-1$, while $\nabla_\theta f(\omega_k,\pi_{\theta_k}) - \nabla_\theta f(\omega_k,\pi_{\theta_k},\xi_k)$ is zero in expectation over ξ_k. By the law of total expectation,

$$
\begin{aligned}
\mathbb{E}[\textcircled{3}] = \mathbb{E}\Big[2\Big\langle u_k - u^\star(\omega_k,\theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k})\Big), \\
\eta_k\mathbb{E}\Big[\nabla_\theta f(\omega_k,\theta_k) - \nabla_\theta f(\omega_k,\pi_{\theta_k},\xi_k) \mid \mathcal{F}_{k-1}\Big]\Big\rangle\Big] \\
= 0.
\end{aligned}
$$

Bounding ④. By Lemma 10, we have

$$
\begin{aligned}
\mathbb{E}[\textcircled{4}] &\le \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\textcircled{1}] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^\theta] \\
&\quad + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^u] + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k} \\
&\le \frac{\underline{\lambda}\eta_k}{4}\cdot(1 - \frac{5\underline{\lambda}\eta_k}{4})\mathbb{E}[\mathrm{res}_k^u] + \frac{\underline{\lambda}\eta_k}{4}\cdot\frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{\underline{\lambda}\eta_k}{4}\cdot\frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} \\
&\quad + \frac{\underline{\lambda}\eta_k}{4}\cdot 2(B_H^2+L_f^2)\eta_k^2 + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^\theta] \\
&\quad + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^u] + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k} \\
&\le \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\mathrm{res}_k^u] + \frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2 \\
&\quad + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^\theta] \\
&\quad + \frac{48L_D^2L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^u] + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k} \\
&\le \frac{\underline{\lambda}\eta_k}{2}\mathbb{E}[\mathrm{res}_k^u] + \frac{8(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] \\
&\quad + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2 + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k},
\end{aligned}
$$

where the second inequality again plugs in the bound on ① from (46), the third inequality follows from the step size condition $\eta_k \leq \frac{4}{\underline{\lambda}}$, and the last inequality simplifies and combines terms under the condition $\frac{\alpha_k}{\eta_k} \leq \min\{\frac{\underline{\lambda}}{8\sqrt{3}L_D L_u}, \frac{\sqrt{L_f^2+L_H^2}}{2\sqrt{3}L_D L_u}\}$.

Bounding ⑤. Since $\|\bar{H}(\omega,\theta,u,\varepsilon)\| \leq B_H$ and $\|\hat{H}(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,a^+,a^+)\| \leq \frac{B_H}{\varepsilon}$, we have

$$\begin{aligned}⑤ &\leq 3\eta_k^2\Big(2\|\bar{H}(\omega_k,\theta_k,u_k,\varepsilon_k)\|^2 + 2\|\hat{H}(\omega_k,\theta_k,u_k,\varepsilon_k,A_k,A_k^+,s_k,a_k,s_k^+,a_k^+)\|^2\Big)\\ &\leq \frac{12B_H^2\eta_k^2}{\varepsilon_k^2},\end{aligned}$$

under the step size condition $\varepsilon_k \leq 1$.

Bounding ⑥. Similarly, we have from the boundedness of $\nabla_\theta f$ and $\tilde{\nabla}_\theta f$

$$\begin{aligned}⑥ &\leq 3\eta_k^2\Big(2\|\nabla_\theta f(\omega_k,\pi_{\theta_k})\| + \|\nabla_\theta f(\omega_k,\pi_{\theta_k},\xi_k)\|^2\Big)\\ &\leq 12L_f^2\eta_k^2.\end{aligned}$$

Bounding ⑦. From Lemma 9,

$$⑦ \leq 3L_u^2\|\omega_k - \omega_{k+1}\|^2 \leq 3B_D^2L_u^2\alpha_k^2.$$

Combining the bounds on ①-⑦, we have in expectation

$$\begin{aligned}\mathbb{E}[\mathrm{res}_{k+1}^u] &\leq (1-\frac{5\underline{\lambda}\eta_k}{4})\mathbb{E}[\mathrm{res}_k^u] + \frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2\\ &\quad + \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\mathrm{res}_k^u] + \frac{4(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big]\\ &\quad + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2\\ &\quad + \frac{\underline{\lambda}\eta_k}{2}\mathbb{E}[\mathrm{res}_k^u] + \frac{8(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2]\\ &\quad + \frac{16B_u^6L_V^2\varepsilon_k^4\eta_k}{9\underline{\lambda}} + 2(B_H^2+L_f^2)\eta_k^2 + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k}\\ &\quad + \frac{12B_H^2\eta_k^2}{\varepsilon_k^2} + 12L_f^2\eta_k^2 + 3B_D^2L_u^2\alpha_k^2\\ &\leq (1-\frac{\underline{\lambda}\eta_k}{2})\mathbb{E}[\mathrm{res}_k^u] + \frac{24L_u^2\alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{16(L_f^2+L_H^2)\eta_k}{\underline{\lambda}}\mathbb{E}[\mathrm{res}_k^\theta]\\ &\quad + \frac{8L_H^2\eta_k}{\underline{\lambda}\varepsilon_k^2}\mathbb{E}\Big[\mathrm{res}_k^{A,-} + \mathrm{res}_k^{A,+}\Big] + \frac{16B_u^6L_V^2L_u^2\varepsilon_k^4\eta_k}{3\underline{\lambda}} + \frac{20(B_H^2+L_f^2)\eta_k^2}{\varepsilon_k^2}\\ &\quad + \frac{8B_u^6L_V^2L_u^2\alpha_k^2\varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32B_D^4L_u^2\alpha_k^4}{\underline{\lambda}\varepsilon_k^4\eta_k} + 3B_D^2L_u^2\alpha_k^2.\end{aligned}$$

■

G.2 Proof of Proposition 2

It is shown in Lemma 8 that F has L_F-Lipschitz gradients. This implies

$$\begin{aligned}&F(\omega_{k+1}) - F(\omega_k)\\ &\leq \langle\nabla F(\omega_k), \omega_{k+1}-\omega_k\rangle + \frac{L_F}{2}\|\omega_{k+1}-\omega_k\|^2\end{aligned}$$

$$
\begin{aligned}
&= -\alpha_k \langle \nabla F(\omega_k), \hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\rangle + \frac{L_F \alpha_k^2}{2} \|\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\|^2 \\
&= -\alpha_k \|\nabla F(\omega_k)\|^2 + \alpha_k \langle \nabla F(\omega_k), \nabla F(\omega_k) - \hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\rangle \\
&\quad + \frac{L_F \alpha_k^2}{2} \|\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\|^2. \qquad (47)
\end{aligned}
$$

