Achieving Alignment Through Adaptive Play: Helping Optimize Objectives Without Observing Them

Jason T. Isa[1] **Samuel A. Burden**[1] **Lillian J. Ratliff**[1]

[1]Electrical and Computer Engineering, University of Washington, Seattle, Washington, USA

Abstract

We study problems where two agents seek to minimize an objective that is known only to one of the agents. This setting arises in human-machine and AI-AI interactions where one agent's objective is private knowledge inaccessible to the other, referred to as the "helper". We propose a game-theoretic learning algorithm that provably converges to optimal policies through repeated interaction *without solving an inverse problem* to recover the unknown objective. Importantly, the helper agent has no access to the cost function, its values, or its gradients, and operates under action-only feedback. Despite the information limitation, we establish convergence guarantees under Polyak-Łojasiewicz and Lipschitz-gradient assumptions. We validate the approach through two sets of experiments: human-machine interaction and a cart-pole environment with a reinforcement learning agent. In each of these experiments, the helper uses our proposed algorithm. Across scalar and multidimensional action spaces, we demonstrate consistent convergence under action-only feedback.

1 INTRODUCTION

Modern AI systems increasingly operate in interactive *multi-agent* environments rather than in isolation. In many such settings, one agent seeks to assist, coordinate with, or adapt to another agent whose objective is hidden or unobserved. This arises prominently in human-machine interaction, where machines must assist humans without access to internal objectives, but also in AI-AI interaction, such as modular AI systems composed of independently trained components and multi-agent learning settings with heterogeneous objectives. In all of these cases, implicit coordination is needed to achieve alignment.

Machines—including robots and learning algorithms—are increasingly moving from the confines of labs and factories to interacting with humans in daily life (Cannan and Hu, 2011; Gorecky et al., 2014; Hoc, 2000; Kun, 2018). Adaptive machines offer an exciting potential to assist humans in everyday work and activities such as tele- or co-robots (Nikolaidis et al., 2017), interfaces between computers and the brain or body (De Santis, 2021; Perdikis and Millan, 2020), and devices such as exoskeletons or prosthetics (Felt et al., 2015; Slade et al., 2022; Zhang et al., 2017). Yet designing learning algorithms that can safely interact with humans and constantly adapt to a dynamic environment remains an open problem in robotics, neuroengineering, and machine learning (Nikolaidis et al., 2017; Perdikis and Millan, 2020; Recht, 2019). Recent work has highlighted the challenges of human–agent coordination, preference learning, and synergy in sequential decision making (Candon et al., 2023; Chen et al., 2025; Liu et al., 2024; Shoresh and Loewenstein, 2025), emphasizing the need for principled learning algorithms that adapt effectively in interactive settings. Importantly, many of these challenges also arise in AI–AI interaction, where independently trained agents or modules must coordinate without shared objectives or reward functions, as in modular AI systems, decentralized multi-agent learning, and foundation-model-based agents interacting with task-specific controllers. Recent work studies coordination and alignment in such asymmetric-information AI–AI settings, including interaction-based adaptation, hidden or nonstationary objectives, and learning to cooperate without shared rewards (Chaudhary et al., 2025; Frankel et al., 2025; Kang et al., 2026; Li et al., 2025; Liu et al., 2025; Smith et al., 2025).

A fundamental asymmetry arises in many multi-agent interactions: one agent directly optimizes its own objective using rich internal feedback, while a helper agent observes only actions and must adapt without access to the objective itself. This asymmetry is especially salient in human-machine collaboration: humans receive rich feedback about comfort, effort, or performance and can adjust accordingly through

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2336–2377.

various mechanisms—e.g., rational best responses, gradient-like adjustments, or trial-and-error exploration. On the other hand, the machine observes only the human's actions, not the underlying preferences or the costs driving those actions. Similar asymmetries also arise in AI-AI systems when independently trained modules interact without sharing reward functions, gradients, or internal loss signals.

We study repeated interaction between two agents with asymmetric information, where Agent 1 seeks to minimize a private cost function $f(x, u)$ which Agent 2 neither knows nor can directly observe or evaluate. Nonetheless Agent 2 (aka *the helper*) has the role of assisting Agent 1 using only feedback from Agent 1's actions. This interaction is inherently strategic and sequential, making a game-theoretic formulation natural (Neumann and Morgenstern, 1944). Prior work has applied such models to human-machine systems (Cao et al., 2020; Chasnov et al., 2025; Crandall et al., 2018; Li et al., 2016, 2019; March, 2021; Nikolaidis et al., 2017). Our approach defines an asymmetric-information, *reverse Stackelberg game* in which Agent 2 (leader) commits to a policy and Agent 1 (follower) repeatedly best-responds or adapts to this policy through interaction. Despite this information constraint, we show that a simple iterative subspace alignment scheme induced by Agent 2's fixed policies provably converges to an approximate minimizer of f under benign smoothness and regularity conditions.

Traditional online learning approaches assume either bandit feedback (e.g., evaluative signals like ratings or clicks) or full information (i.e., direct cost observations). This direction complements recent research on preference-based reinforcement learning (Hussonnois et al., 2025; Kou et al., 2025; Ma et al., 2023), which relies on direct feedback on preferences. However, in many human-machine interactions, even bandit feedback is unavailable as humans rarely provide explicit numerical evaluations of every interaction, and implicit signals may be absent, delayed, or unreliable. More generally, in multi-agent systems, reward sharing may be infeasible due to architectural separation or nonstationarity. Our framework addresses the following question in this action-only feedback regime: *can a helper agent guide a joint multi-agent system toward the optimizer's optimum using only observations of the optimizing agent's actions?* We show that the answer is yes by exploiting the agent's ability to interact with and optimize on the constrained subspace imposed by the helper agent. Whether the agent best-responds, follows gradients, or explores via zeroth-order methods, our algorithm provably converges to a minimum of the hidden objective *without ever observing the cost*. Importantly, this result suggests a general design principle for collaborative AI and multi-agent systems: *rather than trying to infer, model, or learn another agent's objective, systems can leverage the structure of adaptation itself to achieve alignment.*

Contributions.

- **Algorithm: action-only alignment via affine subspace interaction.** We introduce an action-only learning algorithm in which a cost-blind helper commits to affine response policies and updates solely from observed actions, without access to cost values or gradients. The induced subspace geometry yields simple outer-loop updates with provable convergence and exposes tunable algorithm parameters for trading off convergence rate, steady-state error, number of iterations, and the amount of interaction required per iteration.
- **Theory: provable convergence with noisy or approximate inner updates.** We provide formal convergence guarantees for the proposed method even when the optimizing agent adapts imperfectly or stochastically, showing that the interaction protocol remains stable and convergent up to a predictable steady-state error.
- **Empirical validation in interactive settings.** Through human-subject experiments and simulations with optimizer agents using deterministic gain optimization and gain-space Proximal Policy Optimization (PPO), we demonstrate consistent improvement of the hidden objective across nonlinear and interactive environments.

To our knowledge, no existing methods directly address this setting, in which the helper has no *a priori* knowledge of the optimizer's objective, does not infer it, and observes only actions. This stands in contrast to gradient-based, bandit, and preference-learning approaches that require cost values, evaluative feedback, or preference comparisons.

2 PRELIMINARIES

Let $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ and define the joint action

$$z := (x, u) \in \mathbb{R}^d, \quad d := d_1 + d_2,$$

and by a slight abuse of notation $f(z) := f(x, u)$. Define the global optimum and corresponding cost by

$$f^* := \min_{z \in \mathbb{R}^d} f(z), \qquad z^* \in \arg\min_{z \in \mathbb{R}^d} f(z).$$

We impose the following regularity assumptions to characterize the problem class and the geometric properties of the induced affine subspaces.

Assumption 1. (Regularity). The function f satisfies:

a. f is β-smooth, i.e., $f(z') \le f(z) + \nabla f(z)^\top (z' - z) + \frac{\beta}{2}\|z' - z\|^2 \ \forall z, z'$.
b. The Polyak-Łojasiewicz (PŁ) inequality holds, i.e., $\|\nabla f(z)\|^2 \ge 2\mu_{\mathrm{PL}}\big(f(z) - f^*\big) \ \forall z$.

Let $\mathcal{L} = \{L_1, \dots, L_N\}$ with $L_i \in \mathbb{R}^{d_2 \times d_1}$ be the set of policy parameters for Agent 2 (the *helper*), and define

$$V_i := \begin{bmatrix} I_{d_1} \\ L_i \end{bmatrix} \in \mathbb{R}^{d \times d_1}, \qquad \mathcal{S}_i := \mathrm{range}(V_i) \subset \mathbb{R}^d.$$

Let $P_i := V_i (V_i^\top V_i)^{-1} V_i^\top$ denote the orthogonal projection onto $\mathcal{S}_i$.

Assumption 2. (Alignment of subspaces). Assume there exists $\lambda \in (0, 1]$ such that for all $g \in \mathbb{R}^d$,

$$\max_{i=1,\dots,N} \|P_i g\|^2 \geq \lambda \|g\|^2.$$

Moreover, set $V_i^\top V_i = I_{d_1} + L_i^\top L_i$ and define $\kappa := \max_{i=1,\dots,N} \lambda_{\max}(V_i^\top V_i) < \infty$.

This alignment condition ensures that the collection of subspaces $\{\mathcal{S}_i\}$ covers the ambient space: every direction has a nontrivial projection onto at least one subspace. Lemma 6 and Corollary 6 provide a constructive procedure for designing the set $\mathcal{L}$ to satisfy Assumption 2.

2.1 ALGORITHM

Algorithm 1 Two-timescale affine-response method (Uniform Sampling)

Initialize $(x_0, u_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.
for $k = 1, \dots, K$ **do**
  Sample L_k uniformly from $\mathcal{L}$ and set $L_k = L_{i_k}$.
  Define $\pi_k(x) = u_{k-1} + L_k(x - x_{k-1})$.
  Set $x_{k,0} \leftarrow x_{k-1}$ and $u_{k,0} \leftarrow u_{k-1}$.
  for $t = 0, \dots, T-1$ **do**
    $x_{k,t+1} \leftarrow \mathcal{A}(x_{k,t}; L_k, u_{k-1}, x_{k-1})$
    $u_{k,t+1} \leftarrow \pi_k(x_{k,t+1})$ ▷ policy-implied update
  end for
  $(x_k, u_k) \leftarrow (x_{k,T}, u_{k,T})$.
end for

Algorithm 2 Two-timescale affine-response method (Round-Robin Sampling with Averaging)

Initialize $(x_0, u_0) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$.
for $k = 1, \dots, K$ **do**
  for $i = 1, \dots, N$ **do** ▷ cycle through all $L_i \in \mathcal{L}$
    Set $L_{i_k} \leftarrow L_i$.
    Define $\pi_k^{(i)}(x) = u_{k-1} + L_{i_k}(x - x_{k-1})$.
    Set $x_{k,0}^{(i)} \leftarrow x_{k-1}$ and $u_{k,0}^{(i)} \leftarrow u_{k-1}$.
    for $t = 0, \dots, T-1$ **do**
      $x_{k,t+1}^{(i)} \leftarrow \mathcal{A}(x_{k,t}^{(i)}; L_{i_k}, u_{k-1}, x_{k-1})$.
      $u_{k,t+1}^{(i)} \leftarrow \pi_k^{(i)}(x_{k,t+1}^{(i)})$ ▷ policy-implied update
    end for
    $(x_k^{(i)}, u_k^{(i)}) \leftarrow (x_{k,T}^{(i)}, u_{k,T}^{(i)})$.
  end for
  $(x_k, u_k) \leftarrow \frac{1}{N} \sum_{i=1}^N (x_k^{(i)}, u_k^{(i)})$ ▷ average outputs
end for

Algorithms 1 and 2 summarize two variants of the proposed action-only feedback interaction algorithm: uniform sampling and round-robin sampling with averaging, respectively. In the theoretical analysis that follows, we focus on the uniform sampling variant.

At outer-loop epoch k, Agent 2 (the *helper*) samples $i_k \in \{1, \dots, N\}$ uniformly and commits to the fixed affine policy

$$\pi_k(x) := u_{k-1} + L_{i_k}(x - x_{k-1}),$$

which induces the composite objective for the inner loop,

$$\varphi_k(x) := f\big(x, \pi_k(x)\big).$$

Starting from $x_{k,0} := x_{k-1}$, Agent 1 interacts with this fixed policy and performs T_k steps of an inner-loop update $\mathcal{A}$, which may be stochastic, producing

$$x_k := x_{k,T_k}, \qquad u_k := \pi_k(x_k), \qquad z_k := (x_k, u_k).$$

In summary, the algorithm alternates between (i) fixing an affine policy for the helper, which induces a subspace-restricted objective, and (ii) allowing Agent 1 to approximately optimize this objective over a finite inner loop, yielding a joint update that lies on the induced affine subspace.

Stochastic Framework. We formalize the randomness in the outer-loop sampling and inner-loop updates via a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_k)_{k \geq 0}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. For each epoch $k \geq 1$:

a. The sampled index $i_k : \Omega \to \{1, \dots, N\}$ is $\mathcal{F}_k$-measurable, and $\mathcal{F}_{k-1} \subseteq \bar{\sigma}(\mathcal{F}_{k-1}, i_k) \subseteq \mathcal{F}_k$ where $\mathfrak{F}(\cdot)$ is the sigma algebra generator.
b. The inner-loop randomness in epoch k is represented by a collection $\{\omega_{k,t}\}_{t=0}^{T_k - 1}$ of random variables and a within-epoch filtration $(\mathcal{F}_{k,t})_{t=0}^{T_k}$ satisfying $\mathcal{F}_{k,0} := \bar{\sigma}(\mathcal{F}_{k-1}, i_k)$, $\mathcal{F}_{k,t+1} := \bar{\sigma}(\mathcal{F}_{k,t}, \omega_{k,t})$, $t = 0, \dots, T_k - 1$, and $\mathcal{F}_k := \mathcal{F}_{k,T_k}$.
c. The epoch iterates are adapted: $z_k : \Omega \to \mathbb{R}^d$ is $\mathcal{F}_k$-measurable, and the inner iterates $z_{k,t}$ (and $x_{k,t}$) are $\mathcal{F}_{k,t}$-measurable for all t.

This framework specifies the probabilistic setting used in the analysis.

Let $\varphi_k^* := \min_x \varphi_k(x)$ and $\Delta_k^\varphi(x) := \mathbb{E}_k[\varphi_k(x) - \varphi_k^*]$ where $\mathbb{E}_k[\cdot] \equiv \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$.

Assumption 3. (One-step inner-loop function-gap contraction). For each fixed epoch k, the inner-loop algorithm generates iterates $\{x_{k,t}\}_{t \geq 0}$ with initialization $x_{k,0} := x_{k-1}$ and there exist constants $q_k \in [0, 1)$ and $\delta_k \geq 0$ such that for every inner iteration $t \geq 0$, the bound holds:

$$\Delta_k^\varphi(x_{k,t+1}) \leq q_k \, \Delta_k^\varphi(x_{k,t}) \; + \; (1 - q_k) \, \delta_k.$$

Assumption 3 is satisfied by a broad class of optimization and learning procedures under standard regularity conditions, including smoothness and the PŁ condition. It is well known that deterministic gradient descent and proximal-gradient methods enjoy global linear convergence

for PŁ functions (Garrigós and Gower, 2023; Karimi et al., 2016). Variance-reduced methods such as SAGA and Prox-SVRG admit local linear convergence under comparable regularity conditions (Poon et al., 2018). Extensions to momentum-based methods, including stochastic momentum and Polyak's heavy-ball method, retain linear or accelerated local convergence rates under the PŁ inequality (Kassing and Weissmann, 2024; Liang et al., 2023). Importantly, adaptive methods such as RMSProp and Adam-type optimizers, which are widely used in training modern AI agents, can converge under Łojasiewicz-type conditions under additional moment assumptions (Bensaid et al., 2024).

As a representative example, consider stochastic gradient descent (SGD) with stochastic gradients $g_k(x_{k,t})$ satisfying, for all t, $\mathbb{E}[g_k(x_{k,t}) \mid \mathcal{F}_{k,t}] = \nabla\varphi_k(x_{k,t})$, $\mathbb{E}\big[\|g_k(x_{k,t}) - \nabla\varphi_k(x_{k,t})\|^2 \mid \mathcal{F}_{k,t}\big] \le \sigma^2$. Equivalently, writing $g_k(x_{k,t}) = \nabla\varphi_k(x_{k,t}) + \omega_{k,t}$, the noise $\omega_{k,t}$ is conditionally mean-zero with bounded second moment. Under these standard assumptions, SGD satisfies Assumption 3 under the PŁ condition with constants $q_k = 1 - \eta\mu_{\mathrm{PL}} \in (0,1)$ and noise floor $\delta_k = \eta\beta\kappa\sigma^2/(2\mu_{\mathrm{PL}})$ (Garrigós and Gower, 2023; Karimi et al., 2016).

3 THEORETICAL GUARANTEES

This section establishes convergence guarantees for the two-timescale affine-response method using uniform-sampling in Algorithm 1.

Intuition. Before getting into the main details, let us give the general idea behind why the proposed affine policy for Agent 2 (aka helper) suffices to ensure convergence even in the absence of information related to Agent 1's objective function, as long as Agent 1 is making progress on the induced objective $f(x, \pi_k(x))$ in each epoch.

Indeed, for each epoch k, the sampled matrix L_{i_k} induces an affine subspace $S_k = z_{k-1} + \mathcal{S}_{i_k}$ with $\mathcal{S}_{i_k} := \operatorname{range}\left(\begin{bmatrix} I_{d_1} \\ L_{i_k} \end{bmatrix}\right)$. Due to the policy constraint $u = \pi_k(x)$, all inner-loop iterates remain on S_k (Appendix Lemma 3), so the inner loop is progressively approximating the minimizer of f *restricted to this subspace*—i.e., the subspace gap $f(z_{k-1}) - f^*_{i_k}(z_{k-1})$ is progressively decreasing, where $f^*_{i_k}(z) := \min_{y \in z + \mathcal{S}_{i_k}} f(y)$. Even if the algorithm run by Agent 1 in each epoch is stochastic, this subspace constraining lemma holds, and progress is made towards the noise floor. By smoothness of f (Assumption 1.a), this gap is lower bounded by the projected gradient norm $\|P_{i_k}\nabla f(z_{k-1})\|^2$, and the alignment condition (Assumption 2) ensures that, in expectation over the sampled subspace, the lower bound $\mathbb{E}_{i_k}[\|P_{i_k}\nabla f(z)\|^2] \ge (\lambda/N)\|\nabla f(z)\|^2$ holds. The PŁ condition (Assumption 1.b) then links gradient norm to global suboptimality via $\|\nabla f(z)\|^2 \ge 2\mu_{\mathrm{PL}}(f(z) - f^*)$. Combining these ingredients yields a linear contraction of the expected objective gap across epochs; in the regime where Agent 1 runs a stochastic algorithm within each epoch, stochastic inner-loop convergence holds up to a steady-state error determined by the noise floor in Assumption 3, while in the deterministic setting the noise floor does not exist, yielding linear convergence to the global minimizer.

3.1 STOCHASTIC AGENT 1 ALGORITHM

We first consider the general stochastic setting in which the inner-loop update $\mathcal{A}$ may be noisy (e.g., SGD or noisy best response), but satisfies the function-gap contraction property in Assumption 3. The following lemma quantifies the one-epoch progress of the outer loop in expectation.

Define $f_i^*(z) := \min_{y \in z + \mathcal{S}_i} f(y)$. Set $\varphi_k^* := \min_x \varphi_k(x)$, and observe that $\varphi_k^* = f_i^*(z)$ and $f(z_k) = \varphi_k(x_k)$. Define $\Delta_k^f(z_k, z) := \mathbb{E}\left[f(z_k) - f_i^*(z) \mid \mathcal{F}_{k,0}\right]$.

Lemma 1. (One-epoch progress). Fix an epoch k, and set $z := z_{k-1}$. Under Assumptions 1, 2, and 3, the estimate holds:

$$\Delta_k^f(z_k, z) \le q_k^{T_k}\Delta_k^{\varphi}(x_{k-1}) + \delta_k\big(1 - q_k^{T_k}\big). \tag{1}$$

Building on Lemma 1, we obtain global linear convergence of the outer loop up to a noise floor.

Theorem 1. (Outer-loop linear convergence under a global PŁ condition). Suppose Assumptions 1, 2, and 3 hold. The estimate holds:

$$\mathbb{E}[f(z_k) - f^*] \le \rho_k\,\mathbb{E}[f(z_{k-1}) - f^*] + (1 - q_k^{T_k})\delta_k,$$

where $\rho_k := 1 - (1 - q_k^{T_k})\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N} \in (0,1)$. If $(1 - q_k^{T_k}) = \theta$ and $\delta_k = \delta$ are constant across k, then

$$\mathbb{E}[f(z_k) - f^*] \le \rho^k\big(f(z_0) - f^*\big) + \delta\theta/(1 - \rho).$$

The proof is given in Appendix B.2.2. The critical observation that leads to the proof of this theorem is the fact that the iterates $z_{k,t} = (x_{k,t}, u_{k,t})$ remain in the constrained subspace S_k for each epoch as we show in the Appendix Lemma 3.

Observe that $f(z_0) - f^*$ is the optimization error and $\delta_k\theta_k$ is the error due to the stochasticity in the inner problem that Agent 1 faces. The optimization error decays exponentially across epochs since $\rho_k < 1$. The noise term accumulates as follows:

$$\mathbb{E}[f(z_k) - f^*] \le \bar{\rho}^k\big(f(z_0) - f^*\big) + \textstyle\sum_{t=1}^{k}(1 - q_t^{T_t})\,\delta_t\bar{\rho}^{k-t}$$

where $\bar{\rho} := \max_{s \le k}\rho_s$. Hence a critical question is how the helper agent (Agent 2) can devise a *control strategy* to mitigate this error accumulation. The challenge is that

the optimization error and the noise terms tradeoff in the following sense: increasing T_t causes (i) $1-q_t^{T_t}$ to increase and thereby increase the noise $(1-q_t^{T_t})\delta_t$, and (ii) $\bar{\rho}$ to decrease which discounts past noise more. Hence, longer epochs solve the inner problem more accurately, but also result in hitting the noise floor sooner with potentially more noise per epoch if δ_k does not shrink with T_k. This motivates designing an adaptive algorithm that can control for δ_k in each epoch.

