Online Learning in Stabilized Linear Dynamical Games with Adversarial Disturbances

Anas Barakat
SUTD
anas_barakat@sutd.edu.sg

John Lazarsfeld
Georgia Tech
jlazarsfeld@gatech.edu

Georgios Piliouras
SUTD
georgios@sutd.edu.sg

Antonios Varvitsiotis
SUTD, CQT NUS, Archimedes
antonios@sutd.edu.sg

Abstract

We study online learning in linear dynamical games where multiple strategic agents act on a shared state evolving according to a linear dynamical system subject to adversarial disturbances. This setting lies beyond both single-agent nonstochastic online control and classical linear-quadratic games, which typically focus on quadratic objectives and noiseless or stochastic dynamics. Each agent seeks to minimize its own sequence of convex losses under full-state observability. Following the stabilizing-baseline paradigm in nonstochastic online control, we assume access to linear controllers that stabilize the noiseless system and focus on online adaptation by learning disturbance-action corrections. This extends adversarial online control to strategic multi-agent shared-state systems, where each agent's actions shape the state trajectory and hence the realized losses of all learners. Under state-only and aggregate-input feedback models, we analyze agents running online gradient descent with memory to update their own disturbance-action policies. We prove per-agent regret bounds that are sublinear and near-optimal in the time horizon, with population dependence that varies under the specified feedback, opponent policy, and stability assumptions. In the common-interest case, where all agents have identical cost functions, we show that the induced online problem forms a time-varying potential game and derive equilibrium tracking guarantees. Together, our results provide a theoretical framework for adversarial online learning in stabilized linear dynamical games, connecting online control with learning in games.

1 Introduction

Linear dynamical games are a classical model for strategic interaction in dynamical systems [8]. In the linear-quadratic case, multiple agents act on a shared linear system and optimize individual quadratic costs; this setting has long served as a benchmark for equilibrium analysis in control and, more recently, for studying learning dynamics and policy-gradient methods in multi-agent reinforcement learning [49, 33]. By contrast, little is known about linear dynamical games in the online nonstochastic regime, where multiple strategic agents must act causally on a shared state while both disturbances and costs may vary adversarially over time. Adversarial disturbances provide a worst-case model for misspecification, unmodeled dynamics, abrupt shocks, or strategic perturbations, without requiring distributional assumptions on the noise. This perspective is standard in nonstochastic online control [38] and is especially relevant in multi-agent systems, where the environment faced by each learner is shaped by both exogenous perturbations and the actions of other agents.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We study an online learning variant of linear dynamical games. The shared state evolves according to:

$$x_{t+1} = Ax_t + B_1 u_t^1 + \cdots + B_N u_t^N + w_t, \tag{LDS}$$

where N agents simultaneously choose controls $(u_t^i)_{i\in\{1,\cdots,N\}}$, A and $(B_i)_{i\in\{1,\cdots,N\}}$ are time-invariant transition matrices, and w_t is an arbitrary bounded disturbance. We call this an online linear dynamical game: agents act on a common linear system, but each agent evaluates the resulting trajectory through its own time-varying convex cost. At each time step t, each agent $i \in \{1, \cdots, N\}$ observes the full state x_t, selects its own control input u_t^i according to a causal policy π^i mapping the observation history to controls, and incurs an individual time-varying convex cost $c_t^i(x_t, u_t^i)$. Thus, the agents are coupled through the shared state: one agent's action affects the state trajectory, and hence the realized costs incurred by the other agents. This shared-state coupling makes the online problem inherently nonstationary from each agent's perspective: exogenous disturbances perturb the dynamics, while the adaptive controls of other agents change the state trajectory on which costs are evaluated.

Our formulation follows the stabilizing-baseline paradigm in nonstochastic online control [38]. We separate stabilization from online adaptation: agents are given nominal controllers that stabilize the noiseless system ($w_t = 0$), and the learning problem is to adapt around this stable backbone through disturbance-action corrections. As in the single-agent nonstochastic control literature, these stabilizing controllers are treated as part of the problem input; our focus is on the online adaptation problem induced by adversarial disturbances, changing costs, and shared-state coupling. We refer to this setting as a *stabilized linear dynamical game*: the baseline closed-loop dynamics are stable, while agents learn online correction policies in response to the realized adversarial environment.