By the law of total expectation,

$$
\begin{aligned}
&\mathbb{E}[F(\omega_{k+1}) - F(\omega_k)] \\
&\le -\alpha_k \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{L_F \alpha_k^2}{2} \mathbb{E}[\|\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\|^2] \\
&\quad + \alpha_k \mathbb{E}[\langle \nabla F(\omega_k), \nabla F(\omega_k) - \mathbb{E}[\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) \mid \mathcal{F}_{k-1}]\rangle] \\
&= -\alpha_k \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{L_F \alpha_k^2}{2} \mathbb{E}[\|\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k)\|^2] \\
&\quad + \alpha_k \mathbb{E}[\langle \nabla F(\omega_k), \nabla F(\omega_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\rangle] \\
&\le -\alpha_k \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} \\
&\quad + \frac{\alpha_k}{2} \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{\alpha_k}{2} \mathbb{E}[\|\nabla F(\omega_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\|^2] \\
&\le -\frac{\alpha_k}{2} \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \alpha_k \mathbb{E}[\|\nabla F(\omega_k) - \bar{D}(\omega_k, \theta^\star(\omega_k), u(\omega_k, \theta^\star(\omega_k)), \varepsilon_k)\|^2] \\
&\quad + \alpha_k \mathbb{E}[\|\bar{D}(\omega_k, \theta^\star(\omega_k), u(\omega_k, \theta^\star(\omega_k)), \varepsilon_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\|^2] \\
&\le -\frac{\alpha_k}{2} \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \alpha_k \mathbb{E}[\frac{L_V^2 \varepsilon_k^4}{9} \|u^\star(\omega_k, \theta^\star(\omega_k))\|^6] \\
&\quad + 2B_D^2 \alpha_k \mathbb{E}[\|\theta_k - \theta^\star(\omega_k)\|^2 + \|u_k - u(\omega_k, \theta^\star(\omega_k))\|^2] \\
&\le -\frac{\alpha_k}{2} \mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{B_D^2 L_F \alpha_k^2}{2\varepsilon_k^2} + \frac{B_u^6 L_V^2 \varepsilon_k^4 \alpha_k}{9} + 2B_D^2 \alpha_k \mathbb{E}[\mathrm{res}_k^\theta + \mathrm{res}_k^u],
\end{aligned}
$$

where the fourth inequality is due to Lemma 5 and the Lipschitz continuity of $\bar{D}$.

■

H Proof of Lemmas

H.1 Proof of Lemma 1

The lower-level first-order equilibrium condition is

$$\nabla_1 J_\omega\big(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}\big) = 0,$$

where ∇_1 denotes differentiation with respect to the parameter of the first policy argument. Differentiating both sides with respect to ω yields

$$\nabla^2_{1,\omega} J_\omega + \nabla^2_{1,1} J_\omega \frac{\partial \theta^\star(\omega)}{\partial \omega} + \nabla^2_{1,2} J_\omega \frac{\partial \theta^\star(\omega)}{\partial \omega} = 0,$$

where all derivatives of J_ω are evaluated at $(\theta^\star(\omega), \theta^\star(\omega))$. Hence, we have

$$\frac{\partial \theta^\star(\omega)}{\partial \omega} = -\left(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\right)^{-1} \nabla^2_{1,\omega} J_\omega. \qquad (48)$$

Applying the chain rule,

$$\nabla F(\omega) = \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) + \left(\frac{\partial \theta^\star(\omega)}{\partial \omega}\right)^\top \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}). \qquad (49)$$

Substituting the above expression for $\partial\theta^\star(\omega)/\partial\omega$ gives

$$\begin{aligned}\nabla F(\omega) &= \nabla_\omega f\big(\omega, \pi_{\theta^\star(\omega)}\big) - \big(\nabla^2_{1,\omega} J_\omega\big)^\top \big(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\big)^{-\top} \nabla_\theta f\big(\omega, \pi_{\theta^\star(\omega)}\big) \\ &= \nabla_\omega f\big(\omega, \pi_{\theta^\star(\omega)}\big) - \nabla^2_{\omega,1} J_\omega \big(\nabla^2_{1,1} J_\omega + \nabla^2_{2,1} J_\omega\big)^{-1} \nabla_\theta f\big(\omega, \pi_{\theta^\star(\omega)}\big),\end{aligned}$$

where the last equality follows from

$$\big(\nabla^2_{1,\omega} J_\omega\big)^\top = \nabla^2_{\omega,1} J_\omega, \quad \big(\nabla^2_{1,2} J_\omega\big)^\top = \nabla^2_{2,1} J_\omega,$$

since J_ω is twice continuously differentiable.

■

H.2 Proof of Lemma 2

We skip the proof as many variants of the lemmas have been established in the literature. For example, with $\pi_1' = \pi_2'$ (i.e. the problem reduces to a bi-level optimization problem where the lower-level objective is the standard RL objective), Zeng et al. (2026a)[Lemma 3] shows the Lipschitz continuity of J and the value function, as well as their (higher-order) gradients, with respect to ω and the first θ argument. The additional work required here is to show the Lipschitz continuity with respect to the second θ argument, which is trivial given that the reference policy does not affect the state transition and the reward model is Lipschitz in the reference policy by Assumption 3.

H.3 Proof of Lemma 3

We first bound $\hat{H}$. By definition,

$$\hat{H} = \hat{H}_1 + \hat{H}_2,$$

where

$$\hat{H}_1 = \frac{A^+(s^+, a^+)\nabla_\theta \log \pi_{\theta+\varepsilon u}(a^+ \mid s^+) - A^-(s^-, a^-)\nabla_\theta \log \pi_{\theta-\varepsilon u}(a^- \mid s^-)}{2(1-\gamma)\varepsilon},$$

and

$$\hat{H}_2 = \frac{\nabla_{\theta'} r_\omega(s^+, a^+ \mid \pi_{\theta'}) - \nabla_{\theta'} r_\omega(s^-, a^- \mid \pi_{\theta'})}{2(1-\gamma)\varepsilon}.$$

Due to $-B_V \le A^+(s^+, a^+), A^-(s^-, a^-) \le B_V$ and the fact that $\|\nabla_\theta \log \pi_\theta\| \le 2$ for any θ under the softmax parameter, we have

$$\|\hat{H}_1\| \le \frac{2B_V}{(1-\gamma)\varepsilon}.$$

Assumption 3 implies $\|\nabla_\theta r_\omega(s, a \mid \pi_\theta)\| \le L_r$ for all ω, θ, s, a, which further leads to

$$\|\hat{H}_2\| \le \frac{L_r}{(1-\gamma)\varepsilon}.$$

Hence, we have

$$\|\hat{H}\| \le \frac{2B_V + L_r}{(1-\gamma)\varepsilon}.$$

Recall the definition of $\hat{G}$

$$\hat{G}(\omega, \theta, \varepsilon, s^-, a^-, s^+, a^+) = \frac{\nabla_\omega r_\omega(s^+, a^+ \mid \pi_{\theta'}) - \nabla_\omega r_\omega(s^-, a^- \mid \pi_{\theta'})}{2(1-\gamma)\varepsilon}.$$

Assumption 3 implies $\|\nabla_\omega r_\omega(s, a \mid \pi_\theta)\| \le L_r$ for all ω, θ, s, a. Therefore,

$$\|\hat{G}\| \le \frac{L_r}{(1-\gamma)\varepsilon}.$$

For $\hat{D}$,

$$\begin{aligned}\|\hat{D}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+,\xi)\| &= \|\nabla_\omega f(\omega,\pi_\theta,\xi) - \hat{G}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+)\| \\ &\le \|\nabla_\omega f(\omega,\pi_\theta,\xi)\| + \|\hat{G}(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+)\| \\ &\le L_f + \frac{L_r}{(1-\gamma)\varepsilon} \\ &\le \frac{(1-\gamma)L_f + L_r}{(1-\gamma)\varepsilon},\end{aligned}$$

where the second inequality follows from Assumption 2, and the last inequality is due to $\varepsilon \in (0,1]$.