Before designing such an algorithm, it is helpful to see what these constants are in the context of a common inner algorithm class. Indeed, as a concrete instantiation of Theorem 1, consider Agent 1 running stochastic gradient descent.

Corollary 1. (Outer-loop convergence with SGD $\mathcal{A}$). Suppose Assumptions 1 and 2 hold, and that Agent 1 runs stochastic gradient descent with stepsize $\eta < 1/(\beta\kappa)$ and with unbiased gradients of bounded variance σ^2. Fix constants $q := 1-\eta\mu_{\mathrm{PL}}$, $\theta := 1-q^T$, and assume a fixed inner-loop length $T_k \equiv T$ for all epochs. The outer-loop iterates satisfy

$$\mathbb{E}\left[f(z_k)-f^*\right] \le \rho^k\left(f(z_0)-f^*\right) + \tfrac{N\beta}{\lambda}\cdot\tfrac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}^2} \qquad (2)$$

where $\rho := 1-\theta\tfrac{\mu_{\mathrm{PL}}}{\beta}\tfrac{\lambda}{N} \in (0,1)$.

Proposition 1. (Iteration Complexity). Assume the setting of Corollary 1. Set constants $\rho := 1-\theta c \in (0,1)$ with $c := \mu_{\mathrm{PL}}\lambda/(\beta N)$ and $\delta := \beta\kappa\eta\sigma^2/(2\mu_{\mathrm{PL}})$ so that the noise floor in (2) is given by $\epsilon_\infty := \delta\theta/(1-\rho)$. Consequently, for any target accuracy $\epsilon_\star > \epsilon_\infty$, it suffices to choose

$$k \ge \frac{1}{\log(1/\rho)}\log\left(\frac{f(z_0)-f^*}{\epsilon_\star-\epsilon_\infty}\right) \qquad (3)$$

to guarantee $\mathbb{E}\left[f(z_k)-f^*\right] \le \epsilon_\star$.

Observe that (3) reduces to $k \ge \frac{1}{c(1-q^T)}\log\left(\frac{f(z_0)-f^*}{\epsilon_\star-\epsilon_\infty}\right)$ so that as T increases the lower bound on k decreases. Moreover, we have that the total number of iterations (across all epochs) is $kT \ge (c\eta\mu_{\mathrm{PL}})^{-1}\log((f(z_0)-f^*)/(\epsilon_*-\epsilon_\infty))$.

Controlling Noise-vs-Optimization Error Tradeoff. Now let us consider how we can design algorithms for the helper agent to control for the noise at the inner level and trade that off with performance error. At outer epoch k, sample $i_k \sim \mathrm{Unif}(\{1,\dots,N\})$ and set $L_k := L_{i_k}$ as before, but now introduce a response-gain parameter $\nu_k \in (0,1]$ and define the epoch-k affine policy

$$\pi_k(x) := u_{k-1} + \nu_k L_k(x-x_{k-1}). \qquad (4)$$

Theorem 2. (Outer-loop recursion) Under Assumptions 1, 2, and 3, the outer iterates satisfy

$$\Delta_k^\star \le \big(1-\theta_k(\nu_k)\Gamma_k(\nu_k)\big)\Delta_{k-1}^\star + \theta_k(\nu_k)\,\delta_k(\nu_k), \qquad (5)$$

where $\Delta_k^\star := \mathbb{E}[f(z_k)-f^*]$, $\Gamma_k(\nu_k) := \tfrac{\mu_{\mathrm{PL}}}{\beta}\cdot\tfrac{\lambda(\nu_k)}{N}$, $\theta_k(\nu_k) := 1-q_k(\nu_k)^{T_k}$ and $\delta_k(\nu_k)$ is the one-step noise-floor parameter appearing in Assumption 3.

Although ν_k enters the contraction factor through $\Gamma_k(\nu_k)$ via the alignment constant $\lambda(\nu_k)$, it also enters the additive term through $\theta_k(\nu_k)$ and $\delta_k(\nu_k)$, since both quantities depend on the inner objective $\varphi_k(x) = f(x,\pi_k(x))$, whose geometry—e.g., smoothness/curvature and noise amplification—varies with ν_k. To see this, let us specialize to the SGD case.

Corollary 2. (Outer recursion with SGD $\mathcal{A}$). Assume that Agent 1 runs stochastic gradient descent with step size $\eta \le 1/\beta_k(\nu_k)$ and that φ_k is $\mu_k(\nu_k)$-PŁ and $\beta_k(\nu_k)$-smooth, with unbiased stochastic gradients of variance at most σ^2. Then Assumption 3 holds with $q_k(\nu_k) = 1-\eta\mu_k(\nu_k)$, $\theta_k(\nu_k) = 1-(1-\eta\mu_k(\nu_k))^{T_k}$, and $\delta_k(\nu_k) = \tfrac{\beta_k(\nu_k)\eta\sigma^2}{2\mu_k(\nu_k)}$. Hence (5) reduces to

$$\Delta_k^\star \le \big(1-\theta_k(\nu_k)\Gamma_k(\nu_k)\big)\Delta_{k-1}^\star + \theta_k(\nu_k)\cdot\tfrac{\beta_k(\nu_k)\eta\sigma^2}{2\mu_k(\nu_k)}.$$

Increasing ν_k improves scaled alignment by increasing $\lambda(\nu_k)$, which increases $\Gamma_k(\nu_k)$ and decreases the contraction factor $1-\theta_k\Gamma_k(\nu_k)$, leading to faster reduction of the optimization error. However, larger ν_k can also amplify the stochastic noise term: the restricted smoothness scales as $\beta\lambda_{\max}(I+\nu_k^2 L_k^\top L_k)$, which increases the noise floor. In standard optimization this tradeoff could be controlled through the stepsize η, but η is chosen by Agent 1 not Agent 2. This motivates either adaptive or parameter-free inner methods that estimate some of these unknown quantities. Another approach is to optimize (ν_k, T_k) to control the error due to optimization and noise. Below we illustrate this in the most simple case where constant schedules are chosen; however, we note that more complex schedules from optimization such as step-size halving and staged-based algorithms are also employable.

Corollary 3. (Rate–floor tradeoff via ν_k) Assume the setting of Theorem 2 and Corollary 2. Suppose Agent 2 chooses a constant response gain $\nu_k \equiv \nu \in (0,1]$ and a constant inner-loop length $T_k \equiv T$. Let $\lambda(\nu)$ denote the scaled alignment constant, and suppose $\|L_i\| \le \bar{L}$ for all $i \in [N]$. Define $\Gamma(\nu) := \tfrac{\mu_{\mathrm{PL}}}{\beta}\tfrac{\lambda(\nu)}{N}$ and $\Psi(\nu) := \tfrac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)}$. Then the asymptotic noise floor is

$$\epsilon_\infty(\nu) := \frac{\Psi(\nu)}{\Gamma(\nu)} = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)}\cdot\frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)}.$$

Moreover, suppose $\lambda(\nu) \ge \frac{1}{2(1+\nu^2\bar{L}^2)}\min\{1,\nu^2\lambda_L\}$, where $\lambda_L := \inf_{\|b\|=1}\max_{i\in[N]}\|L_i^\top b\|^2$, and suppose $\beta(\nu) \le \beta(1+\nu^2\bar{L}^2)$. Then

$$\epsilon_\infty(\nu) \lesssim \frac{N\beta^2\eta\sigma^2}{\mu_{\mathrm{PL}}\mu(\nu)}\cdot\frac{(1+\nu^2\bar{L}^2)^2}{\min\{1,\nu^2\lambda_L\}}.$$

In particular, if $\mu(\nu)$ is uniformly lower bounded, then this upper bound scales like ν^{-2} when $\nu^2\lambda_L \leq 1$.

Corollary 3 captures the tradeoff induced by the response gain ν. Smaller ν can reduce the restricted smoothness through the factor $1+\nu^2\bar{L}^2$, but it can also make the scaled alignment constant $\lambda(\nu)$ small, leading to slower convergence and a larger noise floor. When $\nu^2\lambda_L \gtrsim 1$, the alignment term saturates, and the remaining dependence is governed by the conditioning of the restricted inner objective through $\beta(\nu)/\mu(\nu)$. Thus, ν balances noise amplification against scaled alignment.

Corollary 4. (Optimizing epoch length and finite-time decay) Assume the setting of Theorem 2 and Corollary 2. Suppose Agent 2 chooses a constant response gain $\nu_k \equiv \nu \in (0,1]$ and a constant inner-loop length $T_k \equiv T$. Define $\Gamma(\nu) := \frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda(\nu)}{N}$, $q(\nu) := 1-\eta\mu(\nu)$, and $\theta(\nu,T) := 1-q(\nu)^T$. Let $\epsilon_\infty(\nu) := \frac{\Psi(\nu)}{\Gamma(\nu)} = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)}$ denote the corresponding asymptotic noise floor. Fix a target accuracy $\epsilon_\star > \epsilon_\infty(\nu)$ and a horizon $k \in \mathbb{N}$. Then it suffices to choose T so that

$$T \geq \frac{1}{\log q(\nu)} \log\left(1 - \frac{1}{k\Gamma(\nu)} \log\left(\frac{\Delta_0^\star}{\epsilon_\star - \epsilon_\infty(\nu)}\right)\right), \tag{6}$$

provided the logarithm is well-defined, i.e., the argument of $log(\cdot)$ lies in $(0,1)$. With this choice, the outer-loop error satisfies $\Delta_k^\star \leq \epsilon_\star$.

Corollary 4 shows how to choose the epoch length T after fixing the response gain ν. For fixed ν, larger T increases $\theta(\nu,T) = 1-q(\nu)^T$, decreases the contraction factor $\rho(\nu,T) = 1-\theta(\nu,T)\Gamma(\nu)$, and therefore reduces the number of outer epochs needed to reach a target accuracy above the noise floor. However, the floor $\epsilon_\infty(\nu) = \Psi(\nu)/\Gamma(\nu)$ is independent of T, so changing T affects only the rate of approach, not the limiting accuracy. Together, Corollaries 3 and 4 suggest a two-stage design: (i) choose ν to balance scaled alignment against noise amplification, (ii) then choose T large enough to reach the desired accuracy within the allotted number of outer epochs.

3.2 DETERMINISTIC AGENT 1 ALGORITHM

We now consider the noise-free regime, where Agent 1's update $\mathcal{A}$ is deterministic. In this case, the outer loop converges exactly to the global minimizer.

Corollary 5. (Exact linear convergence with deterministic inner loop). Under Assumptions 1 and 2, suppose the inner loop satisfies Assumption 3 with $\delta_k = 0$ for all k. Then the outer-loop iterates converge linearly:

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k\big(f(z_0) - f^*\big), \tag{7}$$

where $\rho = 1-\theta\,\frac{\mu_{\mathrm{PL}}}{\beta}\,\frac{\lambda}{N} \in (0,1)$.

This establishes exact linear convergence for any deterministic inner loop that achieves subspace contraction.

4 EXPERIMENTS

We evaluate the proposed interaction algorithms across two experimental environments: (i) a human-machine interaction task with human subjects recruited using a crowdsourcing platform and (ii) a two-agent cart-pole system in which Agent 1 adapts using either deterministic gain optimization or learning-based gain-space PPO. These environments span increasing levels of realism and complexity—from convex objectives to nonlinear dynamics and learning-based policies—and are designed to test generality, robustness under challenging dynamics, and compatibility with learning-based agents. Across all environments, the helper agent is cost-blind and updates using action-only feedback, consistent with the information constraints assumed in our theory. Full methods and implementation details are provided in Appendix D & E and code can be found at (Isa et al., 2026).

4.1 HUMAN-MACHINE INTERACTION TASK

We evaluate the uniform sampling (Algorithm 1) and round-robin (Algorithm 2) variants of the algorithm in a *dual-control* human-machine task, where both the human H and the machine M jointly influence the outcome. Experiments span four action-space conditions, $d_H \times d_M \in \{1\times1,\ 1\times2,\ 2\times1,\ 2\times2\}$, which determine the dimensionality of each agent's action space. The human optimizes a prescribed quadratic cost over joint actions, visualized on a computer display as the radius of a circle, while the machine is cost-blind and updates using action-only feedback with respect to the private objective (Chasnov et al., 2025; Isa et al., 2024). Participants were recruited via a crowdsource platform (Prolific (Palan and Schitter, 2018)) and were naïve to the task. Each participant completed a single condition and interacted with only one variant of the algorithm. In total, 248 participants were recruited; data from 240 were retained after excluding seven for data-logging errors and one for failed attention checks. Experiments were conducted using procedures approved by the University of Washington Institutional Review Board (UW IRB STUDY00013524).

Human participant experiments converge to optimum. Figure 1 shows results for the 1×1 game using the round-robin sampling variant of the algorithm, compared against simulations with a best-responding agent in place of the human. From each of eight symmetrically arranged initialization points (a), the iterates (h_k, m_k) rapidly converges to the human's prescribed minimizer (h^*, m^*) (b,c), yielding convergence of the cost to its minimum value $f^* = 0$ (d). Convergence rates and trajectories closely match those observed with best-response simulations, with the estimate

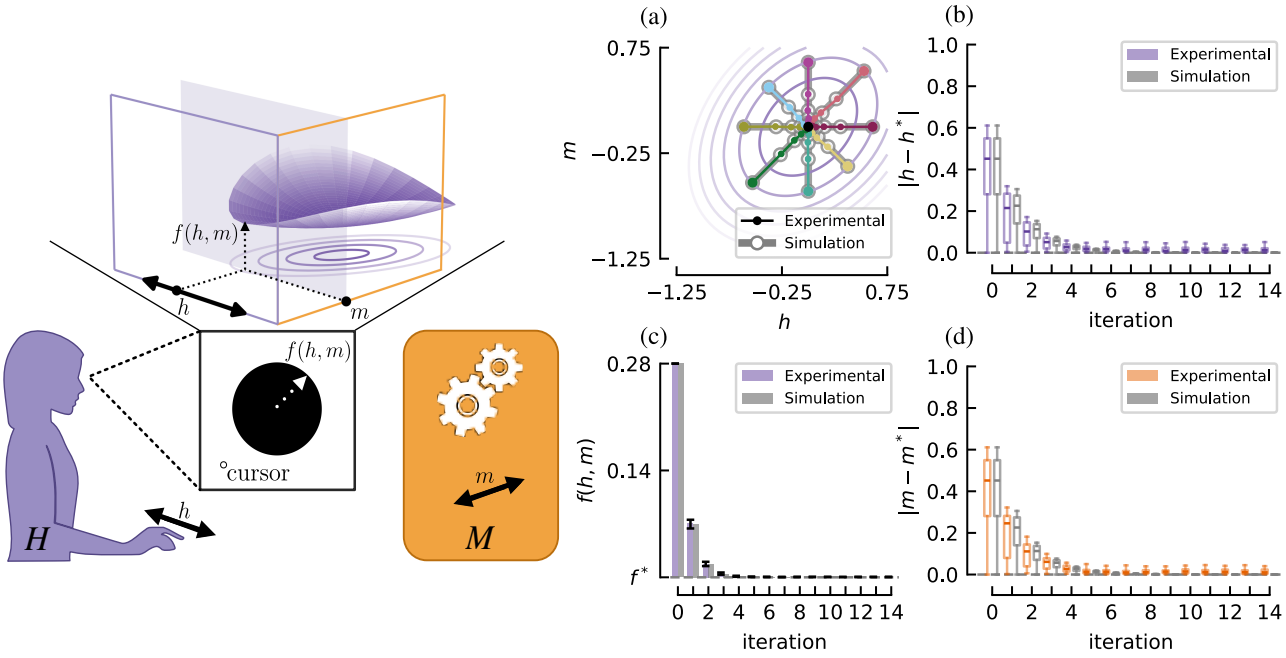

Figure 1: 1×1 Human Participant Experiment vs Simulations ($n = 80$; $n = 10$ per initialization point): Grey symbols correspond to simulation data where H played best response. (a) M's median estimate of h^* and m^* over iterations for each of 8 color-coded initializations overlaid on cost contours. Experimental data shown with solid dots, simulation data shown with open dots. (b) Distributions of ℓ_1 error of M's estimate of h^*; box-and-whiskers plot showing 5th, 25th, 50th, 75th, and 95th percentiles. (c) Cost distributions; bar plots with quartiles. (d) Distributions of ℓ_1 error of M's estimate of m^*; box-and-whiskers plot showing same as (b).

typically becoming indistinguishable from the optimum within four outer iterations. The straight-line paths to the optimum arise from the fixed choice of $\mathcal{L}$ and symmetric initializations; alternative choices of $\mathcal{L}$ yield similarly rapid convergence without straight-line trajectories.

Additional results for the 1×1, 1×2, 2×1, and 2×2 conditions under both the round-robin and uniform sampling variants of the algorithms are reported in Appendix F and exhibit consistent convergence toward the optimum. The uniform sampling method shows higher trajectory variance than the round-robin sampling method, as reflected in the cost distributions over iterations, due to updating with a single randomly selected subspace per iteration rather than averaging over all subspaces in $\mathcal{L}$. Although the uniform sampling method typically requires more outer iterations, each round-robin sampling iteration evaluates all subspaces in $\mathcal{L}$ and thus incurs higher interaction cost. As a result, when measured by the total number of trials, the uniform sampling converges faster in practice. Together, these results highlight a tradeoff between trajectory variance and interaction efficiency when choosing between sampling variants.

4.2 COUPLED CART-POLE GAME

We evaluate the uniform-sampling variant of the algorithm (Algorithm 1) in a closed-loop two-agent cart-pole task with a spring-coupled helper cart. Agent 1 controls the primary cart-pole, while Agent 2 controls a secondary cart coupled to it through spring-damper dynamics, yielding a nonlinear coupled system. Agent 1 observes the quadratic regulation objective on state and control effort and, at each outer iteration, computes a closed-loop response to the helper's current affine coupling.

We initialize Agent 1's policy from the one-player closed-loop LQR controller and consider two gain-space inner solvers. In the deterministic condition, Agent 1 computes an approximate best response by directly optimizing a six-dimensional linear feedback gain on the nonlinear rollout objective. In the learning-based condition, Agent 1 uses gain-space PPO on the full nonlinear system. In both cases, the algorithm acts in the space of policy parameters rather than by perturbing raw control inputs: the outer-loop variables correspond to feedback gains that induce closed-loop control laws. Agent 2 is cost-blind and probes Agent 1 through affine perturbations of these policy parameters. We compare against a baseline in which Agent 2 is present but inactive, with its input fixed to zero, and Agent 1 stabilizes the cart-pole using the one-player closed-loop LQR controller.

We run two additional learning-based condition (PPO) ablations to illustrate the tradeoffs in Corollaries 3 and 4. First, we vary the helper response gain ν, which controls the magnitude of Agent 2's affine response to Agent 1's policy

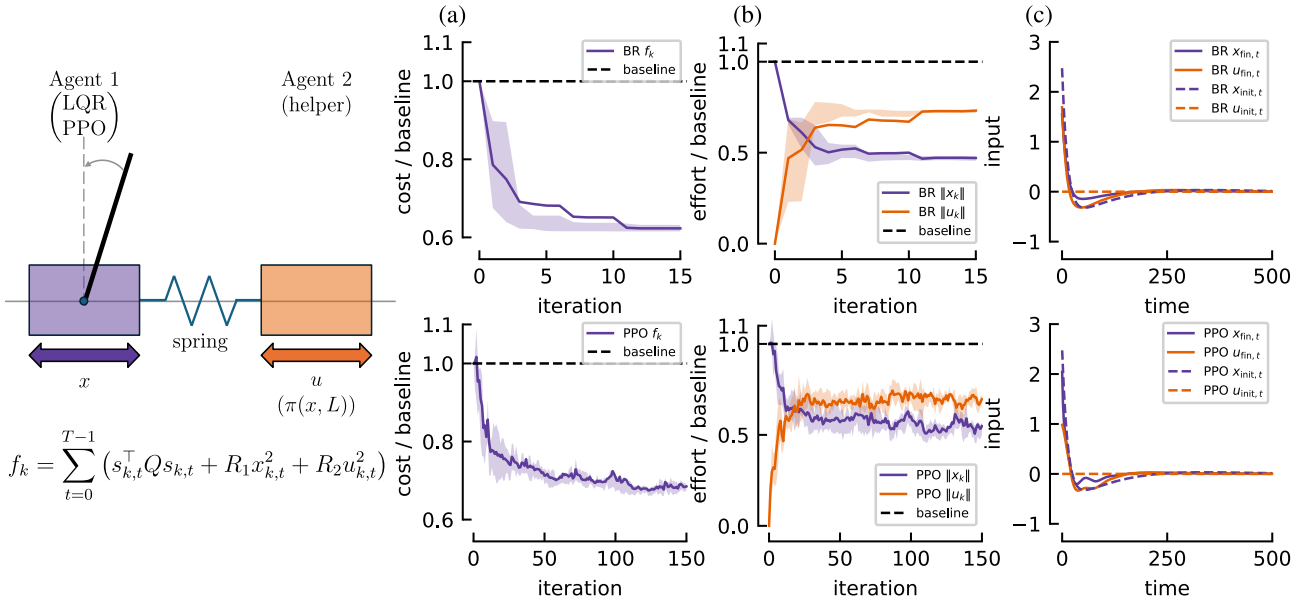

Figure 2: Closed-Loop Two-Agent Balancing Cart-Pole Game (10 seeds): Agent 1 computes best responses (BR) (top row) or PPO (bottom row), while Agent 2 updates using the proposed uniform sampling method. Black curves denote the Agent 1-only LQR baseline (helper disabled). Colored curves denote the two-agent setting, with Agent 1 initialized from the baseline policy. (a) Normalized objective f_k/f_0 versus outer iteration k. Curves show the mean; shaded regions indicate the interquartile range (25th–75th percentiles). (b) Normalized control effort $\|x_k\|_1/\|x_0\|_1$ and $\|u_k\|_1/\|x_0\|_1$ versus k. Curves and shaded regions same as (a). (c) Control inputs $x_{k,t}$ and $u_{k,t}$ at the initial (dashed) and final (solid) outer iterations.

update, isolating its effect on the rate–floor tradeoff in Corollary 3. Second, we vary the PPO inner-loop budget T_{PPO}, defined as the number of Stable-Baselines3 PPO environment steps used to approximate Agent 1's best response at each outer iteration, isolating its effect on the epoch-length tradeoff in Corollary 4. For both ablations, we report normalized cost f_k/f_0 across seeds and use fixed helper-response sequences within each seed, so that differences across curves reflect the varied design parameter rather than different samples of L_k.

The helper agent improves performance in the two-agent cart-pole game. Figure 2 evaluates the proposed method in a nonlinear, spring–damper-coupled two-agent cart-pole task. This experiment tests the algorithm beyond the convex quadratic setting covered by the theory, since the induced optimization over policy parameters is globally nonconvex. Agent 1 computes closed-loop responses using either best response (top row) or a PPO inner loop (bottom row), while the cost-blind helper updates with the proposed affine-response rule. Curves show the mean and interquartile range over $N = 10$ seeds, with costs normalized to the Agent 1-only LQR baseline at $k = 0$.

Introducing the helper reduces the cooperative objective in both conditions (Fig. 2a). The BR condition converges rapidly, dropping from 1.00 to 0.65 by iteration 10 and to 0.62 ± 0.02 by iteration 15. PPO improves more gradually over 150 outer iterations, reaching 0.86 at iteration 10, 0.71 at iteration 50, and 0.69 ± 0.03 at termination. Control-effort traces (Fig. 2b) show that Agent 2 takes on nonzero effort while Agent 1's effort decreases, indicating that the helper relieves the primary controller rather than simply increasing total actuation. Across seeds, terminal rollouts balance the pole for the full episode horizon, showing that the helper improves performance without compromising stability.

Response gain controls the rate–floor tradeoff in the coupled cart-pole game. Figure 3a evaluates the response-gain tradeoff from Corollary 3 by sweeping $\nu \in \{0.5, 0.75, 1.0\}$ with fixed PPO budget $T_{\text{PPO}} = 150$. Across 10 seeds, all curves start at the one-player baseline ($f_k/f_0 = 1$), decrease over the first 10–50 outer iterations, and then flatten. The three gains often reach similar final costs, but $\nu = 1.0$ is less stable: it has the highest mean final cost (0.82 ± 0.32), the largest seed-to-seed spread, and one failure case where $f_{150}/f_0 > 1.6$. In contrast, $\nu = 0.5$ attains the lowest mean floor (0.70 ± 0.09), with $\nu = 0.75$ slightly higher (0.72 ± 0.12). Thus, stronger helper coupling does not reliably improve early progress in this nonlinear PPO setting and can amplify inner-solve errors, while smaller ν yields more conservative and reliable adaptation, consistent with the rate–floor tradeoff in Corollary 3.

The PPO inner-loop budget affects the sample-efficiency of adaptation. Figure 3b evaluates the epoch-length tradeoff from Corollary 4 by sweeping $T_{\text{PPO}} \in \{90, 150, 210\}$

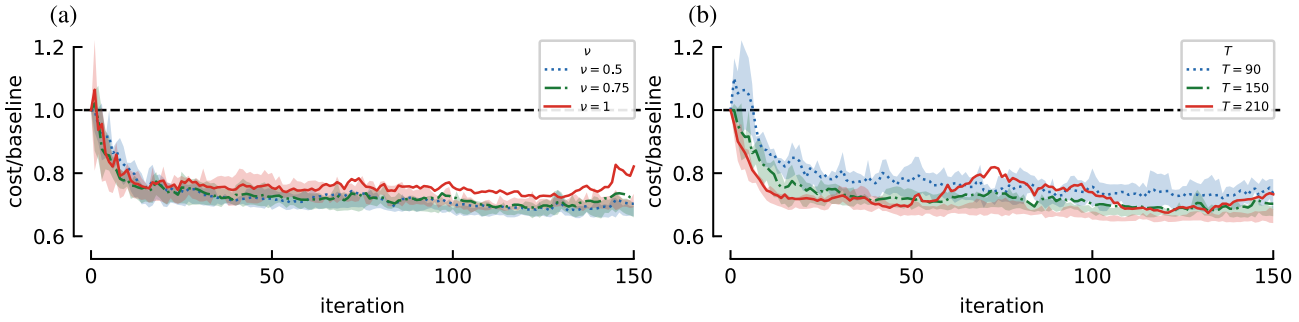

Figure 3: Cart-pole ablations for the rate–floor and epoch-length tradeoffs (10 seeds): We evaluate gain-space PPO in the coupled cart-pole game while varying the design parameters associated with Corollaries 3 and 4. (a) Response-gain ablation: normalized objective f_k/f_0 versus outer iteration k for different helper response gains ν, with the PPO inner-loop budget fixed. (b) Inner-budget ablation: normalized objective f_k/f_0 versus outer iteration k for different PPO inner-loop budgets T_{PPO}. Curves show the mean; shaded regions indicate the interquartile range (25th–75th percentiles).

with fixed $\nu = 0.5$. Plotted against outer iteration, all budgets rapidly improve from $f_k/f_0 = 1$, but they reach different empirical floors. The intermediate budget performs best, with $T_{\mathrm{PPO}} = 150$ achieving the lowest final mean cost (0.70 ± 0.09), compared with 0.74 ± 0.06 for $T_{\mathrm{PPO}} = 90$ and 0.73 ± 0.23 for $T_{\mathrm{PPO}} = 210$. The dependence on budget is not monotone: the smallest budget improves slowest early, while the largest can improve quickly but does not consistently outperform the intermediate budget. These results support the qualitative message of Corollary 4: inner-loop length should balance solve accuracy against update frequency, with $T_{\mathrm{PPO}} = 150$ giving the best empirical tradeoff in this implementation.

5 DISCUSSION

Our analysis establishes linear convergence under Polyak–Łojasiewicz regularity and smoothness, including approximate or stochastic inner-loop updates that converge to a nonzero noise floor. The theory also identifies concrete design levers: the choice and scheduling of helper-induced subspaces, the size and structure of the subspace family, step sizes and time-scale separation, and inner-loop stochasticity. These choices trade off convergence rate, robustness to noise and tracking error, and per-iteration cost. We also give constructive procedures for minimal subspace families whose size depends on rank conditions rather than ambient dimension, allowing the algorithm to be tailored for efficiency, stability, or responsiveness in high-dimensional or noisy settings.

A primary limitation of the human-subject study is that the objective was prescribed by the experimenter. While this provides a clean proof-of-concept by removing uncertainty about the objective, it does not capture the complexity of real-world objectives. Future work will evaluate the algorithm in settings where objectives are endogenous, task-dependent, potentially nonstationary, and where inner-loop optimization is learned rather than specified.

One natural application is assisted mobility, where a component of the human objective is often believed to be minimizing the metabolic cost of transport (Abram et al., 2022; Felt et al., 2015; Slade et al., 2022; Zhang et al., 2017). In this setting, the method can be compared against state-of-the-art *human-in-the-loop* optimization approaches that rely on explicit physiological measurements. Another promising domain is physical human–robot interaction (Li et al., 2016; Nikolaidis et al., 2017), where objectives are latent and vary across users, and action-only alignment may offer an alternative to preference inference. Beyond human–machine interaction, the framework also applies to AI–AI settings in which one agent seeks to align with another agent whose objective is hidden or unobserved, such as modular AI systems and multi-agent reinforcement learning, complementing reward modeling and preference learning approaches (Rastogi et al., 2022; Zhang et al., 2020).

6 CONCLUSION

We introduced a multi-agent learning framework in which a cost-blind helper aligns with another agent's hidden objective using only action feedback, without observing costs, gradients, preferences, or solving an inverse problem. Framed as adaptive play in a reverse Stackelberg game, the helper guides the joint system toward the optimizer of the unobserved objective by inducing structured subspace responses. Under PŁ regularity and smoothness assumptions, we prove linear outer-loop convergence and characterize the noise floor that arises when the optimizing agent adapts approximately or stochastically. Empirically, we validate the method in human-subject experiments and two-agent cart-pole control, demonstrating convergence in both convex static objectives and nonlinear dynamical settings.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant #2447698.

References

Sabrina J. Abram, Katherine L. Poggensee, Natalia Sánchez, Surabhi N. Simha, James M. Finley, Steven H. Collins, and J. Maxwell Donelan. General Variability Leads to Specific Adaptation Toward Optimal Movement Policies. *Current Biology*, 32(10):2222–2232.e5, 2022. doi: 10.1016/j.cub.2022.04.015.

Bilel Bensaid, Gaël Poëtte, and Rodolphe Turpault. Convergence of the iterates for momentum and rmsprop for local smooth functions: Adaptation is the key. *arXiv preprint arXiv:2407.15471*, 2024.

Kate Candon, Jesse Chen, Yoony Kim, Zoe Hsu, Nathan Tsoi, and Marynel Vázquez. Nonverbal human signals can help autonomous agents infer human preferences for their behavior. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 307–316, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

James Cannan and Huosheng Hu. Human-Machine Interaction (HMI): A Survey. *University of Essex*, 27:46–64, 2011.

Mukun Cao, Qing Hu, Melody Y. Kiang, and Hong Hong. A Portfolio Strategy Design for Human-Computer Negotiations in e-Retail. *International Journal of Electronic Commerce*, 24(3):305–337, 2020. doi: 10.1080/10864415.2020.1767428.

B J Chasnov, L J Ratliff, and S A Burden. Human adaptation to adaptive machines converges to game-theoretic equilibria. *Scientific Reports*, 15(1):1–10, 2025. doi: 10.1038/s41598-025-12998-1.

Paresh Chaudhary, Yancheng Liang, Daphne Chen, Simon S Du, and Natasha Jaques. Improving human-ai coordination through online adversarial training and generative models. *arXiv preprint arXiv:2504.15457*, 2025.

Shenghui Chen, Ruihan Zhao, Sandeep Chinchali, and Ufuk Topcu. Human-agent coordination in games under incomplete information via multi-step intent. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 481–489, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Jacob W Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, and Iyad Rahwan. Cooperating with Machines. *Nature Communications*, 9, 2018. doi: 10.1038/s41467-017-02597-8.

Dalia De Santis. A Framework for Optimizing Co-adaptation in Body-Machine Interfaces. *Frontiers in Neurorobotics*, 15, 2021. doi: 10.3389/fnbot.2021.662181.

Wyatt Felt, Jessica C. Selinger, J. Maxwell Donelan, and C. David Remy. "Body-In-The-Loop": Optimizing Device Parameters Using Measures of Instantaneous Energetic Cost. *PLOS ONE*, 10(8):1–21, 08 2015. doi: 10.1371/journal.pone.0135342.

Eric Frankel, Kshitij Kulkarni, Dmitriy Drusvyatskiy, Sewoong Oh, and Lillian J Ratliff. Finite-time convergence rates in stochastic stackelberg games with smooth algorithmic agents. In *Forty-second International Conference on Machine Learning*, 2025.

Guillaume Garrigós and Robert M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

Dominic Gorecky, Mathias Schmitt, Matthias Loskyll, and Detlef Zühlke. Human-machine-interaction in the industry 4.0 era. In *IEEE International Conference on Industrial Informatics (INDIN)*, pages 289–294, 2014. doi: 10.1109/INDIN.2014.6945523.

Jean-Michel Hoc. From human-machine interaction to human-machine cooperation. *Ergonomics*, 43(7):833–843, 2000. doi: 10.1080/001401300409044.

Maxence Hussonnois, Thommen George Karimpanal, and Santu Rana. Human-aligned skill discovery: Balancing behaviour exploration and alignment. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 1025–1033, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Jason T. Isa, Bohan Wu, Qirui Wang, Yilin Zhang, Samuel A. Burden, Lillian J. Ratliff, and Benjamin J. Chasnov. Effect of Adaptation Rate and Cost Display in a Human-AI Interaction Game, 2024. URL https://arxiv.org/abs/2408.14640.

Jason T. Isa, Samuel A. Burden, and Lillian J. Ratliff. Achieving alignment through adaptive play: Helping optimize objectives without observing them (code), 2026. URL https://doi.org/10.24433/CO.8589291.v1.

Sehyeok Kang, Minu Kim, Jihwan Oh, and Se-Young Yun. Dual preference learning for multi-agent reinforcement

learning. *IEEE Access*, 14:113–130, 2026. doi: 10.1109/ACCESS.2025.3645778.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak–łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, Cham, Switzerland, 2016. Springer.

Sebastian Kassing and Simon Weissmann. Polyak's heavy ball method achieves accelerated local rate of convergence under polyak–łojasiewicz inequality. *arXiv preprint arXiv:2410.16849*, 2024.

Qian Kou, Mingyang Li, Zeyang Liu, Long Qian, Zhuoran Chen, Lipeng Wan, Xingyu Chen, and Xuguang Lan. Offline multi-agent preference-based reinforcement learning with agent-aware direct preference optimization. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 1181–1190, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.

Andrew L. Kun. Human-Machine Interaction for Vehicles: Review and Outlook. *Foundations and Trends® in Human–Computer Interaction*, 11(4):201–293, 2018. doi: 10.1561/1100000069.

Benjamin Li, Shuyang Shi, Lucia Romero, Huao Li, Yaqi Xie, Woojun Kim, Stefanos Nikolaidis, Michael Lewis, Katia Sycara, and Simon Stepputtis. Adaptively coordinating with novel partners via learned latent strategies. *arXiv preprint arXiv:2511.12754*, 2025.

Yanan Li, Keng Peng Tee, Rui Yan, Wei Liang Chan, and Yan Wu. A Framework of Human–Robot Coordination Based on Game Theory and Policy Iteration. *IEEE Transactions on Robotics*, 32(6):1408–1418, 2016. doi: 10.1109/TRO.2016.2597322.

Yanan Li, Gerolamo Carboni, Franck Gonzalez, Domenico Campolo, and Etienne Burdet. Differential game theory for versatile physical human-robot interaction. *Nature Machine Intelligence*, 1(1):36–43, 2019. doi: 10.1038/s42256-018-0010-3.

Yuqing Liang, Jinlan Liu, and Dongpo Xu. Stochastic momentum methods for non-convex learning without bounded assumptions. *Neural Networks*, 165:830–845, 2023.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *arXiv preprint arXiv:2506.24119*, 2025.

Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. LLM-Powered Hierarchical Language Agent for Real-time Human-AI Coordination. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 1219–1228, 2024.

Haozhe Ma, Thanh Vinh Vo, and Tze-Yun Leong. Hierarchical reinforcement learning with human-ai collaborative sub-goals optimization. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 2310–2312, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450394321.

Christoph March. Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology*, 87:102426, 2021. doi: 10.1016/j.joep.2021.102426.

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Stefanos Nikolaidis, Swaprava Nath, Ariel D. Procaccia, and Siddhartha Srinivasa. Game-Theoretic Modeling of Human Adaptation in Human-Robot Collaboration. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 323–331. Association for Computing Machinery, 2017. doi: 10.1145/2909824.3020253.

Stefan Palan and Christian Schitter. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. doi: 10.1016/j.jbef.2017.12.004.

Serafeim Perdikis and Jose del R. Millan. Brain-machine interfaces: A tale of two learners. *IEEE Systems, Man, and Cybernetics Magazine*, 6(3):12–19, 2020. doi: 10.1109/MSMC.2019.2958200.

Clarice Poon, Jingwei Liang, and Carola Schoenlieb. Local convergence properties of SAGA/Prox-SVRG and acceleration. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4124–4132. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/poon18a.html.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW1), 2022. doi: 10.1145/3512930.

Benjamin Recht. A Tour of Reinforcement Learning: The View from Continuous Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019. doi: 10.1146/annurev-control-053018-023825.

David Shoresh and Yonatan Loewenstein. Modeling the centaur: Human-machine synergy in sequential decision making. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '25, page 1941–1949. International Foundation for Autonomous Agents and Multiagent Systems, 2025. ISBN 9798400714269.

Patrick Slade, Mykel J. Kochenderfer, Scott L. Delp, and Steven H. Collins. Personalizing exoskeleton assistance while walking in the real world. *Nature*, 610:277–282, 2022. doi: 10.1038/s41586-022-05191-1.

Chandler Smith, Marwa Abdulhai, Manfred Diaz, Marko Tesic, Rakshit S Trivedi, Alexander Sasha Vezhnevets, Lewis Hammond, Jesse Clifton, Minsuk Chang, Edgar A Duéñez-Guzmán, et al. Evaluating generalization capabilities of llm-based agents in mixed-motive scenarios using concordia. *arXiv preprint arXiv:2512.03318*, 2025.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. URL https://arxiv.org/abs/2407.17032.

Juanjuan Zhang, Pieter Fiers, Kirby A. Witte, Rachel W. Jackson, Katherine L. Poggensee, Christopher G. Atkeson, and Steven H. Collins. Human-in-the-loop optimization of exoskeleton assistance during walking. *Science*, 356(6344):1280–1284, 2017. doi: 10.1126/science.aal5054.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 295–305. Association for Computing Machinery, 2020. doi: 10.1145/3351095.3372852.

A ORGANIZATION OF APPENDIX

The appendix contains several sections; we provide a brief contents list to help the reader navigate the material.

- Appendix B: **Extended Proofs.** Complete proofs of the theoretical results in the main paper, including convergence guarantees for deterministic and stochastic inner-loop updates, auxiliary lemmas, and technical bounds used in the analysis.
- Appendix C: **Additional Algorithm Variants.** We provide schematic illustrations for both the uniform-sampling and round-robin variants in an idealized convex best-response setting. We then prove linear convergence to a noise floor for the round-robin variant. We also show the algorithm for "Outer-controlled drift via response gain ν_k and inner budget T_k"
- Appendix D: **Supplemental Methods: Human Participant Experiments.** Experimental protocol, task design and interface details, cost specification, algorithmic parameters, participant recruitment, and data processing procedures for the human-subject studies.
- Appendix E: **Supplemental Methods: Closed-Loop Two-Agent Balancing Cart-Pole Game.** System dynamics, spring–damper coupling model, cost function, gain-space policy parameterization, LQR-initialized, deterministic gain optimization, gain-space PPO implementation, and training and evaluation procedures for the cart-pole experiments.
- Appendix F: **Additional Results: Human Participant Experiments.** Extended quantitative results, additional plots across all action-space conditions, ablations comparing uniform and round-robin subspace selection, and robustness analyses for the human-subject experiments.
- Appendix G: **Additional Experiment: Open-Loop LQ Game with Non-Minimum Phase Plant.** Methods including plant dynamics, cost formulation, transfer function and state-space models, controller implementation details, algorithm parameters, and simulation settings for the non-minimum-phase LQ experiments. Results and figures.

B EXTENDED PROOFS

Complete proofs of the theoretical results in the main paper, including convergence guarantees for deterministic and stochastic inner-loop updates, auxiliary lemmas, and technical bounds used in the analysis.

B.1 PRELIMINARIES

Let $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ and define the joint action

$$z := (x, u) \in \mathbb{R}^d, \quad d := d_1 + d_2,$$

and by a slight abuse of notation $f(z) := f(x, u)$. Define the global optimum and corresponding cost by

$$f^* := \min_{z \in \mathbb{R}^d} f(z), \qquad z^* \in \arg\min_{z \in \mathbb{R}^d} f(z).$$

We impose the following regularity assumptions to characterize the problem class and the geometric properties of the induced affine subspaces.

Assumption 1. (Regularity). The function f satisfies:

a. f is β-smooth, i.e., $f(z') \le f(z) + \nabla f(z)^\top (z' - z) + \frac{\beta}{2}\|z' - z\|^2 \; \forall z, z'$.
b. The Polyak-Łojasiewicz (PŁ) inequality holds, i.e., $\|\nabla f(z)\|^2 \ge 2\mu_{\mathrm{PL}}\big(f(z) - f^*\big) \; \forall z$.

Let $\mathcal{L} = \{L_1, \dots, L_N\}$ with $L_i \in \mathbb{R}^{d_2 \times d_1}$ be the set of policy parameters for Agent 2 (the *helper*), and define

$$V_i := \begin{bmatrix} I_{d_1} \\ L_i \end{bmatrix} \in \mathbb{R}^{d \times d_1}, \qquad \mathcal{S}_i := \mathrm{range}(V_i) \subset \mathbb{R}^d.$$

Let $P_i := V_i \, (V_i^\top V_i)^{-1} V_i^\top$ denote the orthogonal projection onto $\mathcal{S}_i$.

Assumption 2. (Alignment of subspaces). Assume there exists $\lambda \in (0, 1]$ such that for all $g \in \mathbb{R}^d$,

$$\max_{i=1,\dots,N} \|P_i g\|^2 \geq \lambda \|g\|^2.$$

Moreover, set $V_i^\top V_i = I_{d_1} + L_i^\top L_i$ and define $\kappa := \max_{i=1,\dots,N} \lambda_{\max}(V_i^\top V_i) < \infty$.

Let $\varphi_k^* := \min_x \varphi_k(x)$ and $\Delta_k^\varphi(x) := \mathbb{E}_k[\varphi_k(x) - \varphi_k^*]$ where $\mathbb{E}_k[\cdot] \equiv \mathbb{E}[\cdot \mid \mathcal{F}_{k-1}]$.