We then prove the bounds on $\bar{H}, \bar{G}, \bar{D}$. Recall the definition of $\bar{H}$ in (26). By Lemma 2, the maps $\theta \mapsto \nabla_1 J_\omega(\pi_\theta, \pi_{\theta'})$, $\theta \mapsto \nabla_2 J_\omega(\pi_\theta, \pi_{\theta'})$, and $\theta \mapsto \nabla_\omega J_\omega(\pi_\theta, \pi_{\theta'})$ are L_V-Lipschitz in the policy parameters. by the mean-value theorem,

$$\|\bar{H}_1(\omega,\theta,u,\varepsilon)\| = \left\|\frac{\nabla_1 J_\omega(\pi_{\theta+\varepsilon u},\pi_\theta) - \nabla_1 J_\omega(\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}\right\| \le L_V\|u\| \le B_u L_r,$$

$$\|\bar{H}_2(\omega,\theta,u,\varepsilon)\| = \left\|\frac{\nabla_2 J_\omega(\pi_{\theta+\varepsilon u},\pi_\theta) - \nabla_2 J_\omega(\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}\right\| \le L_V\|u\| \le B_u L_V,$$

and hence

$$\|\bar{H}(\omega,\theta,u,\varepsilon)\| \le \|\bar{H}_1(\omega,\theta,u,\varepsilon)\| + \|\bar{H}_2(\omega,\theta,u,\varepsilon)\| \le 2B_u L_V.$$

Similarly,

$$\|\bar{G}(\omega,\theta,u,\varepsilon)\| = \left\|\frac{\nabla_\omega J_\omega(\pi_{\theta+\varepsilon u},\pi_\theta) - \nabla_\omega J_\omega(\pi_{\theta-\varepsilon u},\pi_\theta)}{2\varepsilon}\right\| \le L_V\|u\| \le B_u L_V.$$

For $\bar{D}$, we have from (34)

$$\begin{aligned}\|\bar{D}(\omega,\theta,u,\varepsilon,\xi)\| &= \|\nabla_\omega f(\omega,\pi_\theta,\xi) - \bar{G}(\omega,\theta,u,\varepsilon)\| \\ &\le L_f + \|\bar{G}(\omega,\theta,u,\varepsilon)\| \\ &\le L_f + B_u L_V,\end{aligned}$$

where the first inequality follows from Assumption 2.

H.4 Proof of Lemma 4

We prove the three claimed bounds separately.

Lipschitz continuity of $\hat{H}$ with respect to the advantage inputs. Fix $\omega, \theta, u, \varepsilon$ and the samples (s^-, a^-, s^+, a^+). By definition,

$$\begin{aligned}&\hat{H}(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+) \\ &= \hat{H}_1(\omega,\theta,u,\varepsilon,A^-,A^+,s^-,a^-,s^+,a^+) + \hat{H}_2(\omega,\theta,\varepsilon,s^-,a^-,s^+,a^+),\end{aligned}$$

where only $\hat{H}_1$ depends on A^- and A^+. Hence, for any two pairs (A_1^-, A_1^+) and (A_2^-, A_2^+),

$$\begin{aligned}&\|\hat{H}(\omega,\theta,u,\varepsilon,A_1^-,A_1^+,s^-,a^-,s^+,a^+) - \hat{H}(\omega,\theta,u,\varepsilon,A_2^-,A_2^+,s^-,a^-,s^+,a^+)\| \\ &= \|\hat{H}_1(\omega,\theta,u,\varepsilon,A_1^-,A_1^+,s^-,a^-,s^+,a^+) - \hat{H}_1(\omega,\theta,u,\varepsilon,A_2^-,A_2^+,s^-,a^-,s^+,a^+)\| \\ &\le \frac{1}{2(1-\gamma)\varepsilon}\Big(|A_1^+(s^+,a^+) - A_2^+(s^+,a^+)| \cdot \|\nabla_\theta \log \pi_{\theta+\varepsilon u}(a^+ \mid s^+)\| \\ &\qquad\qquad + |A_1^-(s^-,a^-) - A_2^-(s^-,a^-)| \cdot \|\nabla_\theta \log \pi_{\theta-\varepsilon u}(a^- \mid s^-)\|\Big) \\ &\le \frac{1}{(1-\gamma)\varepsilon}\Big(\|A_1^- - A_2^-\| + \|A_1^+ - A_2^+\|\Big),\end{aligned}$$

where the last inequality is due to $\|\nabla_\theta \log \pi_\theta(a \mid s)\| \le 2$ for any θ under the softmax parameterization.

Lipschitz continuity of $\bar{H}$ in (θ, u). Recall that

$$\bar{H}(\omega, \theta, u, \varepsilon) = \bar{H}_1(\omega, \theta, u, \varepsilon) + \bar{H}_2(\omega, \theta, u, \varepsilon),$$

where for each $i \in \{1, 2\}$,

$$\bar{H}_i(\omega, \theta, u, \varepsilon) = \frac{1}{2\varepsilon}\Big(\nabla_i J_\omega(\pi_{\theta+\varepsilon u}, \pi_\theta) - \nabla_i J_\omega(\pi_{\theta-\varepsilon u}, \pi_\theta)\Big).$$

By the fundamental theorem of calculus,

$$\bar{H}_i(\omega, \theta, u, \varepsilon) = \frac{1}{2}\int_0^1 \Big(\nabla^2_{i,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{i,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta)\Big) u \, dt.$$

For brevity, we define $M_i(t; \theta, u) \triangleq \nabla^2_{i,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{i,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta)$. Then,

$$\bar{H}_i(\omega, \theta, u, \varepsilon) = \frac{1}{2}\int_0^1 M_i(t; \theta, u) u \, dt.$$

This implies

$$\begin{aligned}
&\bar{H}_i(\omega, \theta_1, u_1, \varepsilon) - \bar{H}_i(\omega, \theta_2, u_2, \varepsilon) \\
&= \frac{1}{2}\int_0^1 M_i(t; \theta_1, u_1)(u_1 - u_2)\, dt + \frac{1}{2}\int_0^1 \big(M_i(t; \theta_1, u_1) - M_i(t; \theta_2, u_2)\big) u_2 \, dt.
\end{aligned}$$

We bound the two terms separately.