Assumption 3. (One-step inner-loop function-gap contraction). For each fixed epoch k, the inner-loop algorithm generates iterates $\{x_{k,t}\}_{t\geq 0}$ with initialization $x_{k,0} := x_{k-1}$ and there exist constants $q_k \in [0, 1)$ and $\delta_k \geq 0$ such that for every inner iteration $t \geq 0$, the bound holds:

$$\Delta_k^\varphi(x_{k,t+1}) \leq q_k \, \Delta_k^\varphi(x_{k,t}) \; + \; (1 - q_k) \, \delta_k.$$

Stochastic Framework. We formalize the randomness in the outer-loop sampling and inner-loop updates via a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_k)_{k\geq 0}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$. For each epoch $k \geq 1$:

a. The sampled index $i_k : \Omega \to \{1, \dots, N\}$ is $\mathcal{F}_k$-measurable, and $\mathcal{F}_{k-1} \subseteq \bar{\sigma}(\mathcal{F}_{k-1}, i_k) \subseteq \mathcal{F}_k$ where $\mathfrak{F}(\cdot)$ is the sigma algebra generator.

b. The inner-loop randomness in epoch k is represented by a collection $\{\omega_{k,t}\}_{t=0}^{T_k-1}$ of random variables and a within-epoch filtration $(\mathcal{F}_{k,t})_{t=0}^{T_k}$ satisfying $\mathcal{F}_{k,0} := \bar{\sigma}(\mathcal{F}_{k-1}, i_k)$, $\mathcal{F}_{k,t+1} := \bar{\sigma}(\mathcal{F}_{k,t}, \omega_{k,t})$, $t = 0, \dots, T_k - 1$, and $\mathcal{F}_k := \mathcal{F}_{k,T_k}$.

c. The epoch iterates are adapted: $z_k : \Omega \to \mathbb{R}^d$ is $\mathcal{F}_k$-measurable, and the inner iterates $z_{k,t}$ (and $x_{k,t}$) are $\mathcal{F}_{k,t}$-measurable for all t.

This framework specifies the probabilistic setting used in the analysis.

B.2 PROOFS OF LEMMAS AND THEORIES FROM MAIN PAPER

B.2.1 Lemma 1 (One-epoch progress)

We consider the general stochastic setting in which the inner-loop update $\mathcal{A}$ may be noisy but satisfies the function-gap contraction property in Assumption 3.

Define $f_i^*(z) := \min_{y \in z + \mathcal{S}_i} f(y)$. Set $\varphi_k^* := \min_x \varphi_k(x)$, and observe that $\varphi_k^* = f_i^*(z)$ and $f(z_k) = \varphi_k(x_k)$. Define $\Delta_k^f(z_k, z) := \mathbb{E}\left[f(z_k) - f_i^*(z) \mid \mathcal{F}_{k,0}\right]$.

Lemma 1. (One-epoch progress). Fix an epoch k, and set $z := z_{k-1}$. Under Assumptions 1, 2, and 3, the estimate holds:

$$\Delta_k^f(z_k, z) \leq q_k^{T_k} \Delta_k^\varphi(x_{k-1}) + \delta_k\big(1 - q_k^{T_k}\big). \tag{1}$$

Proof. Condition on $\mathcal{F}_{k,0} = \sigma(\mathcal{F}_{k-1}, i_k)$ so that $i = i_k$ is fixed. By construction of the epoch-k inner problem, the inner objective φ_k minimizes f over the affine subspace $z + \mathcal{S}_i$ (via the parametrization $x \mapsto (x, \pi_k(x))$), hence $\varphi_k^* = f_i^*(z)$. Moreover, the epoch output satisfies $z_k = (x_k, \pi_k(x_k))$ and thus $f(z_k) = \varphi_k(x_k)$. Therefore,

$$\mathbb{E}[f(z_k) - f_i^*(z) \,|\, \mathcal{F}_{k,0}] = \mathbb{E}[\varphi_k(x_k) - \varphi_k^* \,|\, \mathcal{F}_{k,0}]\,.$$

It remains to prove the contraction bound. Let $\Delta_{k,t} := \mathbb{E}[\varphi_k(x_{k,t}) - \varphi_k^* \mid \mathcal{F}_{k,0}]$. Assumption 3 (applied with conditioning $\mathcal{F}_{k,0}$, which fixes φ_k) gives, for each $t \geq 0$,

$$\Delta_{k,t+1} \leq q_k \, \Delta_{k,t} + (1 - q_k)\delta_k, \qquad \Delta_{k,0} = \varphi_k(x_{k-1}) - \varphi_k^*,$$

since $x_{k,0} = x_{k-1}$ is $\mathcal{F}_{k,0}$-measurable.

Unrolling this linear recursion yields

$$\begin{aligned}\Delta_{k,T_k} &\leq q_k^{T_k}\Delta_{k,0} + (1 - q_k)\delta_k \sum_{j=0}^{T_k-1} q_k^j \\ &= q_k^{T_k}\Delta_{k,0} + \delta_k(1 - q_k^{T_k}),\end{aligned}$$

which is the result after substituting $\Delta_{k,0} = \varphi_k(x_{k-1}) - \varphi_k^*$ and $x_k = x_{k,T_k}$. □

B.2.2 Theorem 1 (Outer-loop linear convergence under a global PŁ condition)

Building on Lemma 1, we obtain global linear convergence of the outer loop up to a noise floor.

To prove Theorem 1, we must first introduce Lemma 2.

Lemma 2. (Subspace gap lower bound from smoothness). Fix a point $z \in \mathbb{R}^d$ and a linear subspace $\mathcal{S}_i \subseteq \mathbb{R}^d$ with orthogonal projection matrix P_i. Suppose that Assumption 1.a holds—i.e., f is β smooth. Define the subspace optimum $f_i^*(z) := \min_{y \in z+\mathcal{S}_i} f(y)$. Then

$$f(z) - f_i^*(z) \;\geq\; \frac{1}{2\beta}\,\|P_i \nabla f(z)\|^2.$$

Proof. By β–smoothness, for any $s \in \mathcal{S}_i$ we have

$$f(z+s) \leq f(z) + \nabla f(z)^\top s + \frac{\beta}{2}\|s\|^2.$$

Since $s \in \mathcal{S}_i$ and P_i is the orthogonal projection matrix onto $\mathcal{S}_i$, we have that

$$\nabla f(z)^\top s = (P_i \nabla f(z))^\top s.$$

Consider the quadratic upper bound which is given by

$$\phi(s) := (P_i \nabla f(z))^\top s + \frac{\beta}{2}\|s\|^2, \qquad s \in \mathcal{S}_i.$$

This function is minimized over $\mathcal{S}_i$ at

$$s^\star = -\frac{1}{\beta} P_i \nabla f(z),$$

and the minimum value is

$$\min_{s \in \mathcal{S}_i} \phi(s) = \phi(s^\star) = -\frac{1}{2\beta}\|P_i \nabla f(z)\|^2.$$

Therefore, we deduce that

$$f_i^*(z) = \min_{s \in \mathcal{S}_i} f(z+s) \leq \min_{s \in \mathcal{S}_i} \Big(f(z) + \phi(s) \Big) = f(z) - \frac{1}{2\beta}\|P_i \nabla f(z)\|^2.$$

Rearranging yields the claim:

$$f(z) - f_i^*(z) \;\geq\; \frac{1}{2\beta}\|P_i \nabla f(z)\|^2.$$

□

Theorem 1. (Outer-loop linear convergence under a global PŁ condition). Suppose Assumptions 1, 2, and 3 hold. The estimate holds:

$$\mathbb{E}[f(z_k) - f^*] \leq \rho_k\, \mathbb{E}[f(z_{k-1}) - f^*] + (1 - q_k^{T_k})\delta_k,$$

where $\rho_k := 1 - (1 - q_k^{T_k})\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N} \in (0,1)$. If $(1 - q_k^{T_k}) = \theta$ and $\delta_k = \delta$ are constant across k, then

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k \big(f(z_0) - f^* \big) + \delta\theta/(1-\rho).$$

Proof. Fix k and condition on $z := z_{k-1}$ and $i := i_k$—i.e., $\mathcal{F}_{k,0} := \sigma(\mathcal{F}_{k-1}, i_k)$. By Lemma 1, we have that

$$\mathbb{E}[f(z_k) - f_i^*(z) \mid \mathcal{F}_{k,0}] \leq \tilde{q}_k^{T_k} \big(f(z) - f_i^*(z) \big) + \delta_k (1 - \tilde{q}_k^{T_k}).$$

Let $\theta_k := 1 - \tilde{q}_k^{T_k}$, and rearrange the above inequality to get that

$$\begin{aligned} \mathbb{E}[f(z_k) \mid \mathcal{F}_{k,0}] &\leq f_i^*(z) + (1-\theta_k)\big(f(z) - f_i^*(z) \big) + \theta_k \delta_k \\ &= f(z) - \theta_k \big(f(z) - f_i^*(z) \big) + \theta_k \delta_k. \end{aligned}$$

Now, subtract f^* from both sides to get that

$$\mathbb{E}[f(z_k) - f^* \mid \mathcal{F}_{k,0}] \leq f(z) - f^* - \theta_k\big(f(z) - f_i^*(z)\big) + \theta_k\delta_k.$$

Using the subspace gap bound from Lemma 2, we have that

$$f(z) - f_i^*(z) \geq \frac{1}{2\beta}\|P_i\nabla f(z)\|^2,$$

so that

$$\mathbb{E}[f(z_k) - f^* \mid \mathcal{F}_{k,0}] \leq f(z) - f^* - \theta_k\frac{1}{2\beta}\|P_i\nabla f(z)\|^2 + \theta_k\delta_k.$$

Now taking the expectation over uniform sampling of $i_k = i \in [N]$ and combining this with the alignment condition from Assumption 2, we have that

$$\mathbb{E}_i\|P_i\nabla f(z)\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|P_i\nabla f(z)\|^2 \geq \frac{\lambda}{N}\|\nabla f(z)\|^2.$$

Thus, we conclude

$$\mathbb{E}_i[f(z_k) - f^* \mid \mathcal{F}_{k-1}] \leq f(z) - f^* - \theta_k\frac{\lambda}{2\beta N}\|\nabla f(z)\|^2 + \theta_k\delta_k.$$

Applying the global PŁ inequality, gives us that

$$\mathbb{E}_i[f(z_k) - f^* \mid \mathcal{F}_{k-1}] \leq \left(1 - \theta_k\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N}\right)(f(z) - f^*) + \theta_k\delta_k.$$

Setting $\rho_k := 1 - \theta_k\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N}$ and taking the total expectation over $z = z_{k-1}$ yields the first claimed inequality:

$$\mathbb{E}[f(z_k) - f^*] \leq \rho_k\,\mathbb{E}[f(z_{k-1}) - f^*] + \theta_k\delta_k.$$

If $\rho_k = \rho$ and $\theta_k = \theta$ are constant, unrolling the recursion gives

$$\begin{aligned}\mathbb{E}[f(z_k) - f^*] &\leq \rho^k(f(z_0) - f^*) + \theta\delta\sum_{j=0}^{k-1}\rho^j \\ &= \rho^k(f(z_0) - f^*) + \frac{\theta}{1-\rho}\delta(1 - \rho^k) \\ &\leq \rho^k(f(z_0) - f^*) + \frac{\theta}{1-\rho}\delta,\end{aligned}$$

which concludes the proof. □

B.2.3 Corollary 1 (Outer-loop convergence with SGD)

As a concrete instantiation of Theorem 1, consider Agent 1 running stochastic gradient descent.

To prove Corollary 1, we must first introduce Lemma 3 & 4.

Let the randomness be only the outer samples $\{L_k\}$. Define the filtration

$$\mathcal{F}_k := \sigma(L_1, \ldots, L_k, u_0).$$

so that u_k is $\mathcal{F}_k$-measurable and L_{k+1} is independent of $\mathcal{F}_k$ since the samples are i.i.d. uniform.

Lemma 3. (Iterates remain in the induced affine subspace). Fix an epoch k and let $L_k := L_{i_k}$. Let

$$z_{k,t} := \begin{bmatrix} x_{k,t} \\ u_{k,t} \end{bmatrix} \quad \text{and} \quad z_{k,0} := \begin{bmatrix} x_{k-1} \\ u_{k-1} \end{bmatrix}.$$

Then, for all t, the iterates remain in the subspace: namely, we have

$$z_{k,t} \in z_{k,0} + \mathcal{S}_i \quad \forall\, t \in \{0, \ldots, T\}.$$

Proof. By construction, we have that

$$u_{k,t} = \pi_k(x_{k,t}) = u_{k-1} + L_i(x_{k,t} - x_{k-1}).$$

Define $v_{k,t} := x_{k,t} - x_{k-1} \in \mathbb{R}^{d_1}$. Then

$$z_{k,t} - z_{k,0} = \begin{bmatrix} x_{k,t} - x_{k-1} \\ u_{k,t} - u_{k-1} \end{bmatrix} = \begin{bmatrix} v_{k,t} \\ L_i v_{k,t} \end{bmatrix} = V_i v_{k,t} \in \mathrm{range}(V_i) = \mathcal{S}_i.$$

□

Lemma 4. (One-epoch expected progress toward the subspace optimum with SGD under PL). Suppose Assumptions 1.a, 1.b, 2 and the Stochastic Framework hold. Fix epoch k and condition on $\mathcal{F}_{k,0}$ so that $z = z_{k-1}$, $i = i_k$, and hence φ_k are fixed. Let $\varphi_k^* := \min_x \varphi_k(x) = f_i^*(z)$. Suppose the inner loop is stochastic gradient descent with step size $\eta \le 1/\beta\kappa$. Set $q := 1 - \eta\mu_{\mathrm{PL},k} \in (0,1)$ and $\theta := 1 - q^{T_k}$. Then the inner output $x_k := x_{k,T_k}$ satisfies

$$\mathbb{E}\left[f(z_k) - f_i^*(z) \,|\, \mathcal{F}_{k,0}\right] = \mathbb{E}\left[\varphi_k(x_{k,T_k}) - \varphi_k^* \,|\, \mathcal{F}_{k,0}\right] \le q^{T_k}\left(\varphi_k(x_{k,0}) - \varphi_k^*\right) + \frac{\beta_k \eta \sigma^2}{2\mu_{\mathrm{PL},k}}\left(1 - q^{T_k}\right).$$

Proof. Fix k and condition on $\mathcal{F}_{k,0}$ so that φ_k is deterministic. By β_k-smoothness of φ_k, for any x and y, we have that

$$\varphi_k(y) \le \varphi_k(x) + \nabla\varphi_k(x)^\top (y - x) + \frac{\beta_k}{2}\|y - x\|^2.$$

Applying this bound with $x = x_{k,t}$ and $y = x_{k,t+1} = x_{k,t} - \eta g_k(x_{k,t})$, we have that

$$\varphi_k(x_{k,t+1}) \le \varphi_k(x_{k,t}) - \eta\,\nabla\varphi_k(x_{k,t})^\top g_k(x_{k,t}) + \frac{\beta_k \eta^2}{2}\|g_k(x_{k,t})\|^2.$$

Now, we take the conditional expectation given $\mathcal{F}_{k,t}$, and using the fact that the gradient estimate is unbiased to deduce that

$$\mathbb{E}\left[\nabla\varphi_k(x_{k,t})^\top g_k(x_{k,t}) \mid \mathcal{F}_{k,t}\right] = \|\nabla\varphi_k(x_{k,t})\|^2.$$

Moreover, we have that

$$\begin{aligned}\mathbb{E}\left[\|g_k(x_{k,t})\|^2 \mid \mathcal{F}_{k,t}\right] &= \|\nabla\varphi_k(x_{k,t})\|^2 + \mathbb{E}\left[\|g_k(x_{k,t}) - \nabla\varphi_k(x_{k,t})\|^2 \mid \mathcal{F}_{k,t}\right] \\ &\le \|\nabla\varphi_k(x_{k,t})\|^2 + \sigma^2.\end{aligned}$$

Therefore, we further deduce that

$$\begin{aligned}\mathbb{E}[\varphi_k(x_{k,t+1}) \mid \mathcal{F}_{k,t}] &\le \varphi_k(x_{k,t}) - \eta\|\nabla\varphi_k(x_{k,t})\|^2 + \frac{\beta_k\eta^2}{2}\left(\|\nabla\varphi_k(x_{k,t})\|^2 + \sigma^2\right) \\ &= \varphi_k(x_{k,t}) - \left(\eta - \frac{\beta_k\eta^2}{2}\right)\|\nabla\varphi_k(x_{k,t})\|^2 + \frac{\beta_k\eta^2}{2}\sigma^2.\end{aligned}$$

Under $\eta \le 1/\beta\kappa$ (where $\beta_k \le \beta\kappa$ for all k), we have $\eta - \frac{\beta_k\eta^2}{2} \ge \eta/2$, so

$$\mathbb{E}[\varphi_k(x_{k,t+1}) \mid \mathcal{F}_{k,t}] \le \varphi_k(x_{k,t}) - \frac{\eta}{2}\|\nabla\varphi_k(x_{k,t})\|^2 + \frac{\beta_k\eta^2}{2}\sigma^2.$$

Subtracting φ_k^* and applying the PŁ inequality, namely $\|\nabla\varphi_k(x_{k,t})\|^2 \ge 2\mu_{\mathrm{PL},k}(\varphi_k(x_{k,t}) - \varphi_k^*)$, we have that

$$\mathbb{E}[\varphi_k(x_{k,t+1}) - \varphi_k^* \mid \mathcal{F}_{k,t}] \le \left(1 - \eta\mu_{\mathrm{PL},k}\right)\left(\varphi_k(x_{k,t}) - \varphi_k^*\right) + \frac{\beta_k\eta^2}{2}\sigma^2.$$

Letting $q := 1 - \eta\mu_{\mathrm{PL},k}$ and iterating for $t = 0, \ldots, T_k - 1$, we have

$$\begin{aligned}\mathbb{E}[\varphi_k(x_{k,T_k}) - \varphi_k^* \mid \mathcal{F}_{k,0}] &\le q^{T_k}\left(\varphi_k(x_{k,0}) - \varphi_k^*\right) + \frac{\beta_k\eta^2}{2}\sigma^2\sum_{s=0}^{T_k-1} q^s \\ &= q^{T_k}\left(\varphi_k(x_{k,0}) - \varphi_k^*\right) + \frac{\beta_k\eta^2}{2}\sigma^2 \cdot \frac{1 - q^{T_k}}{1 - q}.\end{aligned}$$

Since $1-q=\eta\mu_{\mathrm{PL},k}$ and $x_{k,0}=x_{k-1}$, this yields

$$\mathbb{E}[\varphi_k(x_{k,T_k})-\varphi_k^* \mid \mathcal{F}_{k,0}] \le q^{T_k}\big(\varphi_k(x_{k,0})-\varphi_k^*\big)+\frac{\beta_k\eta\sigma^2}{2\mu_{\mathrm{PL},k}}(1-q^{T_k}),$$

which proves the first inequality. The second displayed inequality follows by substituting $\varphi_k^*=f_i^*(z)$, $\varphi_k(x_{k,0})=f(z)$, $\varphi_k(x_{k,T_k})=f(z_k)$ (which follows from Lemma 3) and rearranging with $\theta:=1-q^{T_k}$. □

Corollary 1. (Outer-loop convergence with SGD $\mathcal{A}$). Suppose Assumptions 1 and 2 hold, and that Agent 1 runs stochastic gradient descent with stepsize $\eta<1/(\beta\kappa)$ and with unbiased gradients of bounded variance σ^2. Fix constants $q:=1-\eta\mu_{\mathrm{PL}}$, $\theta:=1-q^T$, and assume a fixed inner-loop length $T_k\equiv T$ for all epochs. The outer-loop iterates satisfy

$$\mathbb{E}\left[f(z_k)-f^*\right] \le \rho^k\big(f(z_0)-f^*\big)+\tfrac{N\beta}{\lambda}\cdot\tfrac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}^2} \tag{2}$$

where $\rho:=1-\theta\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N}\in(0,1)$.

Proof. Fix k and condition on $z:=z_{k-1}$ and $i:=i_k$—i.e., $\mathcal{F}_{k,0}=\sigma(\mathcal{F}_{k-1},i_k)$. By Lemma 4, replacing $\mu_{\mathrm{PL},k}$ with the global μ_{PL} and using $\beta_k\le\beta\kappa$, we have that

$$\mathbb{E}\left[f(z_k)\mid\mathcal{F}_{k,0}\right]\le f(z)-\theta\big(f(z)-f_i^*(z)\big)+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Now, subtract f^* from both sides to get

$$\mathbb{E}\left[f(z_k)-f^*\mid\mathcal{F}_{k,0}\right]\le f(z)-f^*-\theta\big(f(z)-f_i^*(z)\big)+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Applying the subspace gap lower bound from Lemma 2, we get that

$$\mathbb{E}\left[f(z_k)-f^*\mid\mathcal{F}_{k,0}\right]\le f(z)-f^*-\theta\frac{1}{2\beta}\|P_i\nabla f(z)\|^2+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Using the alignment condition from Assumption 2, we have that

$$\mathbb{E}_i\|P_ig\|^2=\frac{1}{N}\sum_{i=1}^N\|P_ig\|^2\ge\frac{1}{N}\max_i\|P_ig\|^2\ge\frac{\lambda}{N}\|g\|^2,$$

so that

$$\mathbb{E}_i\left[f(z_k)-f^*\mid z\right]\le f(z)-f^*-\theta\frac{\lambda}{2\beta N}\|\nabla f(z)\|^2+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Applying the global PŁ inequality for f, we have that

$$\|\nabla f(z)\|^2\ge2\mu_{\mathrm{PL}}\big(f(z)-f^*\big),$$

Therefore we deduce that

$$\mathbb{E}_i\left[f(z_k)-f^*\mid\mathcal{F}_{k,0}\right]\le\left(1-\theta\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N}\right)\left(f(z)-f^*\right)+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Define $\rho:=1-\theta\frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda}{N}$. Taking full expectation gives the claimed recursion:

$$\mathbb{E}[f(z_k)-f^*]\le\rho\,\mathbb{E}[f(z_{k-1})-f^*]+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.$$

Finally, we unroll the recursion to get that

$$\begin{aligned}\mathbb{E}[f(z_k)-f^*]&\le\rho^k\big(f(z_0)-f^*\big)+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}\sum_{j=0}^{k-1}\rho^j\\&=\rho^k\big(f(z_0)-f^*\big)+\theta\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}\cdot\frac{1-\rho^k}{1-\rho}\\&\le\rho^k\big(f(z_0)-f^*\big)+\frac{\theta}{1-\rho}\cdot\frac{\beta\kappa\eta\sigma^2}{2\mu_{\mathrm{PL}}}.\end{aligned}$$

□

Remark 1 (Strongly convex global case)**.** If f is globally α_f-strongly convex (instead of only PL), then f satisfies the PL inequality with $\mu_{\mathrm{PL}} = \alpha_f$. Thus Corollary 1 holds with μ_{PL} replaced by α_f.

B.2.4 Proposition 1 (Iteration Complexity)

Proposition 1. (Iteration Complexity). Assume the setting of Corollary 1. Set constants $\rho := 1 - \theta c \in (0, 1)$ with $c := \mu_{\mathrm{PL}}\lambda/(\beta N)$ and $\delta := \beta\kappa\eta\sigma^2/(2\mu_{\mathrm{PL}})$ so that the noise floor in (2) is given by $\epsilon_\infty := \delta\theta/(1-\rho)$. Consequently, for any target accuracy $\epsilon_\star > \epsilon_\infty$, it suffices to choose

$$k \geq \frac{1}{\log(1/\rho)} \log\left(\frac{f(z_0) - f^*}{\epsilon_\star - \epsilon_\infty}\right) \tag{3}$$

to guarantee $\mathbb{E}\left[f(z_k) - f^*\right] \leq \epsilon_\star$.

Proof. From Corollary 1 we have that

$$\mathbb{E}\left[f(z_k) - f^*\right] \leq \rho^k\big(f(z_0) - f^*\big) + \frac{\theta}{1-\rho}\,\delta,$$

with $\rho = 1 - \theta c$ and $\theta = 1 - q^T$. Set $\epsilon_\infty := \frac{\theta}{1-\rho}\delta$. Then

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k(f(z_0) - f^*) + \epsilon_\infty.$$

If $\epsilon_\star > \epsilon_\infty$, it suffices that

$$\rho^k(f(z_0) - f^*) \leq \epsilon_\star - \epsilon_\infty,$$

equivalently

$$k \log\rho \leq \log\left(\frac{\epsilon_\star - \epsilon_\infty}{f(z_0) - f^*}\right).$$

Since $\log\rho < 0$, this yields (3) as desired. □

Remark 2 (Inner-loop length T and the speed–computation tradeoff)**.** With $\theta = 1 - q^T$, increasing T increases θ and decreases

$$\rho(T) = 1 - c(1 - q^T),$$

so the number of outer epochs needed to reach a fixed $\epsilon_\star > \epsilon_\infty$ decreases like $1/(1-q^T)$ according to (3). However, each epoch costs T inner SGD steps. Thus the *total* inner steps needed scales as

$$kT \gtrsim \frac{T}{c(1-q^T)} \log\left(\frac{f(z_0) - f^*}{\epsilon_\star - \epsilon_\infty}\right),$$

highlighting the tradeoff between longer inner solves (larger T) and fewer outer epochs (smaller k).

B.2.5 Theorem 2 (Outer-loop recursion)

To prove Theorem 2, we must first introduce Lemma 5.

Lemma 5. (One-epoch progress toward the subspace optimum under one-step contraction). Suppose Assumptions 1.a, 1.b, 2, 3, and the Stochastic Framework hold. Fix epoch k and condition on $z := z_{k-1}$ and on i_k. Set $q_k := \tilde{q}_k \in (0, 1)$ and $\theta_k := 1 - q_k^{T_k}$. The estimate holds:

$$\mathbb{E}\left[f(z_k) \mid \mathcal{F}_{k,0}\right] \leq f(z) - \theta_k\big(f(z) - f_k^*(z)\big) + \theta_k\,\delta_k. \tag{8}$$

Proof. Define $\Delta_{k,t} := \mathbb{E}[\varphi_k(x_{k,t}) - \varphi_k^* \,|\, \mathcal{F}_{k,0}]$. Since $x_{k,0} = x_{k-1}$ is $\mathcal{F}_{k,0}$-measurable, $\Delta_{k,0} = \varphi_k(x_{k-1}) - \varphi_k^* = f(z) - f_k^*(z)$. By Assumption 3, for every $t \geq 0$, we have that

$$\mathbb{E}\left[\varphi_k(x_{k,t+1}) - \varphi_k^* \,|\, \mathcal{F}_{k,0}\right] \leq \tilde{q}_k\,\mathbb{E}\left[\varphi_k(x_{k,t}) - \varphi_k^* \,|\, \mathcal{F}_{k,0}\right] + (1 - \tilde{q}_k)\delta_k,$$

that is, $\Delta_{k,t+1} \le q_k \Delta_{k,t} + (1-q_k)\delta_k$. Unrolling this linear recursion gives

$$\Delta_{k,T_k} \le q_k^{T_k} \Delta_{k,0} + (1-q_k)\delta_k \sum_{s=0}^{T_k-1} q_k^s \le q_k^{T_k} \Delta_{k,0} + \delta_k (1 - q_k^{T_k}).$$

Since $x_{k,T_k} = x_k$ and $\varphi_k(x_k) = f(z_k)$, we have that

$$\mathbb{E}\left[f(z_k) - f_k^*(z) \,|\, \mathcal{F}_{k,0}\right] \le q_k^{T_k}\big(f(z) - f_k^*(z)\big) + \delta_k(1 - q_k^{T_k}).$$

Rewriting with $\theta_k = 1 - q_k^{T_k}$ yields

$$\mathbb{E}[f(z_k) \mid \mathcal{F}_{k,0}] \le f(z) - \theta_k\big(f(z) - f_k^*(z)\big) + \theta_k \delta_k.$$

□

At outer epoch k, sample $i_k \sim \mathrm{Unif}(\{1, \dots, N\})$ and set $L_k := L_{i_k}$ as before, but now introduce a response-gain parameter $\nu_k \in (0, 1]$ and define the epoch-k affine policy

$$\pi_k(x) := u_{k-1} + \nu_k L_k (x - x_{k-1}). \tag{9}$$

For each response-gain value $\nu \in (0, 1]$, define the scaled subspace matrix

$$V_i(\nu) := \begin{bmatrix} I_{d_1} \\ \nu L_i \end{bmatrix}, \qquad \mathcal{S}_i(\nu) := \mathrm{range}(V_i(\nu)).$$

Let

$$P_i(\nu) := V_i(\nu)\big(V_i(\nu)^\top V_i(\nu)\big)^{-1} V_i(\nu)^\top$$

denote the orthogonal projection onto $\mathcal{S}_i(\nu)$. We define the corresponding scaled alignment constant by

$$\lambda(\nu) := \inf_{g \neq 0} \frac{\max_{i=1,\dots,N} \|P_i(\nu) g\|^2}{\|g\|^2}.$$

Equivalently, $\lambda(\nu)$ is the largest constant such that

$$\max_{i=1,\dots,N} \|P_i(\nu) g\|^2 \ge \lambda(\nu) \|g\|^2 \qquad \forall g \in \mathbb{R}^d.$$

When the epoch-k response gain is ν_k, the sampled affine policy induces the affine subspace

$$z_{k-1} + \mathcal{S}_{i_k}(\nu_k),$$

and the one-epoch progress bound uses the projection $P_{i_k}(\nu_k)$ and alignment constant $\lambda(\nu_k)$.

Theorem 2. (Outer-loop recursion) Under Assumptions 1, 2, and 3, the outer iterates satisfy

$$\Delta_k^\star \le \big(1 - \theta_k(\nu_k)\Gamma_k(\nu_k)\big)\Delta_{k-1}^\star + \theta_k(\nu_k)\, \delta_k(\nu_k), \tag{5}$$

where $\Delta_k^\star := \mathbb{E}[f(z_k) - f^*]$, $\Gamma_k(\nu_k) := \frac{\mu_{\mathrm{PL}}}{\beta} \cdot \frac{\lambda(\nu_k)}{N}$, $\theta_k(\nu_k) := 1 - q_k(\nu_k)^{T_k}$ and $\delta_k(\nu_k)$ is the one-step noise-floor parameter appearing in Assumption 3.

Proof. Fix k and condition on $\mathcal{F}_{k,0} = \sigma(\mathcal{F}_{k-1}, i_k)$ so that $z := z_{k-1}$ and i_k (hence φ_k) are fixed. By Lemma 5, we have that

$$\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k,0}\right] \le f(z) - f^* - \theta_k\big(f(z) - f_k^*(z)\big) + \theta_k\, \delta_k.$$

Applying Lemma 2 with $P_k = P_{i_k}(\nu_k)$, we lower bound the subspace gap as follows:

$$f(z) - f_k^*(z) \ge \frac{1}{2\beta} \|P_k \nabla f(z)\|^2.$$

Thus, we have that

$$\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k,0}\right] \leq f(z) - f^* - \theta_k \frac{1}{2\beta}\|P_k \nabla f(z)\|^2 + \theta_k\, \delta_k.$$

Now, taking the expectation over $i_k \sim \mathrm{Unif}(\{1, \ldots, N\})$ using the tower property, we have that

$$\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k-1}\right] = \mathbb{E}\left[\,\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k,0}\right] | \, \mathcal{F}_{k-1}\right].$$

By Assumption 2, we have that

$$\mathbb{E}\left[\|P_{i_k}(\nu_k) g\|^2 \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{1}{N}\sum_{i=1}^{N} \|P_i(\nu_k) g\|^2 \geq \frac{\lambda(\nu_k)}{N}\|g\|^2 = \frac{\lambda_k}{N}\|g\|^2.$$

With $g = \nabla f(z)$, this yields

$$\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k-1}\right] \leq f(z) - f^* - \theta_k \frac{\lambda_k}{2\beta N}\|\nabla f(z)\|^2 + \theta_k\, \delta_k.$$

Applying the PŁ inequality, we have that

$$\|\nabla f(z)\|^2 \geq 2\mu_{\mathrm{PL}}\big(f(z) - f^*\big),$$

and therefore we deduce that

$$\mathbb{E}\left[f(z_k) - f^* \,|\, \mathcal{F}_{k-1}\right] \leq \Big(1 - \theta_k \frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda_k}{N}\Big)\big(f(z) - f^*\big) + \theta_k\, \delta_k = (1 - \theta_k \Gamma_k)\big(f(z) - f^*\big) + \theta_k\, \delta_k.$$

Finally, take total expectation over $\mathcal{F}_{k-1}$ and use $\mathbb{E}[f(z) - f^*] = \mathbb{E}[f(z_{k-1}) - f^*]$ to obtain (5). □

B.2.6 Corollary 2 (Outer recursion with SGD $\mathcal{A}$)

As a concrete instantiation of Theorem 2, consider Agent 1 running stochastic gradient descent.

Define

$$\Gamma_k := \frac{\mu_{\mathrm{PL}}}{\beta} \cdot \frac{\lambda_k}{N}, \qquad \lambda_k = \lambda(\nu_k).$$

Set constants $\lambda_k = \lambda(\nu_k)$, where $\lambda(\nu_k)$ indicates the dependence of the alignment parameter on ν_k.

Corollary 2. (Outer recursion with SGD $\mathcal{A}$). Assume that Agent 1 runs stochastic gradient descent with step size $\eta \leq 1/\beta_k(\nu_k)$ and that φ_k is $\mu_k(\nu_k)$-PŁ and $\beta_k(\nu_k)$-smooth, with unbiased stochastic gradients of variance at most σ^2. Then Assumption 3 holds with $q_k(\nu_k) = 1 - \eta\mu_k(\nu_k)$, $\theta_k(\nu_k) = 1 - (1 - \eta\mu_k(\nu_k))^{T_k}$, and $\delta_k(\nu_k) = \frac{\beta_k(\nu_k)\eta\sigma^2}{2\mu_k(\nu_k)}$. Hence (5) reduces to

$$\Delta_k^\star \leq \big(1 - \theta_k(\nu_k)\Gamma_k(\nu_k)\big)\Delta_{k-1}^\star + \theta_k(\nu_k) \cdot \tfrac{\beta_k(\nu_k)\eta\sigma^2}{2\mu_k(\nu_k)}.$$

B.2.7 Corollary 3 (Rate-floor tradeoff via ν_k)

Corollary 3. (Rate–floor tradeoff via ν_k) Assume the setting of Theorem 2 and Corollary 2. Suppose Agent 2 chooses a constant response gain $\nu_k \equiv \nu \in (0, 1]$ and a constant inner-loop length $T_k \equiv T$. Let $\lambda(\nu)$ denote the scaled alignment constant, and suppose $\|L_i\| \leq \bar{L}$ for all $i \in [N]$. Define $\Gamma(\nu) := \frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda(\nu)}{N}$ and $\Psi(\nu) := \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)}$. Then the asymptotic noise floor is

$$\epsilon_\infty(\nu) := \frac{\Psi(\nu)}{\Gamma(\nu)} = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)}.$$

Moreover, suppose $\lambda(\nu) \geq \frac{1}{2(1+\nu^2\bar{L}^2)}\min\{1, \nu^2\lambda_L\}$, where $\lambda_L := \inf_{\|b\|=1}\max_{i\in[N]}\|L_i^\top b\|^2$, and suppose $\beta(\nu) \leq \beta(1 + \nu^2\bar{L}^2)$. Then

$$\epsilon_\infty(\nu) \lesssim \frac{N\beta^2\eta\sigma^2}{\mu_{\mathrm{PL}}\mu(\nu)} \cdot \frac{(1 + \nu^2\bar{L}^2)^2}{\min\{1, \nu^2\lambda_L\}}.$$

In particular, if $\mu(\nu)$ is uniformly lower bounded, then this upper bound scales like ν^{-2} when $\nu^2\lambda_L \leq 1$.

Proof. Under the stationary design $\nu_k \equiv \nu$ and $T_k \equiv T$, the quantities appearing in Theorem 2 are constant across epochs. In particular,

$$\theta(\nu, T) = 1 - q(\nu)^T, \qquad \Gamma(\nu) = \frac{\mu_{\mathrm{PL}}}{\beta} \frac{\lambda(\nu)}{N}.$$

By Corollary 2, the one-step inner-loop noise-floor parameter is

$$\Psi(\nu) = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)}.$$

Therefore, Theorem 2 gives the recursion

$$\Delta_k^\star \le \big(1 - \theta(\nu, T)\Gamma(\nu)\big)\Delta_{k-1}^\star + \theta(\nu, T)\Psi(\nu).$$

Define

$$\rho(\nu, T) := 1 - \theta(\nu, T)\Gamma(\nu).$$

Then

$$\Delta_k^\star \le \rho(\nu, T)\Delta_{k-1}^\star + \theta(\nu, T)\Psi(\nu).$$

Iterating this recursion yields

$$\begin{aligned}
\Delta_k^\star &\le \rho(\nu, T)^k \Delta_0^\star + \theta(\nu, T)\Psi(\nu) \sum_{j=0}^{k-1} \rho(\nu, T)^j \\
&= \rho(\nu, T)^k \Delta_0^\star + \theta(\nu, T)\Psi(\nu) \frac{1 - \rho(\nu, T)^k}{1 - \rho(\nu, T)} \\
&\le \rho(\nu, T)^k \Delta_0^\star + \frac{\theta(\nu, T)\Psi(\nu)}{1 - \rho(\nu, T)}.
\end{aligned}$$

Since

$$1 - \rho(\nu, T) = \theta(\nu, T)\Gamma(\nu),$$

the additive term simplifies to

$$\frac{\theta(\nu, T)\Psi(\nu)}{1 - \rho(\nu, T)} = \frac{\Psi(\nu)}{\Gamma(\nu)}.$$

Thus,

$$\Delta_k^\star \le \rho(\nu, T)^k \Delta_0^\star + \frac{\Psi(\nu)}{\Gamma(\nu)}.$$

This proves the finite-time bound. Taking the asymptotic term gives the noise floor

$$\epsilon_\infty(\nu) := \frac{\Psi(\nu)}{\Gamma(\nu)} = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)}.$$

It remains to derive the stated upper bound on $\epsilon_\infty(\nu)$. For each i, the response-gain policy has linear part

$$V_i(\nu) = \begin{bmatrix} I \\ \nu L_i \end{bmatrix}.$$

Since f is β-smooth, the restricted objective has smoothness bounded by

$$\beta(\nu) \le \beta \max_i \lambda_{\max}\big(V_i(\nu)^\top V_i(\nu)\big).$$

Moreover,

$$V_i(\nu)^\top V_i(\nu) = I + \nu^2 L_i^\top L_i.$$

Therefore, if $\|L_i\| \le \bar{L}$ for all $i \in [N]$, then

$$\beta(\nu) \le \beta(1 + \nu^2 \bar{L}^2).$$

Combining this with the assumed lower bound

$$\lambda(\nu) \ge \frac{1}{2(1 + \nu^2 \bar{L}^2)} \min\{1, \nu^2 \lambda_L\},$$

we obtain

$$\begin{aligned}\epsilon_\infty(\nu) &= \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)} \\ &\le \frac{\beta(1+\nu^2\bar{L}^2)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}} \cdot \frac{2(1+\nu^2\bar{L}^2)}{\min\{1,\nu^2\lambda_L\}} \\ &= \frac{N\beta^2\eta\sigma^2}{\mu_{\mathrm{PL}}\mu(\nu)} \cdot \frac{(1+\nu^2\bar{L}^2)^2}{\min\{1,\nu^2\lambda_L\}}.\end{aligned}$$

Finally, when $\nu^2\lambda_L \le 1$, the denominator satisfies

$$\min\{1,\nu^2\lambda_L\} = \nu^2\lambda_L,$$

so the upper bound scales like ν^{-2}, provided $\mu(\nu)$ remains uniformly lower bounded. This proves the claim. □

B.2.8 Corollary 4 (Optimizing epoch length and finite-time decay)

Corollary 4. (Optimizing epoch length and finite-time decay) Assume the setting of Theorem 2 and Corollary 2. Suppose Agent 2 chooses a constant response gain $\nu_k \equiv \nu \in (0,1]$ and a constant inner-loop length $T_k \equiv T$. Define $\Gamma(\nu) := \frac{\mu_{\mathrm{PL}}}{\beta}\frac{\lambda(\nu)}{N}$, $q(\nu) := 1-\eta\mu(\nu)$, and $\theta(\nu,T) := 1-q(\nu)^T$. Let $\epsilon_\infty(\nu) := \frac{\Psi(\nu)}{\Gamma(\nu)} = \frac{\beta(\nu)\eta\sigma^2}{2\mu(\nu)} \cdot \frac{\beta N}{\mu_{\mathrm{PL}}\lambda(\nu)}$ denote the corresponding asymptotic noise floor. Fix a target accuracy $\epsilon_\star > \epsilon_\infty(\nu)$ and a horizon $k \in \mathbb{N}$. Then it suffices to choose T so that

$$T \ge \frac{1}{\log q(\nu)} \log\left(1 - \frac{1}{k\Gamma(\nu)} \log\left(\frac{\Delta_0^\star}{\epsilon_\star - \epsilon_\infty(\nu)}\right)\right), \tag{6}$$

provided the logarithm is well-defined, i.e., the argument of $log(\cdot)$ lies in $(0,1)$. With this choice, the outer-loop error satisfies $\Delta_k^\star \le \epsilon_\star$.

Proof. By Corollary 3, for fixed response gain ν and fixed epoch length T, the outer-loop error satisfies

$$\Delta_k^\star \le \rho(\nu,T)^k\Delta_0^\star + \epsilon_\infty(\nu),$$

where

$$\rho(\nu,T) = 1-\theta(\nu,T)\Gamma(\nu), \qquad \theta(\nu,T) = 1-q(\nu)^T.$$

To guarantee $\Delta_k^\star \le \epsilon_\star$, it suffices to require

$$\rho(\nu,T)^k\Delta_0^\star \le \epsilon_\star - \epsilon_\infty(\nu).$$

Using $\rho(\nu,T) = 1-\theta(\nu,T)\Gamma(\nu)$ and the inequality $(1-x)^k \le \exp(-kx)$ for $x \in [0,1]$, we have

$$\rho(\nu,T)^k = \big(1-\theta(\nu,T)\Gamma(\nu)\big)^k \le \exp\big(-k\theta(\nu,T)\Gamma(\nu)\big).$$

Therefore, it is sufficient that

$$\exp\big(-k\theta(\nu,T)\Gamma(\nu)\big)\Delta_0^\star \le \epsilon_\star - \epsilon_\infty(\nu).$$

Equivalently,

$$\exp\big(-k\theta(\nu,T)\Gamma(\nu)\big) \le \frac{\epsilon_\star - \epsilon_\infty(\nu)}{\Delta_0^\star}.$$

Taking logarithms gives the sufficient condition

$$-k\theta(\nu,T)\Gamma(\nu) \le \log\left(\frac{\epsilon_\star - \epsilon_\infty(\nu)}{\Delta_0^\star}\right).$$

Equivalently,

$$\theta(\nu,T) \ge \frac{1}{k\Gamma(\nu)} \log\left(\frac{\Delta_0^\star}{\epsilon_\star - \epsilon_\infty(\nu)}\right).$$

Define

$$a_k := \frac{1}{k\Gamma(\nu)} \log\left(\frac{\Delta_0^\star}{\epsilon_\star - \epsilon_\infty(\nu)}\right).$$

By assumption, $a_k \in (0,1)$. Since

$$\theta(\nu, T) = 1 - q(\nu)^T,$$

it is sufficient that

$$1 - q(\nu)^T \geq a_k.$$

Equivalently,

$$q(\nu)^T \leq 1 - a_k.$$

Because $q(\nu) \in (0,1)$, we have $\log q(\nu) < 0$. Taking logarithms gives

$$T \log q(\nu) \leq \log(1 - a_k).$$

Dividing by the negative number $\log q(\nu)$ reverses the inequality, so

$$T \geq \frac{\log(1 - a_k)}{\log q(\nu)}.$$

Substituting the definition of a_k yields

$$T \geq \frac{1}{\log q(\nu)} \log\left(1 - \frac{1}{k\Gamma(\nu)} \log\left(\frac{\Delta_0^\star}{\epsilon_\star - \epsilon_\infty(\nu)}\right)\right).$$

Therefore, any integer T satisfying this lower bound guarantees

$$\Delta_k^\star \leq \epsilon_\star.$$

Finally, the asymptotic floor is

$$\epsilon_\infty(\nu) = \frac{\Psi(\nu)}{\Gamma(\nu)},$$

which does not depend on T. The epoch length T only affects $\theta(\nu, T) = 1 - q(\nu)^T$ and hence the finite-time contraction factor

$$\rho(\nu, T) = 1 - \theta(\nu, T)\Gamma(\nu).$$

Thus, increasing T improves the finite-time rate of convergence toward the noise floor, but does not change the floor itself. □

B.2.9 Corollary 5 (Exact linear convergence with deterministic inner loop)

Corollary 5. (Exact linear convergence with deterministic inner loop). Under Assumptions 1 and 2, suppose the inner loop satisfies Assumption 3 with $\delta_k = 0$ for all k. Then the outer-loop iterates converge linearly:

$$\mathbb{E}[f(z_k) - f^*] \leq \rho^k \big(f(z_0) - f^*\big), \tag{7}$$

where $\rho = 1 - \theta \, \frac{\mu_{\mathrm{PL}}}{\beta} \, \frac{\lambda}{N} \in (0,1)$.