For the first term, Lemma 2 implies that the relevant second derivatives of J_ω are uniformly bounded. Therefore,

$$\left\| \frac{1}{2}\int_0^1 M_i(t; \theta_1, u_1)(u_1 - u_2)\, dt \right\| \le \frac{L_V}{2}\|u_1 - u_2\|.$$

For the second term, Lemma 2 gives that $\nabla^2_{i,1} J_\omega$ is Lipschitz in its policy arguments with constant L_V. Due to the 1-Lipschitz continuity of a policy in its softmax parameter and $\varepsilon \le 1$,

$$\|\pi_{\theta_1 \pm t\varepsilon u_1} - \pi_{\theta_2 \pm t\varepsilon u_2}\| \le \|\theta_1 - \theta_2\| + t\varepsilon\|u_1 - u_2\| \le \|\theta_1 - \theta_2\| + \|u_1 - u_2\|.$$

Hence, for $t \in [0, 1]$,

$$\|M_i(t; \theta_1, u_1) - M_i(t; \theta_2, u_2)\| \le 2L_V\big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\big).$$

With $\|u_2\| \le B_u$, we obtain

$$\left\| \frac{1}{2}\int_0^1 \big(M_i(t; \theta_1, u_1) - M_i(t; \theta_2, u_2)\big) u_2 \, dt \right\| \le L_V B_u\big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\big).$$

Combining the two bounds yields

$$\|\bar{H}_i(\omega, \theta_1, u_1, \varepsilon) - \bar{H}_i(\omega, \theta_2, u_2, \varepsilon)\| \le L_V(B_u + \frac{1}{2})\Big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\Big).$$

Summing over $i = 1, 2$ gives

$$\|\bar{H}(\omega, \theta_1, u_1, \varepsilon) - \bar{H}(\omega, \theta_2, u_2, \varepsilon)\| \le L_V(2B_u + 1)\Big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\Big).$$

Lipschitz continuity of $\bar{D}$ in (θ, u). By definition of $\bar{D}$ in (33),

$$\begin{aligned}
&\|\bar{D}(\omega, \theta_1, u, \varepsilon) - \bar{D}(\omega, \theta_2, u, \varepsilon)\| \\
&\le \|\nabla_\omega f(\omega, \pi_{\theta_1}) - \nabla_\omega f(\omega, \pi_{\theta_2})\| + \|\bar{G}(\omega, \theta_1, u, \varepsilon) - \bar{G}(\omega, \theta_2, u, \varepsilon)\|.
\end{aligned}$$

By Assumption 2,

$$\|\nabla_\omega f(\omega, \pi_{\theta_1}) - \nabla_\omega f(\omega, \pi_{\theta_2})\| \le L_f\|\theta_1 - \theta_2\|.$$

Moreover, from the same argument used for showing the Lipschitz continuity of $\bar{H}_i$, we can show

$$\|\bar{G}(\omega, \theta_1, u_1, \varepsilon) - \bar{G}(\omega, \theta_2, u_2, \varepsilon)\| \le L_V(B_u + \frac{1}{2})\Big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\Big).$$

Combining the above bounds yields

$$\|\bar{D}(\omega, \theta_1, u_1, \varepsilon) - \bar{D}(\omega, \theta_2, u_2, \varepsilon)\| \le (L_f + B_u L_V + \frac{L_V}{2})\Big(\|\theta_1 - \theta_2\| + \|u_1 - u_2\|\Big).$$

■

■

H.5 Proof of Lemma 5

By Lemma 1 and (34),

$$\begin{aligned}
&\nabla F(\omega) - \bar{D}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega)), \varepsilon) \\
&= \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) - \nabla^2_{\omega,1} J_\omega \left(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\right)^{-1} \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}) \\
&\quad - \Big(\nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) - \bar{G}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega)), \varepsilon)\Big) \\
&= \bar{G}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega)), \varepsilon) - \tilde{G}\left(\omega, \theta^\star(\omega), \left(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\right)^{-1} \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)})\right) \\
&= \bar{G}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega)), \varepsilon) - \tilde{G}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega))),
\end{aligned} \tag{50}$$

where the second equality follows from the definition of $\tilde{G}$, and the third equality follows from the definition of $u^\star$ in (35).

By the fundamental theorem of calculus,

$$\nabla_\omega J_\omega(\pi_{\theta+\varepsilon u}, \pi) - \nabla_\omega J_\omega(\pi_{\theta-\varepsilon u}, \pi) = \int_0^1 \nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi)(\varepsilon u)\, dt \tag{51}$$

$$+ \int_0^1 \nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi)(\varepsilon u)\, dt. \tag{52}$$

Recall the definition of $\bar{G}$ in (27). Then, (52) implies

$$\bar{G}(\omega, \theta, u, \varepsilon) = \int_0^1 \frac{1}{2}\Big(\nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta)\Big) u\, dt.$$

Therefore,

$$\begin{aligned}
&\bar{G}(\omega, \theta, u, \varepsilon) - \tilde{G}(\omega, \theta, u) \\
&= \int_0^1 \frac{1}{2}\Big(\nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi_\theta)\Big) u\, dt.
\end{aligned}$$

Since $\nabla^3_{\omega,1,1} J_\omega$ is L_V-Lipschitz, we have

$$\nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) - \nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi_\theta) = \nabla^3_{\omega,1,1} J_\omega(\pi_\theta, \pi)(t\varepsilon u) + T^+, \tag{53}$$

where the higher-order term T^+ satisfies $\|T^+\| \le L_V \|t\varepsilon u\|^2$.

Similarly,

$$\nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - \nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi) = -\nabla^3_{\omega,1,1} J_\omega(\pi_\theta, \pi_\theta)(t\varepsilon u) + T^-, \tag{54}$$

where $\|T^-\| \le L_V \|t\varepsilon u\|^2$.

The first terms on the right hand side of (53) and (54) cancel, and we get

$$\nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi) = T^+ + T^-. \tag{55}$$

Therefore,

$$\begin{aligned}
&\|\bar{G}(\omega, \theta, u, \varepsilon) - \tilde{G}(\omega, \theta, u)\| \\
&\le \int_0^1 \frac{1}{2}\Big\|\nabla^2_{\omega,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{\omega,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{\omega,1} J_\omega(\pi_\theta, \pi_\theta)\Big\| \|u\|\, dt \\
&\le \int_0^1 \frac{1}{2} \cdot 2L_V(\|t\varepsilon u\|)^2 \|u\| dt \\
&\le L_V \varepsilon^2 \|u\|^3 \int_0^1 t^2 dt \\
&= \frac{L_V \varepsilon^2}{3} \|u\|^3.
\end{aligned}$$

Combining this with (50),

$$\|\nabla F(\omega) - \bar{D}(\omega, \theta^\star(\omega), u^\star(\omega, \theta^\star(\omega)), \varepsilon)\| \le \frac{L_V \varepsilon^2}{3} \|u^\star(\omega, \theta^\star(\omega))\|^3.$$