Proof. Fix k and condition on $z := z_{k-1}$. Let $i := i_k$. By Assumption 3, we have that

$$f(z_k) \leq f(z) - \theta\big(f(z) - f_i^*(z)\big)$$

where we set $\theta := \min_i \theta_i$. By Assumption 1.a, we have that

$$f(z) - f_i^*(z) \geq \frac{1}{2\beta} \|P_i \nabla f(z)\|^2,$$

so that

$$f(z_k) \leq f(z) - \theta \frac{1}{2\beta} \|P_i \nabla f(z)\|^2.$$

Now subtract f^* to get that

$$f(z_k) - f^* \leq f(z) - f^* - \theta \frac{1}{2\beta} \|P_i \nabla f(z)\|^2.$$

Take the expectation over the uniform random choice of i, we have that

$$\mathbb{E}[f(z_k) - f^* \mid \mathcal{F}_{k-1}] \leq f(z) - f^* - \theta \frac{1}{2\beta} \mathbb{E}_i \|P_i \nabla f(z)\|^2.$$

Using Assumption 2, we have that

$$\mathbb{E}_i \|P_i g\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|P_i g\|^2 \geq \frac{1}{N} \max_i \|P_i g\|^2 \geq \frac{\lambda}{N} \|g\|^2.$$

Thus, the estimate holds:

$$\mathbb{E}[f(z_k) - f^* \mid \mathcal{F}_{k-1}] \leq f(z) - f^* - \theta \frac{\lambda}{2\beta N} \|\nabla f(z)\|^2.$$

Finally applying the PŁ inequality $\|\nabla f(z)\|^2 \geq 2\mu_{\mathrm{PL}}(f(z) - f^*)$, we have that

$$\mathbb{E}[f(z_k) - f^* \mid \mathcal{F}_{k-1}] \leq \left(1 - \theta \frac{\mu_{\mathrm{PL}}}{\beta} \frac{\lambda}{N}\right) (f(z) - f^*).$$

Taking full expectation and iterating over k yields the result thereby concluding the proof. □

B.3 TECHNICAL LEMMAS

B.3.1 Spanning Condition & Minimal Number of Subspaces

Lemma 6. (Spanning condition). Let $\mathcal{L} = \{L_1, \ldots, L_N\} \subset \mathbb{R}^{d_2 \times d_1}$ and define, for each i,

$$V_i := \begin{bmatrix} I_{d_1} \\ L_i \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times d_1}, \qquad S_i := \mathrm{range}(V_i) \subset \mathbb{R}^{d_1+d_2}.$$

Define the block matrix

$$Y := \begin{bmatrix} V_1 & V_2 & \cdots & V_N \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times(Nd_1)}.$$

Then

$$S_1 + S_2 + \cdots + S_N \;=\; \mathbb{R}^{d_1+d_2} \quad \Longleftrightarrow \quad \mathrm{rank}(Y) = d_1 + d_2.$$

Proof. By definition, $S_1 + \cdots + S_N$ is the set of all finite sums $s_1 + \cdots + s_N$ with $s_i \in S_i = \mathrm{range}(V_i)$. Equivalently, it is the column space of the concatenated matrix $Y = [V_1 \ \cdots \ V_N]$. Therefore $S_1 + \cdots + S_N = \mathbb{R}^{d_1+d_2}$ if and only if $\mathrm{col}(Y) = \mathbb{R}^{d_1+d_2}$, i.e., $\mathrm{rank}(Y) = d_1 + d_2$. □

Corollary 6. (Minimal number of subspace matrices). Let $\mathcal{L} = \{L_1, \ldots, L_N\} \subset \mathbb{R}^{d_2 \times d_1}$ and define

$$V_i := \begin{bmatrix} I_{d_1} \\ L_i \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times d_1}, \qquad S_i := \mathrm{range}(V_i) \subset \mathbb{R}^{d_1+d_2}.$$

If the spanning condition of Lemma 6 holds, i.e.,

$$S_1 + \cdots + S_N = \mathbb{R}^{d_1+d_2},$$

then necessarily

$$N \;\geq\; N_{\min} \;:=\; \left\lceil \frac{d_1 + d_2}{d_1} \right\rceil.$$

Conversely, this lower bound is tight: for $N = N_{\min}$, there exists a choice of $\{L_i\}_{i=1}^N$ such that the block matrix

$$Y := [\, V_1 \ V_2 \ \cdots \ V_N \,]$$

has full row rank, i.e., $\mathrm{rank}(Y) = d_1 + d_2$.

Proof. By Lemma 6, the spanning condition $S_1 + \cdots + S_N = \mathbb{R}^{d_1+d_2}$ is equivalent to $\operatorname{rank}(Y) = d_1 + d_2$, where $Y = [V_1 \ \cdots \ V_N] \in \mathbb{R}^{(d_1+d_2)\times(Nd_1)}$. Since $\operatorname{rank}(Y) \le \min\{d_1 + d_2,\ Nd_1\}$, full row rank requires

$$Nd_1 \ \ge \ d_1 + d_2,$$

which implies

$$N \ \ge \ \frac{d_1 + d_2}{d_1}.$$

Because N is an integer, this yields the necessary lower bound $N \ge \left\lceil \frac{d_1+d_2}{d_1} \right\rceil$.

To see that this bound is tight, take $N = N_{\min} := \left\lceil \frac{d_1+d_2}{d_1} \right\rceil$ and choose $\{L_i\}_{i=1}^{N}$ so that the column blocks V_i together form a basis of $\mathbb{R}^{d_1+d_2}$ (e.g., by selecting L_i so that the columns of V_i span disjoint coordinate subspaces whose union covers all coordinates). With this construction, $Y = [V_1 \ \cdots \ V_N]$ has full row rank, i.e., $\operatorname{rank}(Y) = d_1 + d_2$, completing the proof. □

C ADDITIONAL ALGORITHM VARIANTS

Here we provide schematic illustrations for both the uniform-sampling and round-robin variants in the idealized convex best-response setting. We then prove linear convergence to a noise floor for the round-robin variant of the algorithm. We also show the algorithm for "Outer-controlled drift via response gain ν_k and inner budget T_k"

C.1 ALGORITHM VARIANTS: ROUND-ROBIN VS. UNIFORM SAMPLING (CONCEPTUAL ILLUSTRATION)

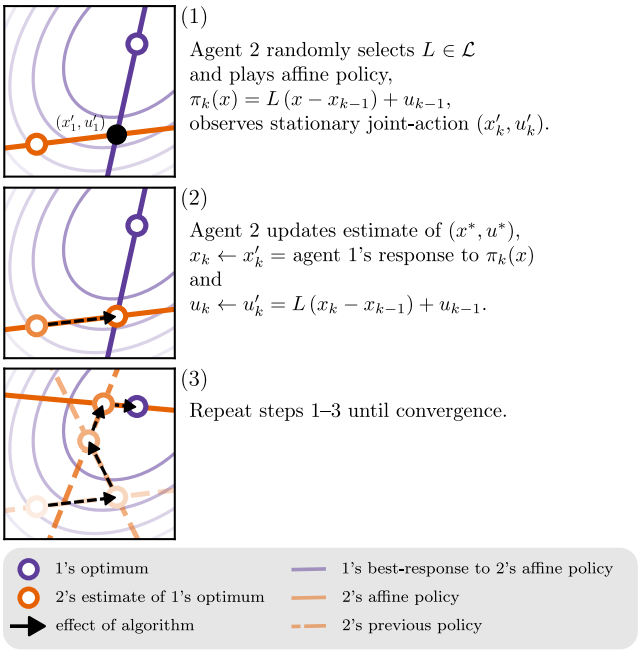

Figure 4: Uniform Sampling Outer Loop Visualization (Algorithm 1)

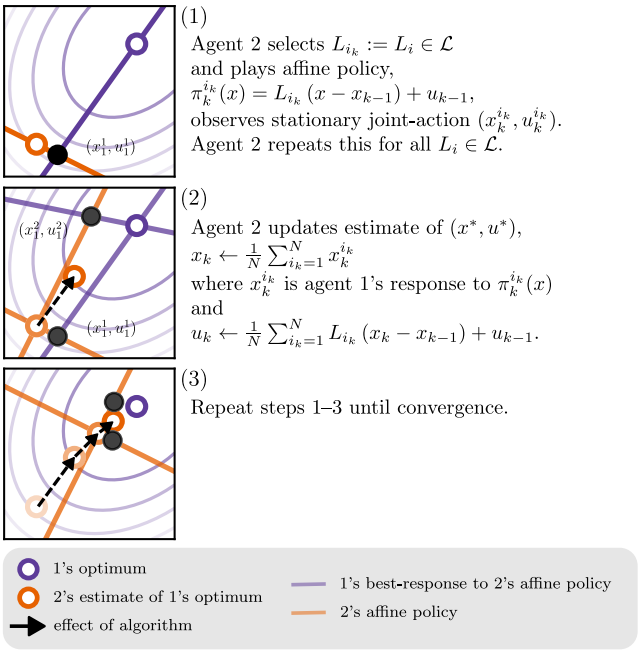

Figure 5: Round-Robin Sampling Outer Loop Visualization (Algorithm 2)

Corollary 7 (Linear convergence to a noise floor for the round-robin batch output (PŁ))**.** Let $\mathcal{L} = \{L_1, \ldots, L_N\}$. At outer iteration k, run the inner procedure once for each $L_{i_k} \in \mathcal{L}$ starting from the same initialization z_k, producing endpoints $\{z_k^{i_k}\}_{i_k=1}^N$. Define the *round-robin batch output value*

$$\bar{f}_{k+1} \;:=\; \frac{1}{N}\sum_{i_k=1}^{N} f(z_k^{i_k}). \tag{10}$$

Suppose Assumptions 1, 2, and 3 hold. Let $\theta := 1 - q^T$. Then the batch output values satisfy the linear-to-noise-floor recursion

$$\mathbb{E}\big[\bar{f}_{k+1} - f^\star\big] \;\le\; \rho_{\mathrm{RR}}\,\mathbb{E}[f(z_k) - f^\star] \;+\; \theta\,\delta, \qquad \rho_{\mathrm{RR}} := 1 - \theta\,\frac{\mu_{\mathrm{PL}}}{\beta}\,\frac{\lambda}{N} \in (0,1). \tag{11}$$

Consequently,

$$\mathbb{E}\big[\bar{f}_k - f^\star\big] \;\le\; (\rho_{\mathrm{RR}})^k\big(f(z_0) - f^\star\big) \;+\; \frac{\theta\,\delta}{1 - \rho_{\mathrm{RR}}}. \tag{12}$$

Equivalently, if $\tilde{z}_{k+1}$ is drawn uniformly from $\{z_k^{i_k}\}_{i_k=1}^N$, then $\mathbb{E}[f(\tilde{z}_{k+1})] = \mathbb{E}[\bar{f}_{k+1}]$ and the same bound holds for $\mathbb{E}[f(\tilde{z}_k) - f^\star]$.

Proof. Fix k and condition on z_k. For each $i_k \in \{1, \ldots, N\}$ define the subspace-restricted optimum

$$f_{i_k}^\star(z_k) \;:=\; \min_{y \in z_k + S_{i_k}} f(y),$$

where S_{i_k} is the subspace induced by L_{i_k} and P_{i_k} is the corresponding orthogonal projector.

Step 1 (One-epoch progress for each cycled subspace). By Assumption 3 applied to the inner run on S_{i_k},

$$\mathbb{E}\big[f(z_k^{i_k}) \,\big|\, z_k\big] \;\le\; f(z_k) \;-\; \theta\big(f(z_k) - f_{i_k}^\star(z_k)\big) \;+\; \theta\,\delta, \qquad \theta := 1 - q^T.$$

Averaging the above inequality over $i_k = 1, \dots, N$ gives

$$\mathbb{E}\big[\bar{f}_{k+1} \,\big|\, z_k\big] \;\le\; f(z_k) \;-\; \theta \cdot \frac{1}{N} \sum_{i_k=1}^{N} \big(f(z_k) - f_{i_k}^\star(z_k)\big) \;+\; \theta\,\delta. \tag{13}$$

Step 2 (Lower bound the average subspace gap). By β-smoothness in Assumption 1, for each i_k,

$$f(z_k) - f_{i_k}^\star(z_k) \;\ge\; \frac{1}{2\beta} \|P_{i_k} \nabla f(z_k)\|^2.$$

Thus

$$\frac{1}{N} \sum_{i_k=1}^{N} \big(f(z_k) - f_{i_k}^\star(z_k)\big) \;\ge\; \frac{1}{2\beta N} \max_{i_k} \|P_{i_k} \nabla f(z_k)\|^2.$$

By Assumption 2, $\max_{i_k} \|P_{i_k} g\|^2 \ge \lambda \|g\|^2$ for all g, so

$$\frac{1}{N} \sum_{i_k=1}^{N} \big(f(z_k) - f_{i_k}^\star(z_k)\big) \;\ge\; \frac{\lambda}{2\beta N} \|\nabla f(z_k)\|^2.$$

Finally, by the PŁ inequality in Assumption 1, $\|\nabla f(z_k)\|^2 \ge 2\mu_{\mathrm{PL}}\big(f(z_k) - f^\star\big)$, hence

$$\frac{1}{N} \sum_{i_k=1}^{N} \big(f(z_k) - f_{i_k}^\star(z_k)\big) \;\ge\; \frac{\mu_{\mathrm{PL}}}{\beta} \frac{\lambda}{N} \big(f(z_k) - f^\star\big). \tag{14}$$

Step 3 (Conclude the recursion). Substituting (14) into (13) yields

$$\mathbb{E}\big[\bar{f}_{k+1} - f^\star \,\big|\, z_k\big] \;\le\; \Big(1 - \theta\, \frac{\mu_{\mathrm{PL}}}{\beta} \frac{\lambda}{N}\Big)\big(f(z_k) - f^\star\big) \;+\; \theta\,\delta.$$

Taking total expectation gives (11). Iterating the recursion yields the stated linear convergence to a noise floor. □

C.2 OUTER-CONTROLLED DRIFT VIA RESPONSE GAIN ν_k AND INNER BUDGET T_k

Algorithm 3 Outer-controlled drift via response gain ν_k and inner budget T_k

Initialize (x_0, u_0).
for $k = 1, \dots, K$ **do**
  Sample $i_k \sim \mathrm{Unif}(\{1, \dots, N\})$ and set $L_k = L_{i_k}$.
  Choose $\nu_k \in (0, 1]$ and $T_k \in \mathbb{N}$.
  Define π_k by (9) and $\varphi_k(x) = f(x, \pi_k(x))$.
  Set $x_{k,0} \leftarrow x_{k-1}$.
  for $t = 0, \dots, T_k - 1$ **do**
    $x_{k,t+1} \leftarrow \mathcal{A}(x_{k,t}; L_k, u_{k-1}, x_{k-1})$.
    $u_{k,t+1} \leftarrow \pi_k(x_{k,t+1})$
  end for
  $x_k \leftarrow x_{k,T_k}, \quad u_k \leftarrow \pi_k(x_k), \quad z_k \leftarrow (x_k, u_k)$.
end for
Return: z_K.

Algorithm 3 shows how Algorithm 1 can be modified to control for noise at the inner level and trade that off with performance error. Related to Theorem 2 and Corollary 2

D SUPPLEMENTAL METHODS: HUMAN PARTICIPANT EXPERIMENTS

Human-subject experiments were conducted using both the uniform and round-robin sampling methods of the algorithm under four action-space conditions, defined by the dimensionality of the human and machine actions: $d_H \times d_M \in \{1 \times 1,\ 1 \times 2,\ 2 \times 1,\ 2 \times 2\}$. A conceptual overview of the interactive task is shown in Figure 1. At each time step t, the instantaneous cost $f(h_{k,t}, m_{k,t})$ was visualized as the radius of a circle, following validated methods from (Chasnov et al., 2025; Isa et al., 2024). Participants were instructed to "keep the circle as small as possible".

D.1 EXPERIMENTAL SETUP

At iteration k, the machine helper updates its estimate (h_k, m_k) of the human-machine optimum (h^*, m^*) by observing the human's response to the affine policy

$$\pi_k(h, L_i) = L_i(h - h_k) + m_k.$$

The cost function used in all experiments is quadratic:

$$f(h, m) = \tfrac{1}{2} h^\top R_H h + h^\top C m + \tfrac{1}{2} m^\top R_M m,$$

where $R_H \in \mathbb{R}^{d_H \times d_H}$, $C \in \mathbb{R}^{d_H \times d_M}$, and $R_M \in \mathbb{R}^{d_M \times d_M}$ are chosen to ensure convexity. Figures 4 and 5 illustrate the main computational steps of both algorithms.

Below we list the parameter values and the corresponding fixed set $\mathcal{L}$ for each action-space condition.

D.1.1 2×2 Condition

$$R_H = \begin{bmatrix} 1.7500 & 0.3536 \\ 0.3536 & 1.2500 \end{bmatrix}, \quad C = \begin{bmatrix} -0.1768 & -0.1768 \\ 0.1768 & -0.1768 \end{bmatrix}, \quad R_M = \begin{bmatrix} 1.8536 & -0.2500 \\ -0.2500 & 1.1464 \end{bmatrix}.$$

$$\mathcal{L} = \left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\}.$$

D.1.2 2×1 Condition

$$R_H = \begin{bmatrix} 1.5000 & 0.3536 \\ 0.3536 & 1.2500 \end{bmatrix}, \quad C = \begin{bmatrix} -0.3536 \\ -0.2500 \end{bmatrix}, \quad R_M = \begin{bmatrix} 1.2500 \end{bmatrix}.$$

$$\mathcal{L} = \{[1, 1], [1, -1]\}.$$

D.1.3 1×2 Condition

$$R_H = [1.7500], \quad C = \begin{bmatrix} -0.0732 & -0.4268 \end{bmatrix}, \quad R_M = \begin{bmatrix} 1.9786 & -0.1250 \\ -0.1250 & 1.2714 \end{bmatrix}.$$

$$\mathcal{L} = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}.$$

D.1.4 1×1 Condition

$$R_H = [1.5000], \quad C = [-0.5000], \quad R_M = [1.5000].$$

$$\mathcal{L} = \{[-1], [1]\}.$$

D.2 EXPERIMENT INITIALIZATIONS

For both the uniform and round-robin sampling methods of the experiments, the machine's initial estimate (h_0, m_0) was sampled uniformly at random from the level set $f(h_0, m_0) = 0.28$, ensuring a consistent starting distance from the optimum. Data were collected from 20 naive participants per condition (1×1, 1×2, 2×1, 2×2) for each algorithm, with each participant completing only one condition.

To compare with simulations, an additional 1×1 experiment tested the round-robin sampling method using eight fixed initialization points on the same level set. Eighty naïve participants (10 per point) completed this condition under the same protocol.

D.3 PARTICIPANT POPULATION

Participants were recruited through the online research platform *Prolific* (Palan and Schitter, 2018). All were naïve to the task and drawn from the standard participant pool. Each participant only interacted with one algorithm and one condition (1×1, 1×2, 2×1, 2×2). In total, 248 participants were recruited; data from 240 were retained after excluding seven for data-logging errors and one for failed attention checks.

D.4 HUMAN INPUT

Participants provided manual input via a computer mouse to continuously control action h. For conditions 1×1 and 1×2, the horizontal cursor position determined the one-dimensional input $h \in [-1, 1] \subset \mathbb{R}$, while for 2×1 and 2×2, both horizontal and vertical positions determined the two-dimensional input $h \in [-1, 1]^2 \subset \mathbb{R}^2$.

The quadratic cost function has its true minimum at $h^* = 0$ (for $d_H = 1$) or $h^* = (0, 0)$ (for $d_H = 2$). To prevent participants from memorizing this fixed location, the mapping between cursor position and action was shifted so that the perceived optimum appeared at $h = 0.25$ for $d_H = 1$ and $h = (0.25, 0.25)$ for $d_H = 2$. This preserved the underlying optimization structure while preventing learning based on spatial memory.

To further randomize the optimum's location, a mirroring effect was applied. For the 1×1 and 1×2 conditions, a random variable $r \in \{-1, +1\}$ was drawn each trial, mapping $h_{k,t} \mapsto r h_{k,t}$. For the 2×1 and 2×2 conditions, independent random variables $r_j \in \{-1, +1\}$ were applied to each dimension, i.e., $h_{k,t,j} \mapsto r_j h_{k,t,j}$ for $j \in \{1, 2\}$. This ensured the optimum appeared in different halves or quadrants of the screen, eliminating consistent spatial biases.

D.5 PROTOCOL

Experiments using the uniform sampling method of the algorithm followed Algorithm 1, and those using the round-robin sampling method of the algorithm followed Algorithm 2. Participants repeated the game for 14 iterations ($K = 14$).

For the round-robin sampling method of the algorithm, each iteration contained 2 trials for the 1×1, 2×1, and 2×2 conditions, and 3 trials for the 1×2 condition, corresponding to the minimal set $\mathcal{L}$ satisfying the Spanning Condition (Lemma 6). For the uniform sampling method of the algorithm, each iteration contained 1 trial, with L_i selected uniformly at random from $\mathcal{L}$.

Trial durations were 15 seconds for 1×1 and 1×2, and 25 seconds for 2×1 and 2×2; the mean action $(h_{k,t}, m_{k,t})$ of the last 5 seconds was used as the trial outcome. Longer durations were used for 2-dimensional conditions due to increased difficulty.

Each experiment included attention check trials at the beginning (before iteration 1), middle (after iteration 7), and end (after iteration 14). These trials were identical in duration, controls, and instructions to normal trials, with optimum actions placed randomly at ± 0.25 for 1×1 and 1×2, and $(\pm 0.25, \pm 0.25)$ for 2×1 and 2×2. During these trials, the machine played constant action $m = [0]$ (or $[0, 0]$). Participants who achieved the mean action within an error of 0.25 proceeded; otherwise, the trial was repeated, with a maximum of 5 attempts before being screened out.

D.6 IMPLEMENTATION AND DATA COLLECTION

Experiments were implemented in JavaScript, HTML, and CSS. Each experimental condition had a dedicated HTML file. A JavaScript loop handled trial execution and data logging. At each time step (60 Hz), the system recorded time t, human input $h_{k,t}$, machine input $m_{k,t}$, instantaneous cost $f(h_{k,t}, m_{k,t})$, machine estimates (h_k, m_k), and $f(h_k, m_k)$. Trial data were serialized into JSON and transmitted to a FastAPI server hosted on Replit, where they were stored in a SQLite database indexed by unique Prolific participant IDs.

D.7 DATA PROCESSING

Raw JSON logs were converted to CSV format for analysis. Rows correspond to time steps and columns to recorded variables (e.g., participant ID, t, $h_{k,t}[0]$, $h_{k,t}[1]$, $m_{k,t}[0]$, $m_{k,t}[1]$). For each trial, only the final 5 seconds of data were retained. Data were processed and visualized using Jupyter notebooks to generate all figures reported in the paper.

D.8 CODE AND DATA AVAILABILITY

Game code for all human participant experimental conditions is included in the CodeOcean Reproducible Capsule under `/code/Human Participant Experiment/Game Code`. The README in `/code/Human Participant Experiment/Game Code/README.md` provides instructions for running the experiments. The data-processing, analysis, and figure-generation scripts are included under `/code/Human Participant Experiment/Data Analysis Code`. The Reproducible Run executes all Python notebooks in this folder and saves the generated outputs to the results directory. The CodeOcean Capsule is available at https://doi.org/10.24433/CO.8589291.v1.

E SUPPLEMENTAL METHODS: CLOSED-LOOP TWO-AGENT BALANCING CART-POLE GAME

Overview. We report implementation details for the closed-loop two-agent cart-pole balancing task with a spring-coupled helper cart. Agent 1 controls the primary cart and has access to the quadratic regulation objective, while Agent 2 is cost-blind and applies the proposed affine-response method using only observed changes in Agent 1's policy parameters. The experiment evaluates whether the helper can improve closed-loop performance when Agent 1 adapts in policy space.

We compare two Agent 1 inner-loop solvers within the same outer alignment loop: (i) a deterministic gain-space optimizer based on L-BFGS-B (BR condition), and (ii) a stochastic gain-space PPO optimizer (PPO condition) implemented with Stable-Baselines3 (Raffin et al., 2021). The environment is a custom Gymnasium-compatible simulator (Towers et al., 2024) that extends classic cart-pole dynamics with a spring–damper-coupled helper cart. All rollouts used for reporting are deterministic simulations from a fixed initial state; PPO is used only inside Agent 1's inner loop.

Notation. Let $s_t \in \mathbb{R}^6$ denote the environment state at time t. Agent 1's scalar control is $x_{k,t}$ and Agent 2's scalar control is $u_{k,t}$ at time t of outer iteration k. Agent 1's policy parameter is a feedback gain $h_k \in \mathbb{R}^6$; Agent 2's accumulated helper gain is $m_k \in \mathbb{R}^6$. We write

$$\mathrm{clip}_{F_{\max}}(z) := \min\{F_{\max}, \max\{-F_{\max}, z\}\},$$

with $F_{\max} = 10$ for both agents. Given anchors (h_{k-1}, m_{k-1}) and sampled matrix $L_k \in \mathbb{R}^{6\times 6}$, the helper's induced gain during Agent 1's inner solve is

$$\pi_k(h) = m_{k-1} + \nu L_k\,(h - h_{k-1}),$$

and the corresponding controls are

$$x_t(h) = \mathrm{clip}_{F_{\max}}(-h^\top s_t), \qquad u_t(h) = \mathrm{clip}_{F_{\max}}(-\pi_k(h)^\top s_t).$$

After Agent 1 commits h_k and Agent 2 updates m_k, the evaluated closed-loop pair is

$$x_{k,t} = \mathrm{clip}_{F_{\max}}(-h_k^\top s_t), \qquad u_{k,t} = \mathrm{clip}_{F_{\max}}(-m_k^\top s_t),$$

since $m_k = m_{k-1} + \nu L_k(h_k - h_{k-1})$.

E.1 ENVIRONMENT AND DYNAMICS

The environment extends the classic cart-pole system by adding a second cart attached to the primary cart through a linear spring–damper coupling. The state is

$$s_t = (p_{1,t}, \dot{p}_{1,t}, \theta_t, \dot{\theta}_t, p_{2,t}, \dot{p}_{2,t}) \in \mathbb{R}^6,$$

where $(p_{1,t}, \dot{p}_{1,t})$ are the primary-cart position and velocity, $(p_{2,t}, \dot{p}_{2,t})$ are the helper-cart position and velocity, and $(\theta_t, \dot{\theta}_t)$ are the pole angle and angular velocity.

The carts interact through

$$F_{\text{spring},t} = -k_s(p_{1,t} - p_{2,t}) - c_s(\dot{p}_{1,t} - \dot{p}_{2,t}),$$

with $k_s = 20$ and $c_s = 2$ unless otherwise noted. Agent 1 applies $x_{k,t}$ to the primary cart–pole system and Agent 2 applies $u_{k,t}$ to the helper cart, so the net forces are $x_{k,t} + F_{\text{spring},t}$ on the primary system and $u_{k,t} - F_{\text{spring},t}$ on the helper cart.

Dynamics follow the standard Gymnasium cart-pole equations (gravity, cart and pole masses, half-pole length $0.5\,\text{m}$) with Euler integration and time step $\tau = 0.02$. Episodes terminate when $|p_{1,t}| > 2.4\,\text{m}$ or $|\theta_t| > 12^\circ$, or after $T = 500$ steps. Initial states are sampled as in Gymnasium: each component of s_0 is drawn independently from $\text{Unif}([-0.05, 0.05])$ using the seed-specific random number generator.

E.2 COST FUNCTION

Agent 1 minimizes a finite-horizon quadratic regulation objective. The stage cost at the *start* of step t is

$$\ell(s_t, x_{k,t}, u_{k,t}) = s_t^\top Q s_t + r_1 x_{k,t}^2 + r_2 u_{k,t}^2, \tag{15}$$

and the rollout objective at outer iteration k is

$$f_k = \sum_{t=0}^{n_k - 1} \ell(s_t, x_{k,t}, u_{k,t}), \tag{16}$$

where $n_k \le T$ is the number of simulated steps before termination. We also write $f(h, m)$ for the cost of simulating gains (h, m) from fixed s_0.

Reported simulations use

$$T = 500, \qquad r_1 = 0.5, \qquad r_2 = 0.25,$$

and

$$Q = \text{diag}(0.1,\ 0.01,\ 10.0,\ 0.1,\ 0.05,\ 0.05).$$

Controls are clipped to $\pm F_{\max}$ with $F_{\max} = 10$.

E.3 POLICY PARAMETERIZATION AND AFFINE HELPER RESPONSE

Both agents are represented by linear state-feedback gains. Agent 1 uses $h \in \mathbb{R}^6$ and Agent 2 accumulates $m \in \mathbb{R}^6$. The outer algorithm therefore acts in policy-parameter space rather than perturbing raw controls directly.

At outer iteration k, Agent 2 samples

$$L_k \sim \text{Unif}(\mathcal{L}), \qquad \mathcal{L} = \{-I_6,\ +I_6\},$$

and commits to the affine gain-space response in (17)–(18):

$$\pi_k(h) = m_{k-1} + \nu L_k(h - h_{k-1}), \tag{17}$$

$$m_k = m_{k-1} + \nu L_k(h_k - h_{k-1}). \tag{18}$$

Agent 2 is cost-blind: it observes only the change $h_k - h_{k-1}$, not f, its gradients, or Agent 1's optimizer.

Initialization. For each seed, we sample s_0, linearize the nonlinear dynamics at s_0 with both controls set to zero, and solve a one-player discrete-time LQR problem (algebraic Riccati equation) with state weight Q and control penalty $R = 0.5$ on Agent 1 only. This yields a stabilizing gain $H_{1\mathrm{p}}$. The two-agent experiment is initialized as

$$h_0 = H_{1\mathrm{p}}, \qquad m_0 = 0.$$

The rollout at $k = 0$ disables Agent 2 ($u_{0,t} \equiv 0$) and therefore matches the Agent 1-only baseline. Note that $R = 0.5$ is used only to compute $H_{1\mathrm{p}}$; reported rollout costs use penalties $r_1 = 0.5$ and $r_2 = 0.25$ on the simulated controls.

E.4 OUTER-LOOP PROTOCOL

For each random seed:

1. Sample s_0 and compute $H_{1\mathrm{p}}$ as above.
2. Record the Agent 1-only baseline rollout with $h_0 = H_{1\mathrm{p}}$, $m_0 = 0$, giving cost f_0 and baseline effort $\|x_{0,\cdot}\|_1$.
3. For $k = 1, \ldots, K$:
   (a) Set anchors (h_{k-1}, m_{k-1}).
   (b) Sample L_k (or read it from a fixed per-seed sequence in ablations).
   (c) Fix the helper response $\pi_k(h) = m_{k-1} + \nu L_k(h - h_{k-1})$.
   (d) Compute Agent 1's approximate best-response gain h_k (L-BFGS-B or PPO).
   (e) Update $m_k = m_{k-1} + \nu L_k(h_k - h_{k-1})$.
   (f) Evaluate (h_k, m_k) by a deterministic rollout from the same fixed s_0.

Each reported trajectory contains $K+1$ cost values, indexed $k = 0, \ldots, K$, where $k = 0$ is the baseline.

In the main comparison experiment we use $K = 15$ for L-BFGS-B and $K = 150$ for PPO. The main L-BFGS-B branch uses $\nu = 1$; the main PPO branch uses $\nu = 0.5$. All other environment and cost parameters are shared across seeds and methods.

E.5 AGENT 1 INNER-LOOP SOLVERS

L-BFGS-B gain optimization ("BR" condition). Agent 1 is restricted to policies $x_t = -h^\top s_t$ with $h \in \mathbb{R}^6$. Given anchors (h_{k-1}, m_{k-1}), sampled L_k, and response gain ν, Agent 1 solves

$$h_k \approx \arg\min_h f(h,\ \pi_k(h))\,, \tag{19}$$

where $\pi_k(h) = m_{k-1} + \nu L_k(h - h_{k-1})$ and f is the exact deterministic rollout cost from fixed s_0. We use `scipy.optimize.minimize` with method `L-BFGS-B`, box constraints $h_i \in [-50, 50]$, initial point h_{k-1}, and at most 100 iterations per outer step.

This branch is initialized from the one-player LQR gain but, after $k = 0$, each update is a nonlinear finite-horizon gain optimization that accounts for clipping, helper coupling, and the full nonlinear dynamics.

Gain-space PPO. Agent 1's stochastic inner loop also optimizes over $\mathbb{R}^6$, but approximates (19) with PPO (Raffin et al., 2021). A fresh PPO agent is trained at every outer iteration.

Constant-gain policy. We use a custom `ConstantGainPolicy` whose actor outputs a single gain vector in $\mathbb{R}^6$ independent of the observation. With residual parameterization, PPO outputs $a \in [-1, 1]^6$ and the evaluated gain is

$$h = \mathrm{clip}_{50}(h_{k-1} + \alpha_{\mathrm{gain}}\, a)\,,$$

where $\alpha_{\mathrm{gain}} = 2.0$ and clip_{50} clips each component to $[-50, 50]$. The policy mean is warm-started to $a = 0$ (so $h \approx h_{k-1}$), with initial log-standard deviation -0.75.

Oneshot inner rollouts. PPO training uses a wrapper with three key features:

1. **Fixed initial state:** every training rollout starts from the same s_0 used for evaluation.

2. **Episode-fixed gain:** one candidate gain is selected and held fixed for the entire rollout.
3. **Oneshot episodes:** each PPO environment step executes one full physics rollout (up to $T = 500$ steps) and returns a single episodic reward

$$r^{\mathrm{PPO}} = -\frac{f(h)}{f_0},$$

where $f(h)$ is the total rollout cost under h and the helper response $\pi_k(h)$.

Thus one PPO environment step corresponds to one complete closed-loop simulation, not one physics time step.

Commit rule. After T_{PPO} training environment steps, Agent 1 commits the gain extracted from the *final* PPO model (deterministic mean action). No evaluation checkpoints or alternative candidate selection are used.

Inner training budget. In the main PPO experiment, $T_{\mathrm{PPO}} = 150$ environment steps per outer iteration. With `n_steps`$= 10$, this yields roughly 15 PPO policy updates per outer step.

E.6 BASELINE: AGENT 1-ONLY CONTROL

Agent 2 is disabled ($u_t \equiv 0$). Agent 1 uses the seed-specific one-player gain $H_{1\mathrm{p}}$ computed at s_0. The resulting rollout cost is f_0 and is used to normalize reported costs f_k/f_0 and PPO rewards. Agent 1 effort $\|x_{0,\cdot}\|_1$ is used to normalize effort traces.

E.7 ABLATION EXPERIMENTS

Corollary 3: response-gain sweep. We fix $T_{\mathrm{PPO}} = 150$ and $K = 150$, vary $\nu \in \{0.5, 0.75, 1.0\}$, and run gain-space PPO with all other settings unchanged.

Corollary 4: inner-budget sweep. We fix $\nu = 0.5$ and $K = 150$, and vary $T_{\mathrm{PPO}} \in \{90, 150, 210\}$.

Paired comparison across ablation conditions. For each seed, the ablations reuse the same s_0, $H_{1\mathrm{p}}$, baseline cost, baseline effort, and a *fixed* helper-response sequence $\{L_k\}_{k=1}^{K}$ generated deterministically from seed $1000 +$ seed. Different ablation settings therefore differ only in the swept parameter (ν or T_{PPO}). The main L-BFGS-B vs. PPO comparison, by contrast, samples L_k independently at each outer iteration.

These ablations illustrate the rate–floor and inner-budget tradeoffs suggested by theory. Because the cart-pole system is nonlinear and PPO is only an approximate inner solver, the results should not be read as direct verification of theoretical constants.

E.8 EVALUATION METRICS

We report, for each outer iteration k:

1. normalized cost f_k/f_0;
2. Agent 1 and Agent 2 effort,

$$\|x_{k,\cdot}\|_1 = \sum_t |x_{k,t}|, \qquad \|u_{k,\cdot}\|_1 = \sum_t |u_{k,t}|,$$

and effort normalized by the baseline $\|x_{0,\cdot}\|_1$;

3. control time series at the first and final outer iterations; and
4. rollout length n_k, with $n_k = T$ indicating that the pole remained within bounds for the full horizon (pole remained upright for full $T = 500$ episode).

Multi-seed plots show the mean across seeds; shaded bands denote the interquartile range across seeds.

E.9 HYPERPARAMETERS AND REPRODUCIBILITY

Shared environment and cost parameters.

$$\tau = 0.02, \quad T = 500, \quad r_1 = 0.5, \quad r_2 = 0.25, \quad k_s = 20, \quad c_s = 2, \quad F_{\max} = 10.$$

$$Q = \mathrm{diag}(0.1,\ 0.01,\ 10.0,\ 0.1,\ 0.05,\ 0.05), \qquad R = 0.5 \ \text{(LQR initialization only)}.$$

Main experiment budgets (Figure 2).

Setting	L-BFGS-B	PPO
Outer iterations K	15	150
Response gain ν	1.0	0.5
Inner budget per outer step	100 L-BFGS-B iterations	$T_{\mathrm{PPO}} = 150$ oneshot env steps

PPO hyperparameters (all PPO experiments). Learning rate 5×10^{-2}; clipping parameter 0.2; `n_steps`$= 10$; batch size 10; 5 epochs per update; $\gamma = 1.0$; GAE $\lambda = 0.95$; entropy coefficient 0; `target_kl`$=$ `None`; residual scale $\alpha_{\mathrm{gain}} = 2.0$; log-std initialization -0.75.

Seeds. All multi-seed results use

$$\{42, 43, 44, 45, 46, 47, 48, 49, 50, 51\}.$$

For each seed, s_0, $H_{1\mathrm{p}}$, f_0, and baseline effort are fixed and shared across compared methods.

E.10 CODE AND DATA AVAILABILITY

Simulation code for all the two-agent cart-pole experiments is included in the CodeOcean Reproducible Capsule under `/code/Closed-Loop Two-Agent Cart-Pole Simulation/notebooks/cartpole.ipynb`. Because of the length of the runtime of the cartpole simulations, the Reproducible Run does not execute the Python notebook, reference `/data/figsoutput` to look at figure outputs. The CodeOcean Capsule is available at https://doi.org/10.24433/CO.8589291.v1.

F ADDITIONAL FIGURES: HUMAN PARTICIPANT EXPERIMENTS

F.1 EXPERIMENTS USING THE ROUND-ROBIN SAMPLING METHOD OF THE ALGORITHM WITH RANDOM INITIALIZATION

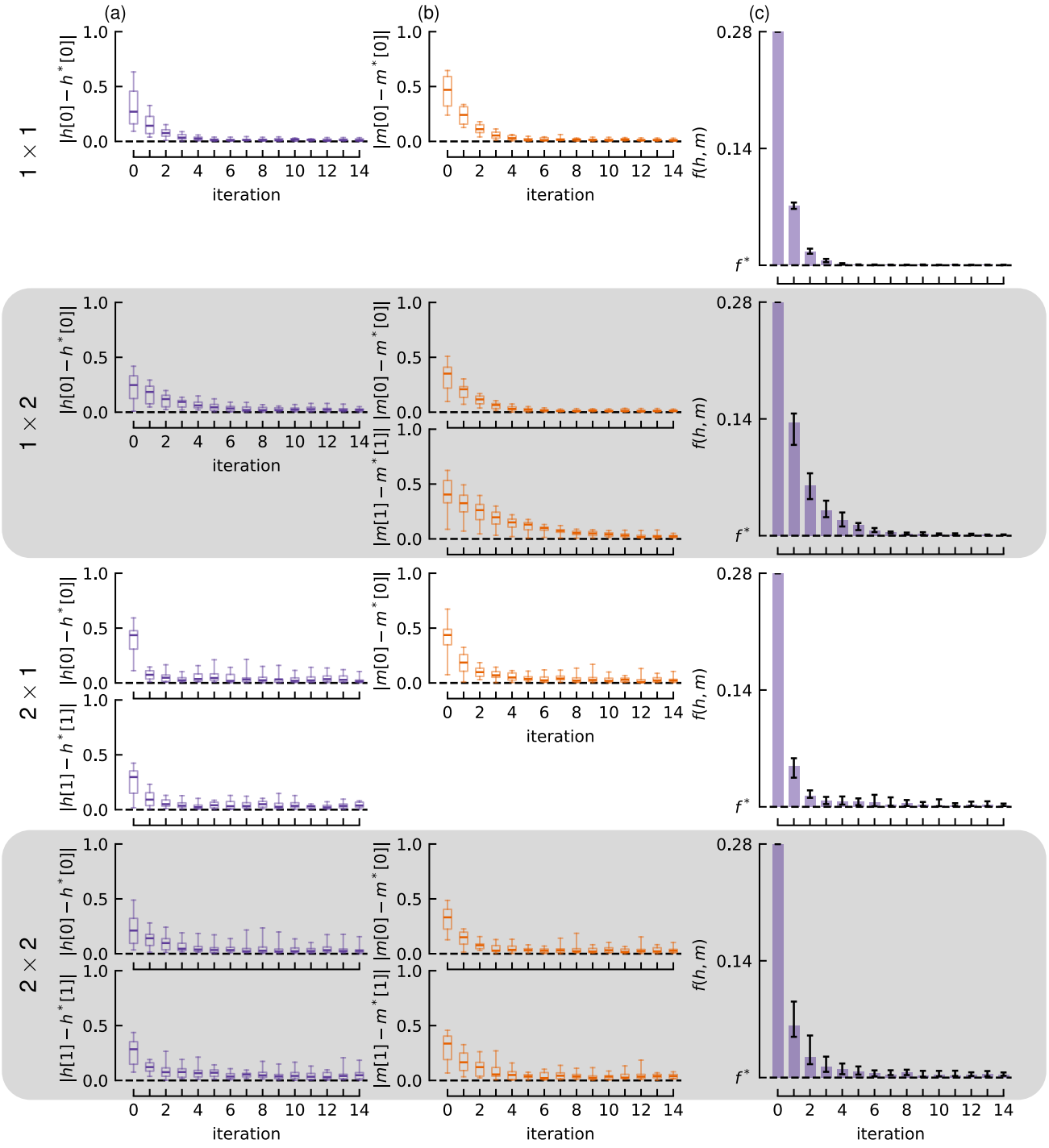

Figure 6: 1×1, 1×2, 2×1, 2×2 Round-Robin Sampling Experiments ($n = 20$ each): (a) Distributions of ℓ_1 error of M's estimate of h^*; box-and-whiskers plot showing 5th, 25th, 50th, 75th, and 95th percentiles. (b) Distributions of ℓ_1 error of M's estimate of m^*; box-and-whiskers plot showing same as (a). (c) Cost distributions; bar plots with quartiles.