■

H.6 Proof of Lemma 6

By the fundamental theorem of calculus,

$$\begin{aligned}
&\nabla_1 J_\omega(\pi_{\theta+\varepsilon u}, \pi_\theta) - \nabla_1 J_\omega(\pi_{\theta-\varepsilon u}, \pi_\theta) \\
&= \int_0^1 \nabla^2_{1,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi)(\varepsilon u)\, dt + \int_0^1 \nabla^2_{1,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta)(\varepsilon u)\, dt. \qquad (56)
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\nabla_2 J_\omega(\pi_{\theta+\varepsilon u}, \pi_\theta) - \nabla_2 J_\omega(\pi_{\theta-\varepsilon u}, \pi_\theta) \\
&= \int_0^1 \nabla^2_{2,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi)(\varepsilon u)\, dt + \int_0^1 \nabla^2_{2,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta)(\varepsilon u)\, dt. \qquad (57)
\end{aligned}$$

Recall the definitions of $\bar{H}, \tilde{H}$ in (26) and (28). We have from (56)-(57)

$$\begin{aligned}
&\bar{H}(\omega, \theta, u, \varepsilon) - \tilde{H}(\omega, \theta, u) \\
&= \frac{1}{2} \int_0^1 \Big(\nabla^2_{1,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{1,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{1,1} J_\omega(\pi_\theta, \pi_\theta) \\
&\quad + \nabla^2_{2,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{2,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{2,1} J_\omega(\pi_\theta, \pi_\theta) \Big) u\, dt. \qquad (58)
\end{aligned}$$

Note that $\nabla^3_{1,1,1} J_\omega(\pi_\theta, \pi_\theta)$ is L_V-Lipschitz continuous in the first argument. By an argument similar to that in the proof of Lemma 5, we have

$$\begin{aligned}
\|\nabla^2_{1,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{1,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{1,1} J_\omega(\pi_\theta, \pi_\theta)\| &\le 2L_V \|t\varepsilon u\|^2, \\
\|\nabla^2_{2,1} J_\omega(\pi_{\theta+t\varepsilon u}, \pi_\theta) + \nabla^2_{2,1} J_\omega(\pi_{\theta-t\varepsilon u}, \pi_\theta) - 2\nabla^2_{2,1} J_\omega(\pi_\theta, \pi_\theta)\| &\le 2L_V \|t\varepsilon u\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\bar{H}(x, \theta, \mu, \varepsilon) - \tilde{H}(x, \theta, \mu)\| &\le \frac{1}{2} \int_0^1 \Big(2L_V (\|t\varepsilon u\|)^2 + 2L_V (\|t\varepsilon u\|)^2 \Big) \|u\| dt \\
&= 2L_V \varepsilon^2 \|u\|^3 \int_0^1 t^2 dt \\
&= \frac{2L_V \varepsilon^2}{3} \|u\|^3.
\end{aligned}$$

■

H.7 Proof of Lemma 7

Recall from (48) that for all ω

$$\frac{\partial \theta^\star(\omega)}{\partial \omega} = -\left(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\right)^{-1} \nabla^2_{1,\omega} J_\omega, \qquad (59)$$

where all derivatives are evaluated at $\left(\pi_{\theta^\star(\omega)}, \pi_{\theta^\star(\omega)}\right)$.

Therefore,

$$\left\| \frac{\partial \theta^\star(\omega)}{\partial \omega} \right\| \le \left\| \left(\nabla^2_{1,1} J_\omega + \nabla^2_{1,2} J_\omega\right)^{-1} \right\| \|\nabla^2_{1,\omega} J_\omega\| \le \frac{L_V}{\underline{\lambda}}, \qquad (60)$$

which implies

$$\|\theta^\star(\omega_1) - \theta^\star(\omega_2)\| \le \frac{L_V}{\underline{\lambda}} \|\omega_1 - \omega_2\|.$$

■

H.8 Proof of Lemma 8

In this section, we use the shorthand notation

$$A(\omega) \triangleq \nabla^2_{1,1} J_\omega(\pi^\star(\omega), \pi^\star(\omega)) + \nabla^2_{2,1} J_\omega(\pi^\star(\omega), \pi^\star(\omega)).$$

By Assumption 1, $A(\omega)$ is invertible, with

$$\|A(\omega)^{-1}\| \le \frac{1}{\underline{\lambda}}.$$

By Lemma 1, we have for all ω

$$\nabla F(\omega) = \nabla_\omega f(\omega, \pi_{\theta^\star(\omega)}) - \nabla^2_{\omega,1} J_\omega(\pi^\star(\omega), \pi^\star(\omega))\, A(\omega)^{-1}\, \nabla_\theta f(\omega, \pi_{\theta^\star(\omega)}).$$

This implies for any ω_1, ω_2

$$\begin{aligned}
\|\nabla F(\omega_1) - \nabla F(\omega_2)\| \le{}& \|\nabla_\omega f(\omega_1, \pi_{\theta^\star(\omega_1)}) - \nabla_\omega f(\omega_2, \pi_{\theta^\star(\omega_2)})\| \\
&+ \Big\| \nabla^2_{\omega,1} J_{\omega_1}(\pi_{\theta^\star(\omega_1)}, \pi_{\theta^\star(\omega_1)}) A(\omega_1)^{-1} \nabla_\theta f(\omega_1, \pi_{\theta^\star(\omega_1)}) \\
&\quad - \nabla^2_{\omega,1} J_{\omega_2}(\pi_{\theta^\star(\omega_2)}, \pi_{\theta^\star(\omega_2)}) A(\omega_2)^{-1} \nabla_\theta f(\omega_2, \pi_{\theta^\star(\omega_2)}) \Big\|. \quad (61)
\end{aligned}$$

To bound the first term of (61), we have from the Lipschitz continuity of $\nabla_\omega f$

$$\begin{aligned}
\|\nabla_\omega f(\omega_1, \pi_{\theta^\star(\omega_1)}) - \nabla_\omega f(\omega_2, \pi_{\theta^\star(\omega_2)})\| &\le L_f \Big(\|\omega_1 - \omega_2\| + \|\pi_{\theta^\star(\omega_1)} - \pi_{\theta^\star(\omega_2)}\| \Big) \\
&\le L_f \Big(\|\omega_1 - \omega_2\| + \|\theta^\star(\omega_1) - \theta^\star(\omega_2)\| \Big) \\
&\le L_f (\|\omega_1 - \omega_2\| + \frac{L_V}{\underline{\lambda}} \|\omega_1 - \omega_2\|) \\
&= \frac{L_f(\underline{\lambda} + L_V)}{\underline{\lambda}} \|\omega_1 - \omega_2\|, \quad (62)
\end{aligned}$$

where the second inequality is due to the 1-Lipschitz continuity of a policy with respect to its softmax parameters, and the third inequality is due to Lemma 7.

For the second term of (61),

$$\begin{aligned}
&\Big\| \nabla^2_{\omega,1} J_{\omega_1} A(\omega_1)^{-1} \nabla_\theta f(\omega_1, \pi_{\theta^\star(\omega_1)}) - \nabla^2_{\omega,1} J_{\omega_2} A(\omega_2)^{-1} \nabla_\theta f(\omega_2, \pi_{\theta^\star(\omega_2)}) \Big\| \\
&\le \|\nabla^2_{\omega,1} J_{\omega_1}(\pi_{\theta^\star(\omega_1)}, \pi_{\theta^\star(\omega_1)}) - \nabla^2_{\omega,1} J_{\omega_2}(\pi_{\theta^\star(\omega_2)}, \pi_{\theta^\star(\omega_2)})\| \cdot \|A(\omega_1)^{-1}\| \cdot \|\nabla_\theta f(\omega_1, \pi_{\theta^\star(\omega_1)})\| \\
&\quad + \|\nabla^2_{\omega,1} J_{\omega_2}\| \cdot \|A(\omega_1)^{-1} - A(\omega_2)^{-1}\| \cdot \|\nabla_\theta f(\omega_1, \pi_{\theta^\star(\omega_1)})\| \\
&\quad + \|\nabla^2_{\omega,1} J_{\omega_2}\| \cdot \|A(\omega_2)^{-1}\| \cdot \|\nabla_\theta f(\omega_1, \pi_{\theta^\star(\omega_1)}) - \nabla_\theta f(\omega_2, \pi_{\theta^\star(\omega_2)})\| \\
&\le L_V (\|\omega_1 - \omega_2\| + 2\|\theta^\star(\omega_1) - \theta^\star(\omega_2))\|) \cdot \frac{1}{\underline{\lambda}} \cdot L_f \\
&\quad + L_V \cdot \|A(\omega_1)^{-1}\| \|A(\omega_2) - A(\omega_1)\| \|A(\omega_2)^{-1}\| \cdot L_f \\
&\quad + L_V \cdot \frac{1}{\underline{\lambda}} \cdot L_f (\|\omega_1 - \omega_2\| + \|\theta^\star(\omega_1) - \theta^\star(\omega_2)\|) \\
&\le \frac{L_f L_V}{\underline{\lambda}} \cdot \frac{\underline{\lambda} + 2L_V}{\underline{\lambda}} \|\omega_1 - \omega_2\| \\
&\quad + \frac{2L_f L_V}{\underline{\lambda}^2} \cdot L_V \Big(\|\omega_1 - \omega_2\| + 2\|\theta^\star(\omega_1) - \theta^\star(\omega_2)\| \Big) \\
&\quad + \frac{L_f L_V}{\underline{\lambda}} \cdot \frac{\underline{\lambda} + L_V}{\underline{\lambda}} \|\omega_1 - \omega_2\| \\
&\le \Big(\frac{L_f L_V (2\underline{\lambda} + 3L_V)}{\underline{\lambda}^2} + \frac{2L_f L_V^2 (\underline{\lambda} + 2L_V)}{\underline{\lambda}^3} \Big) \|\omega_1 - \omega_2\|, \quad (63)
\end{aligned}$$

where the second inequality is due to $A(\omega_1)^{-1} - A(\omega_2)^{-1} = A(\omega_1)^{-1}(A(\omega_2) - A(\omega_1))A(\omega_2)^{-1}$, and the final inequality applies Lemma 7.

Combining (61)-(63) yields the claimed result.

■

H.9 Proof of Lemma 9

By definition,

$$u^\star(\omega_1, \theta_1) = \mathrm{P}_{\underline{\lambda}}\Big(\nabla^2_{1,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1}) + \nabla^2_{2,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1})\Big)^{-1} \nabla_\theta f(\omega_1, \pi_{\theta_1}),$$
$$u^\star(\omega_2, \theta_2) = \mathrm{P}_{\underline{\lambda}}\Big(\nabla^2_{1,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2}) + \nabla^2_{2,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2})\Big)^{-1} \nabla_\theta f(\omega_2, \pi_{\theta_2}).$$

For simplicity of notation, we denote

$$A_i = \mathrm{P}_{\underline{\lambda}}(\nabla^2_{1,1} J_{\omega_i}(\pi_{\theta_i}, \pi_{\theta_i}) + \nabla^2_{2,1} J_{\omega_i}(\pi_{\theta_i}, \pi_{\theta_i})), \quad b_i = \nabla_\theta f(\omega_i, \pi_{\theta_i}), \quad \forall i \in \{1, 2\}.$$

Since $A_i \preceq -\underline{\lambda} I$, we have

$$\|A_i^{-1}\| \leq \frac{1}{\underline{\lambda}}.$$

We can bound the difference between $u^\star(\omega_1, \theta_1)$ and $u^\star(\omega_2, \theta_2)$ as follows

$$\begin{aligned}
&\|u^\star(\omega_1, \theta_1) - u^\star(\omega_2, \theta_2)\| \\
&= \|A_1^{-1} b_1 - A_2^{-1} b_2\| \\
&= \|A_1^{-1}(b_1 - b_2) + (A_1^{-1} - A_2^{-1}) b_2\| \\
&= \|A_1^{-1}(b_1 - b_2) + A_1^{-1}(A_2 - A_1) A_2^{-1} b_2\| \\
&\leq \|A_1^{-1}\| \|b_1 - b_2\| + \|A_1^{-1}\| \|A_2 - A_1\| \|A_2^{-1}\| \|b_2\| \\
&\leq \frac{1}{\underline{\lambda}} \|b_1 - b_2\| + \frac{L_f}{\underline{\lambda}^2} \|A_1 - A_2\| \\
&\leq \frac{1}{\underline{\lambda}} \|\nabla_\theta f(\omega_1, \pi_{\theta_1}) - \nabla_\theta f(\omega_2, \pi_{\theta_2})\| + \frac{L_f}{\underline{\lambda}^2} \|\nabla^2_{1,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1}) - \nabla^2_{1,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2})\| \\
&\qquad + \frac{L_f}{\underline{\lambda}^2} \|\nabla^2_{2,1} J_{\omega_1}(\pi_{\theta_1}, \pi_{\theta_1}) - \nabla^2_{2,1} J_{\omega_2}(\pi_{\theta_2}, \pi_{\theta_2})\| \\
&\leq \Big(\frac{L_f}{\underline{\lambda}} + \frac{2 L_f L_V}{\underline{\lambda}^2}\Big)\Big(\|\omega_1 - \omega_2\| + \|\theta_1 - \theta_2\|\Big),
\end{aligned} \tag{64}$$

where the third inequality follows from the fact that $\mathrm{P}_{\underline{\lambda}}$ is non-expansive.