F.2 EXPERIMENTS USING THE UNIFORM SAMPLING METHOD OF THE ALGORITHM WITH RANDOM INITIALIZATION

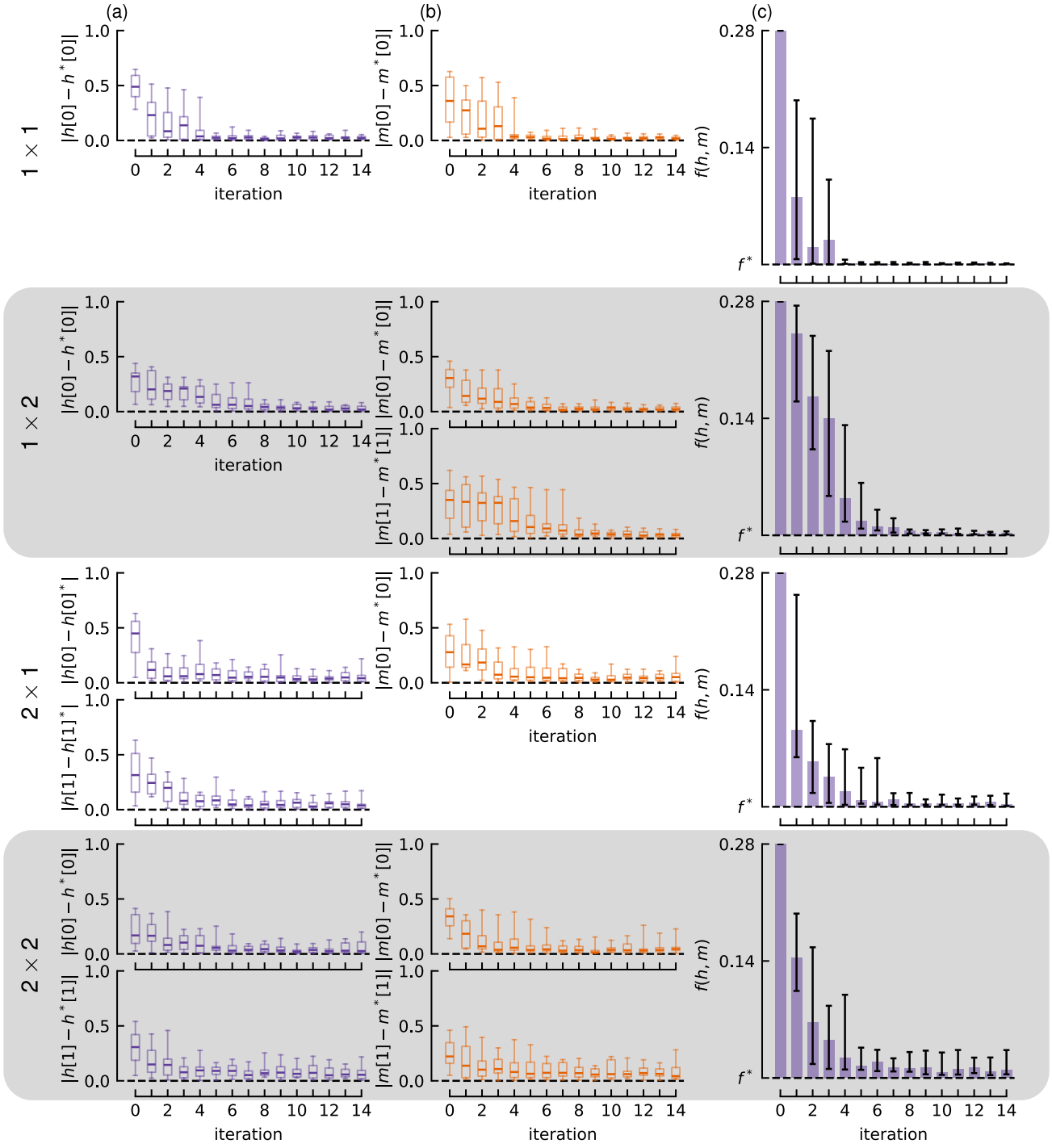

Figure 7: 1×1, 1×2, 2×1, 2×2 Uniform Sampling Experiments ($n = 20$ each): (a) Distributions of ℓ_1 error of M's estimate of h^*; box-and-whiskers plot showing 5th, 25th, 50th, 75th, and 95th percentiles. (b) Distributions of ℓ_1 error of M's estimate of m^*; box-and-whiskers plot showing same as (a). (c) Cost distributions; bar plots with quartiles.

F.3 1×1 EXPERIMENTS USING ROUND-ROBIN SAMPLING METHOD OF THE ALGORITHM WITH 8 FIXED POINT INITIALIZATIONS COMPARED TO SIMULATED RESULTS

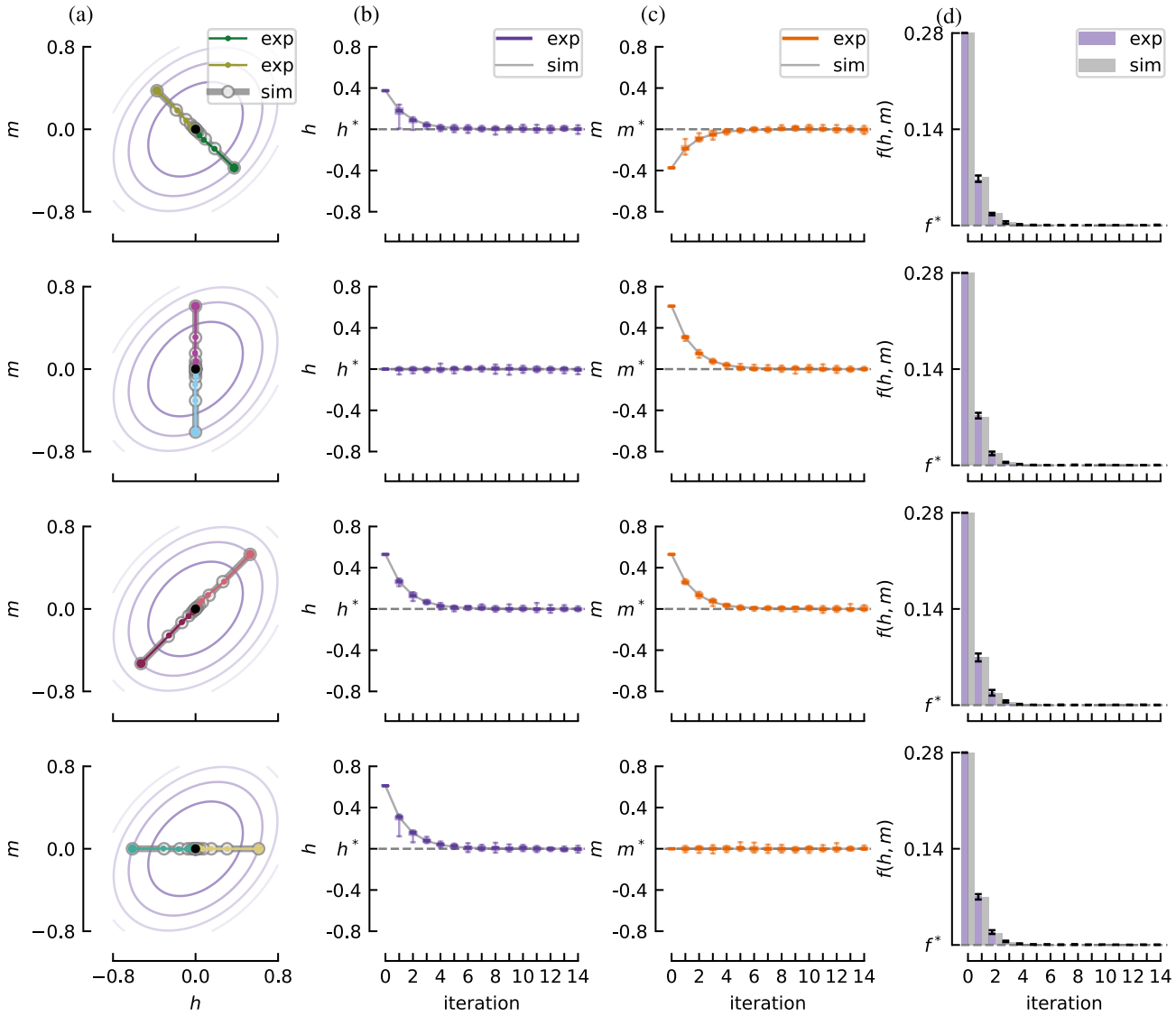

Figure 8: 1×1 Simulation vs Experiment ($n = 80$; $n = 10$ per initialization point): Grey lines correspond to simulation data. (a) M's median estimate of h^* and m^* over iterations for each initialization point. (b) Distributions of M's estimate of h^*; box-and-whiskers plot showing 5th, 25th, 50th, 75th, and 95th percentiles. (c) Distributions of M's estimate of m^*; box-and-whiskers plot showing same as (b). (d) Cost distributions; bar plots with quartiles.

G ADDITIONAL EXPERIMENT: OPEN-LOOP NON-MINIMUM PHASE LQ GAME

G.1 SUPPLEMENTAL METHODS: OPEN-LOOP LQ GAME WITH NON-MINIMUM PHASE PLANT

Overview. We report implementation details for the open-loop two-agent linear–quadratic (LQ) game with a non-minimum-phase (NMP) plant. Agent 1 has access to the objective and computes exact finite-horizon best responses by least squares. Agent 2 is cost-blind and updates via the proposed affine-response rule using only observed changes in Agent 1's open-loop input sequence. This experiment is fully open loop: at each outer iteration both agents commit entire horizon-T input sequences before simulating the plant.

Notation. Let $s_t \in \mathbb{R}^2$ denote the plant state at time t. At outer iteration k, Agent 1 commits the open-loop sequence $x_k = [x_{k,0}, \dots, x_{k,T-1}]^\top \in \mathbb{R}^T$ and Agent 2 commits $u_k = [u_{k,0}, \dots, u_{k,T-1}]^\top \in \mathbb{R}^T$. The total discrete input applied to the plant at time t is $x_{k,t} + u_{k,t}$. The output sequence is $y_k = [y_{k,0}, \dots, y_{k,T-1}]^\top \in \mathbb{R}^T$. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^T$; reported effort traces use the ℓ_1 aggregate $\|x_k\|_1 := \sum_{t=0}^{T-1} |x_{k,t}|$ (and analogously for u_k).

G.2 PLANT MODEL AND DISCRETIZATION

We consider the continuous-time LTI system

$$\dot{s}(t) = A_c s(t) + B_c v(t), \qquad y(t) = C s(t), \tag{20}$$

with scalar input $v(t)$ and scalar output $y(t) \in \mathbb{R}$. The plant is non-minimum phase with transfer function

$$G(s) = \frac{-2s + 4.4}{s^2 + 3.6s + 4}, \tag{21}$$

which has a right-half-plane zero at $s = 2.2$. The continuous-time realization used in code is

$$A_c = \begin{bmatrix} 0 & 1 \\ -4 & -3.6 \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 4.4 & -2 \end{bmatrix}. \tag{22}$$

The plant is discretized by zero-order hold with step size Δt using `scipy.signal.cont2discrete`, yielding (A_d, B_d).

In the two-agent experiment, the discrete input at time t is $v_t = x_{k,t} + u_{k,t}$, so the state evolves as $s_{t+1} = A_d s_t + B_d(x_{k,t} + u_{k,t})$.

G.3 LIFTED OPEN-LOOP MODEL

For a fixed initial state s_0, open-loop sequences $x_k, u_k \in \mathbb{R}^T$ produce the output sequence

$$y_k = O s_0 + G(x_k + u_k), \tag{23}$$

where $O \in \mathbb{R}^{T \times 2}$ and $G \in \mathbb{R}^{T \times T}$ are constructed from (A_d, B_d, C) and horizon T by simulating the impulse response along the horizon. Equation (23) is the batch (lifted) representation used for all best-response algebra.

G.4 OBJECTIVE

Agent 1 minimizes the finite-horizon quadratic cost

$$f(x_k, u_k) = q\|y_k\|_2^2 + r_1\|x_k\|_2^2 + r_2\|u_k\|_2^2, \tag{24}$$

with scalar weights $q, r_1, r_2 > 0$. Using (23), f is a jointly convex quadratic in (x_k, u_k) for fixed s_0.

In the reported simulations,

$$\Delta t = 0.1, \qquad T = 30, \qquad q = 1.0, \qquad r_1 = 0.5, \qquad r_2 = 0.25.$$

G.5 AFFINE HELPER RESPONSE

At outer iteration k, after observing Agent 1's previous sequence x_{k-1}, Agent 2 samples

$$L_k \sim \mathrm{Unif}(\mathcal{L}), \qquad \mathcal{L} = \{0_{T\times T},\, I_T\},$$

and commits to the affine open-loop response

$$u(x) = u_{k-1} + L_k(x - x_{k-1}), \tag{25}$$

defined for any candidate Agent 1 sequence $x \in \mathbb{R}^T$. Agent 2 then updates according to

$$u_k = u_{k-1} + L_k(x_k - x_{k-1}) \tag{26}$$

after Agent 1 commits x_k. Agent 2 does not observe f, (q, r_1, r_2), or Agent 1's solver. Unlike the cart-pole experiments, no response-gain parameter ν is used here; the helper applies the full affine step.

G.6 INITIALIZATION AND BASELINE

For each random seed:

1. Sample $s_0 \sim \mathcal{N}(0, I_2)$.
2. Construct (O, G) from (A_d, B_d, C) and horizon T.
3. Compute the Agent 1-only benchmark with $u \equiv 0$:

$$x^\star = \arg\min_{x \in \mathbb{R}^T} q\|Os_0 + Gx\|_2^2 + r_1\|x\|_2^2, \qquad f_{\min} = f(x^\star, 0). \tag{27}$$

   This is solved in closed form by least squares (`finite_horizon_optimum`).

4. Initialize the two-agent run at outer iteration $k = 0$ by

$$x_0 = x^\star, \qquad u_0 = 0.$$

Thus $f_0 = f_{\min}$ and the initial cooperative rollout matches the Agent 1-only baseline.

G.7 OUTER-LOOP PROTOCOL

Each run performs K outer updates and records $K+1$ cost values indexed $k = 0, \ldots, K$. For $k = 0, \ldots, K-1$:

1. Set anchors (x_k, u_k) from the previous record (at $k = 0$ these are $(x^\star, 0)$).
2. Sample L_k independently and uniformly from $\mathcal{L}$.
3. Agent 1 computes an exact best response x_{k+1} to the induced objective

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^T} f(x,\ u_k + L_k(x - x_k))\,, \tag{28}$$

   equivalently,

$$q\big\|Os_0 + G\big(x + u_k + L_k(x - x_k)\big)\big\|_2^2 + r_1\|x\|_2^2 + r_2\|u_k + L_k(x - x_k)\|_2^2.$$

4. Agent 2 updates $u_{k+1} = u_k + L_k(x_{k+1} - x_k)$.
5. Record the realized cooperative cost $f_{k+1} = f(x_{k+1}, u_{k+1})$.

Reported costs use the committed sequences directly; no helper map is applied at evaluation time.

The reported experiment uses $K = 150$.

G.8 AGENT 1 BEST RESPONSE (LEAST SQUARES)

Substituting (25) into (23)–(24) shows that, for fixed anchors (x_k, u_k) and sampled L_k, the objective is quadratic in x. Agent 1 therefore solves (28) by the normal equations

$$\big(qY^\top Y + r_1 I + r_2 L_k^\top L_k\big)x = -\big(qY^\top y_0 + r_2 L_k^\top b_k\big),$$

where

$$b_k = u_k - L_k x_k, \qquad y_0 = Os_0 + Gb_k, \qquad Y = G(I + L_k).$$

This closed-form solve is implemented in `agent_1_lqr_best_response`. Each outer step is deterministic given (s_0, x_k, u_k, L_k).

G.9 EVALUATION METRICS

For each outer iteration k we report:

1. normalized cost $f_k/f_{\min}$ (per seed);
2. Agent 1 and Agent 2 effort $\|x_k\|_1$, $\|u_k\|_1$, normalized by the seed-specific baseline $\|x^\star\|_1$;
3. open-loop input trajectories $\{x_{k,t}\}$, $\{u_{k,t}\}$, and output trajectories $y_{k,t} = (Os_0 + G(x_k + u_k))_t$ at the initial and final outer iterations; and
4. the Agent 1-only benchmark trajectories $x^\star$ and $y^\star = Os_0 + Gx^\star$.

Multi-seed plots show the mean across seeds with shaded interquartile bands (25th–75th percentiles). The time-series panels use seed 0 as a representative example.

G.10 HYPERPARAMETERS AND REPRODUCIBILITY

Shared parameters.

$$\Delta t = 0.1, \quad T = 30, \quad K = 150, \quad q = 1.0, \quad r_1 = 0.5, \quad r_2 = 0.25.$$

$$\mathcal{L} = \{0_{T\times T}, I_T\}, \qquad \text{initialization: } x_0 = x^\star,\ u_0 = 0.$$

Seeds. The reported multi-seed results use

$$\{0, 1, \ldots, 19\}.$$

For each seed, s_0, $x^\star$, $f_{\min}$, and baseline effort $\|x^\star\|_1$ are fixed for that run. Helper matrices L_k are sampled independently at each outer iteration using the seed-specific random number generator.

G.11 CODE AND DATA AVAILABILITY

Simulation code for the open-loop non-minimum phase LQ game is included in the CodeOcean Reproducible Capsule under `/code/Open-Loop Non-Minimum Phase LQ Simulation/LQGame.ipynb`. The Reproducible Run executes this notebook and saves generated outputs to the results directory. The CodeOcean Capsule is available at https://doi.org/10.24433/CO.8589291.v1.

G.12 OPEN-LOOP NON-MINIMUM PHASE LQ GAME: RESULTS

Performance improves in non-minimum phase LQ game. Figure 9 demonstrates that the proposed uniform sampling method of the algorithm consistently improves performance in the non-minimum phase LQ game relative to the single-agent baseline. In these experiments, Agent 1 is the optimizing agent and solves a finite-horizon LQR problem, while Agent 2 is a cost-blind helper that updates using the proposed uniform sampling method of the algorithm. Starting from the same state initialization, the finite-horizon cost decreases monotonically over outer iterations and converges to a value strictly below that achieved when Agent 1 acts alone. This improvement arises through a redistribution of control effort: as iterations progress, Agent 1's control energy decreases while Agent 2's control energy increases, indicating that the helper progressively assumes a larger share of the actuation burden in a manner that is beneficial to Agent 1's objective. By the end of the time horizon, the combined input $x + u$ yields output trajectories similar to the single-agent baseline, while redistributing control effort from Agent 1 to the helper. These results show that action-only interaction enables a cost-blind helper to meaningfully assist an objective-aware LQR controller.

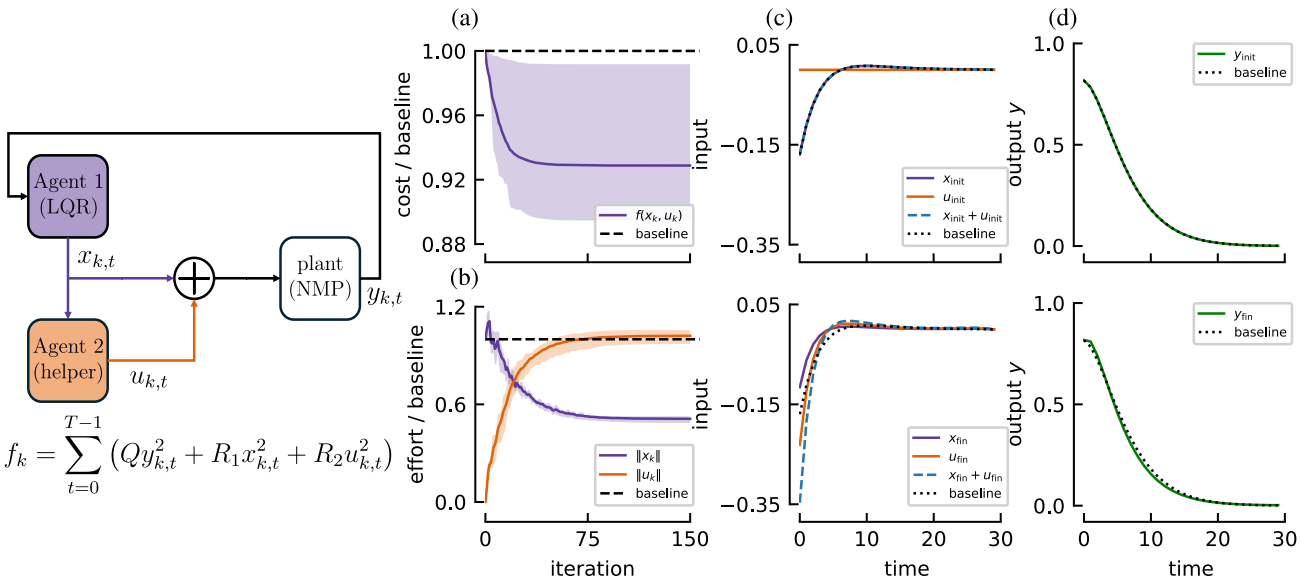

Figure 9: Open-Loop Non-minimum Phase LQ game (20 seeds): Agent 1 uses finite-horizon LQR; Agent 2 updates using the proposed uniform sampling method of the algorithm. Black curves: Agent 1-only baseline. Colored curves: two-agent setting, with Agent 1 initialized at the baseline control sequence. (a) Normalized cost f_k/f_0 over outer iteration. Curves show the mean; shaded regions indicate the interquartile range (25th–75th percentiles). (b) Normalized control effort $\|x_k\|_1/\|x_0\|_1$, $\|u_k\|_1/\|x_0\|_1$ over outer iteration. Curves and shaded regions same as (a). (c) Inputs x, u, $x+u$ at initial (top) and final (bottom) iterations for a single seed. (d) Output y at initial (top) and final (bottom) iterations for a single seed.