By the Lipschitz continuity of $\theta^\star$ shown in Lemma 7,

$$\begin{aligned}
\|u^\star(\omega_1, \theta^\star(\omega_1)) - u^\star(\omega_2, \theta^\star(\omega_2))\| &\leq \Big(\frac{L_f}{\underline{\lambda}} + \frac{2 L_f L_V}{\underline{\lambda}^2}\Big)\Big(\|\omega_1 - \omega_2\| + \|\theta^\star(\omega_1) - \theta^\star(\omega_2)\|\Big) \\
&\leq \Big(\frac{L_f}{\underline{\lambda}} + \frac{2 L_f L_V}{\underline{\lambda}^2}\Big)\Big(\|\omega_1 - \omega_2\| + \frac{L_V}{\underline{\lambda}} \|\omega_1 - \omega_2\|\Big) \\
&\leq \Big(\frac{L_f}{\underline{\lambda}} + \frac{3 L_f L_V}{\underline{\lambda}^2} + \frac{2 L_f L_V^2}{\underline{\lambda}^3}\Big) \|\omega_1 - \omega_2\|.
\end{aligned}$$

We now prove the Lipschitz continuity of $\nabla u^\star$. In the rest of the subsection, we use the notations $\zeta \triangleq (x, \theta)$, $A(\zeta) \triangleq \nabla^2_{1,1} J_\omega(\pi_\theta, \pi_\theta) + \nabla^2_{2,1} J_\omega(\pi_\theta, \pi_\theta)$, $b(\zeta) \triangleq \nabla_\theta f(\omega, \pi_\theta)$, and for (ω, θ) satisfying (37),

$$u^\star(\zeta) \triangleq u^\star(x, \theta) = A(\zeta)^{-1} b(\zeta).$$

The Jacobian can be expressed as follows

$$\nabla u^\star(\zeta) = \nabla A(\zeta)^{-1} b(\zeta) + A(\zeta)^{-1} \nabla b(\zeta)$$

$$= -A(\zeta)^{-1}\nabla A(\zeta)A(\zeta)^{-1}b(\zeta) + A(\zeta)^{-1}\nabla b(\zeta).$$

Then,

$$\begin{aligned}\|\nabla u^\star(\zeta_1) - \nabla u^\star(\zeta_2)\| \le{}& \|A(\zeta_1)^{-1}\nabla A(\zeta_1)A(\zeta_1)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_2)^{-1}b(\zeta_2)\| \\ &+ \|A(\zeta_1)^{-1}\nabla b(\zeta_1) - A(\zeta_2)^{-1}\nabla b(\zeta_2)\|. \end{aligned} \tag{65}$$

To bound the first term of (65),

$$\begin{aligned}&\|A(\zeta_1)^{-1}\nabla A(\zeta_1)A(\zeta_1)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_2)^{-1}b(\zeta_2)\| \\ &\le \|A(\zeta_1)^{-1}\nabla A(\zeta_1)A(\zeta_1)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_1)A(\zeta_1)^{-1}b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\nabla A(\zeta_1)A(\zeta_1)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_1)^{-1}b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_1)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_2)^{-1}b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_2)^{-1}b(\zeta_1) - A(\zeta_2)^{-1}\nabla A(\zeta_2)A(\zeta_2)^{-1}b(\zeta_2)\| \\ &\le \|A(\zeta_1)^{-1} - A(\zeta_2)^{-1}\|\|\nabla A(\zeta_1)\|\|A(\zeta_1)^{-1}\|\|b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\|\|\nabla A(\zeta_1) - \nabla A(\zeta_2)\|\|A(\zeta_1)^{-1}\|\|b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\|\|\nabla A(\zeta_2)\|\|A(\zeta_1)^{-1} - A(\zeta_2)^{-1}\|\|b(\zeta_1)\| \\ &\quad + \|A(\zeta_2)^{-1}\|\|\nabla A(\zeta_2)\|\|A(\zeta_2)^{-1}\|\|b(\zeta_1) - b(\zeta_2)\| \\ &\le 2\|A(\zeta_1)^{-1}\|\|A(\zeta_1) - A(\zeta_2)\|\|A(\zeta_2)^{-1}\| \cdot L_V \cdot \frac{1}{\underline{\lambda}} \cdot L_f \\ &\quad + \frac{1}{\underline{\lambda}} \cdot L_V\|\zeta_1 - \zeta_2\| \cdot \frac{1}{\underline{\lambda}} \cdot L_f \\ &\quad + \frac{1}{\underline{\lambda}} \cdot L_V \cdot \frac{1}{\underline{\lambda}} \cdot L_f\|\zeta_1 - \zeta_2\| \\ &\le \Big(\frac{4L_fL_V^2}{\underline{\lambda}^3} + \frac{2L_fL_V}{\underline{\lambda}^2}\Big)\|\zeta_1 - \zeta_2\|, \end{aligned} \tag{66}$$

where in the third inequality, we bound $\|A(\zeta_1) - A(\zeta_2)\|$ in a manner similar to that in (64).

Similarly, for the second term of (65),

$$\begin{aligned}&\|A(\zeta_1)^{-1}\nabla b(\zeta_1) - A(\zeta_2)^{-1}\nabla b(\zeta_2)\| \\ &\le \|A(\zeta_1)^{-1}\nabla b(\zeta_1) - A(\zeta_1)^{-1}\nabla b(\zeta_2)\| \\ &\quad + \|A(\zeta_1)^{-1}\nabla b(\zeta_2) - A(\zeta_2)^{-1}\nabla b(\zeta_2)\| \\ &\le \|A(\zeta_1)^{-1}\|\|\nabla b(\zeta_1) - \nabla b(\zeta_2)\| + \|A(\zeta_1)^{-1} - A(\zeta_2)^{-1}\|\|\nabla b(\zeta_2)\| \\ &\le \frac{1}{\underline{\lambda}} \cdot L_f\|\zeta_1 - \zeta_2\| + \|A(\zeta_1)^{-1}\|\|A(\zeta_1) - A(\zeta_2)\|\|A(\zeta_2)^{-1}\|\|\nabla b(\zeta_2)\| \\ &\le \Big(\frac{L_f}{\underline{\lambda}} + \frac{2L_fL_V}{\underline{\lambda}^2}\Big)\|\zeta_1 - \zeta_2\|. \end{aligned} \tag{67}$$

Combining (65)-(67) yields the claimed result.

■

H.10 Proof of Lemma 10

We use the following shorthand notations in this section

$$h_k = u_k - u^\star(\omega_k, \theta^\star(\omega_k)) + \eta_k\Big(\bar{H}(\omega_k, \theta_k, u_k, \varepsilon_k) - \nabla_\theta f(\omega_k, \pi_{\theta_k})\Big),$$
$$U(\omega) = u^\star(\omega, \theta^\star(\omega)).$$

The term ④ in (39) can be written as

$$④ = 2\Big\langle h_k, U(\omega_k) - U(\omega_{k+1})\Big\rangle.$$

From Lemma 9 and Assumption 1, we know that $U(\omega)$ is Lipschitz in ω for all $\tau \geq 0$. By the mean value theorem, there exists a scalar $m \in [0,1]$ such that $\omega_{k+1}^m = m\omega_k + (1-m)\omega_{k+1}$ satisfies

$$\begin{aligned}
U(\omega_k) - U(\omega_{k+1}) &= \nabla U(\omega_{k+1}^m)^\top (\omega_k - \omega_{k+1}) \\
&= \alpha_k \nabla U(\omega_{k+1}^m)^\top \hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) \\
&= \alpha_k \nabla U(\omega_{k+1}^m)^\top \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \\
&\quad + \alpha_k \nabla U(\omega_{k+1}^m)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big).
\end{aligned}$$

Then,

$$\begin{aligned}
④ &= 2\alpha_k \Big\langle h_k, \nabla U(\omega_{k+1}^m)^\top \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big\rangle \\
&\quad + 2\alpha_k \Big\langle h_k, \nabla U(\omega_{k+1}^m)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big) \Big\rangle. \quad (68)
\end{aligned}$$

For the first term of (68),

$$\begin{aligned}
&\alpha_k \Big\langle h_k, \nabla U(\omega_{k+1}^m)^\top \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big\rangle \\
&\leq L_u \alpha_k \|h_k\| \|\bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\| \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \|h_k\|^2 + \frac{4L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\|^2 \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \|h_k\|^2 + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) - \bar{D}(\omega_k, \theta^\star(\omega_k), u^\star(\omega_k, \theta^\star(\omega_k)), \varepsilon_k)\|^2 \\
&\quad + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k) - \bar{D}(\omega_k, \theta^\star(\omega_k), u^\star(\omega_k, \theta^\star(\omega_k)), \varepsilon_k)\|^2 + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k)\|^2 \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \|h_k\|^2 + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k)\|^2 + \frac{24L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \Big(\|\theta_k - \theta^\star(\omega_k)\|^2 + \|u_k - u^\star(\omega_k, \theta^\star(\omega_k))\|^2 \Big) \\
&\quad + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k) - \bar{D}(\omega_k, \theta^\star(\omega_k), u^\star(\omega_k, \theta^\star(\omega_k)), \varepsilon_k)\|^2 \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \|h_k\|^2 + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k)\|^2 + \frac{24L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \Big(\mathrm{res}_k^\theta + \mathrm{res}_k^u \Big) \\
&\quad + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \cdot \frac{L_V^2 \varepsilon_k^4}{9} \|u^\star(\omega_k, \theta^\star(\omega_k))\|^6 \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \|h_k\|^2 + \frac{12L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \|\nabla F(\omega_k)\|^2 + \frac{24L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k} \Big(\mathrm{res}_k^\theta + \mathrm{res}_k^u \Big) + \frac{4B_u^6 L_V^2 L_u^2 \alpha_k^2 \varepsilon_k^4}{3\underline{\lambda}\eta_k}, \quad (69)
\end{aligned}$$

where the first inequality is due to $\|\nabla U(\omega)\| \leq L_u$, a consequence of the Lipschitz continuity of $u^\star$ established in Lemma 9, the fourth inequality follows from the Lipschitz continuity of $\bar{D}$ from Lemma 4, and the fifth inequality is due to Lemma 5.

For the second term of (68), we take the expectation

$$\begin{aligned}
&\alpha_k \mathbb{E}\Big[\Big\langle h_k, \nabla U(\omega_{k+1}^m)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big) \Big\rangle \Big] \\
&= \alpha_k \mathbb{E}\Big[\Big\langle h_k, \Big(\nabla U(\omega_{k+1}^m) - \nabla U(\omega_k) \Big)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big) \Big\rangle \Big] \\
&\quad + \alpha_k \mathbb{E}\Big[\Big\langle h_k, \nabla U(\omega_k)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big) \Big\rangle \Big] \\
&= \alpha_k \mathbb{E}\Big[\Big\langle h_k, \Big(\nabla U(\omega_{k+1}^m) - \nabla U(\omega_k) \Big)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \Big) \Big\rangle \Big] \\
&\leq \frac{2B_D \alpha_k}{\varepsilon_k} \mathbb{E}[\|h_k\| \|\nabla U(\omega_{k+1}^m) - \nabla U(\omega_k)\|] \\
&\leq \frac{2B_D L_u \alpha_k}{\varepsilon_k} \mathbb{E}[\|h_k\| \|\omega_{k+1}^m - \omega_k\|]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2B_D L_u \alpha_k}{\varepsilon_k} \mathbb{E}[\|h_k\| \|\omega_{k+1} - \omega_k\|] \\
&\leq \frac{2B_D L_u \alpha_k}{\varepsilon_k} \mathbb{E}[\|h_k\|] \cdot \frac{B_D \alpha_k}{\varepsilon_k} \\
&\leq \frac{\underline{\lambda}\eta_k}{16} \mathbb{E}[\|h_k\|^2] + \frac{16 B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda} \varepsilon_k^4 \eta_k},
\end{aligned} \tag{70}
$$

where the first inequality is due to the bound on the norm of $\hat{D}$ by B_D/ε_k and that on $\bar{D}$ by B_D (which is further upper bounded by B_D/ε_k for $\varepsilon_k \leq 1$), the final inequality is due to Young's inequality, and the second equation is due to

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\langle h_k, \nabla U(\omega_k)^\top \Big(\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k)\Big)\Big\rangle\Big] \\
&= \mathbb{E}\Big[\Big\langle h_k, \nabla U(\omega_k)^\top \mathbb{E}\Big[\hat{D}(\omega_k, \theta_k, \varepsilon_k, s_k^-, a_k^-, s_k^+, a_k^+, \xi_k) - \bar{D}(\omega_k, \theta_k, u_k, \varepsilon_k) \mid \mathcal{F}_{k-1}\Big]\Big\rangle\Big] \\
&= 0.
\end{aligned} \tag{71}
$$

Combining (68), (69), and (70),

$$
\begin{aligned}
\mathbb{E}[④] &\leq \mathbb{E}\left[\frac{\underline{\lambda}\eta_k}{8}\|h_k\|^2 + \frac{24 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k}\|\nabla F(\omega_k)\|^2 + \frac{48 L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k}\Big(\mathrm{res}_k^\theta + \mathrm{res}_k^u\Big)\right] + \frac{8 B_u^6 L_V^2 L_u^2 \alpha_k^2 \varepsilon_k^4}{3\underline{\lambda}\eta_k} \\
&\quad + \frac{\underline{\lambda}\eta_k}{8}\mathbb{E}[\|h_k\|^2] + \frac{32 B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k} \\
&\leq \frac{\underline{\lambda}\eta_k}{4}\mathbb{E}[\|h_k\|^2] + \frac{24 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\nabla F(\omega_k)\|^2] + \frac{48 L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\|\mathrm{res}_k^\theta] + \frac{48 L_D^2 L_u^2 \alpha_k^2}{\underline{\lambda}\eta_k}\mathbb{E}[\mathrm{res}_k^u] \\
&\quad + \frac{8 B_u^6 L_V^2 L_u^2 \alpha_k^2 \varepsilon_k^4}{3\underline{\lambda}\eta_k} + \frac{32 B_D^4 L_u^2 \alpha_k^4}{\underline{\lambda}\varepsilon_k^4 \eta_k}.
\end{aligned}
$$

■