Relative to single-agent nonstochastic control, the main challenge is analyzing counterfactual performance under shared-state coupling. Each agent controls only one component of the joint input, but the state trajectory depends on all agents' controls. Thus, even when evaluating unilateral regret against a fixed realized sequence of other agents' controls, the counterfactual trajectory induced by one agent's policy must be analyzed through the coupled dynamics. Different feedback models determine how much of this coupling each learner can account for, leading to different dependences on the number of agents. In the common-interest case, the same shared-state structure raises a collective question: whether individual online adaptation tracks equilibria of the induced time-varying game.

Thus, our setting sits between classical linear-quadratic games and single-agent nonstochastic online control. Unlike the former, we allow adversarial disturbances and time-varying convex costs; unlike the latter, multiple learners act on the same state. Consequently, even with a stabilizing baseline, it is unclear whether one can obtain individual regret guarantees or collective guarantees from simultaneous gradient updates under shared-state coupling. This motivates our central question:

In stabilized linear dynamical games with adversarial disturbances, can agents update their own policies online so as to achieve sublinear individual regret and collective equilibrium guarantees in the common-interest setting?

Contributions. We obtain the first regret and equilibrium-tracking guarantees for adversarial online learning in shared-state linear dynamical games. Our contributions are as follows.

- **Online learning in stabilized linear dynamical games.** We formulate an online nonstochastic variant of linear dynamical games in which agents act on a shared linear dynamical system, face individual time-varying convex costs, and adapt through disturbance-action corrections on top of stabilizing baseline controllers (Section 2). This framework extends single-agent nonstochastic online control to strategic shared-state systems.
- **Individual policy regret under state-only feedback.** In the state-only feedback model, agents observe the shared state but not the aggregate control contribution of the other agents. We prove a per-agent policy regret bound of $\widetilde{\mathcal{O}}(N^2\sqrt{T})$ (Theorem 1) for online gradient descent (Algorithm 1) over disturbance-action policies. The proof reduces each agent's problem to a single-agent nonstochastic control problem by absorbing the other agents' realized inputs into an effective disturbance. The resulting dependence on N arises from bounding the other agents' aggregate contribution to the effective disturbance.
- **Improved population scaling under aggregate-input feedback.** In the aggregate-input feedback model, agents additionally observe the combined contribution of the other agents' controls. Under compatible individual and joint strong stability, we improve the per-agent policy regret bound

to $\widetilde{\mathcal{O}}(N\sqrt{T})$ (Theorem 2). With an additional Lipschitz assumption on the costs, we remove the polynomial dependence on N, obtaining a $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bound with only polylogarithmic dependence on N (Theorem 3). Technically, the improvement comes from exploiting the structure of the joint input contributions in the shared dynamics, rather than treating the other agents' actions as an unstructured disturbance. Our analysis controls the induced state evolution under simultaneous DAC policies and balances the memory-truncation error with the online-gradient regret term to obtain the sharper dependence on N. By embedding the classical online learning lower bound into our admissible cost and controller classes, we show that the horizon dependence of Theorems 1–3 is optimal up to logarithmic factors (Theorem 4).

- **Equilibrium tracking in common-interest games.** When all agents share identical cost functions, we show that the induced linear dynamical game forms a time-varying potential game. We derive equilibrium-tracking guarantees in which the average equilibrium gap is controlled by the temporal variation of the costs and disturbances (Theorem 5). The proof uses a dedicated time-varying potential argument tailored to our stateful dynamical setting, independent of our individual regret bounds. Our equilibrium gap guarantee extends prior work in time-varying (stateless) normal-form games to continuous convex games whose time-varying costs depend on state trajectories generated by adversarially disturbed dynamics.

Related Work. We give a brief discussion of related works and defer more details to Appendix A.

Online non-stochastic control. Our work builds on a recent and growing line of research focusing on the use of online learning techniques to address control problems with adversarially perturbed dynamical systems [35, 1, 2, 37, 24, 55, 54, 30, 27, 10, 59, 29]. On the one hand, when the dynamical system (LDS) involves only a single agent (i.e., $N = 1$), our setting collapses to (single-agent) *online non-stochastic control*. This problem has been thoroughly studied over the past years, see e.g. [38] and the references therein. On the other hand, most of the works in this line of research are devoted to the control of linear dynamical systems influenced by a *single* controller. We discuss a few exceptions in the next section.

Linear dynamical games and multi-agent control. There is extensive research at the interface of control and game theory, see e.g. [45, 15] for surveys. An important body of this literature has focused on linear-quadratic games [8, 48, 39, 64, 9, 66, 61, 60, 49, 33]. Some of these works typically consider the same (LDS) and assume quadratic costs for systems which are either deterministic ($w_t = 0$) or perturbed by a noise sequence $\{w_t\}$ which is i.i.d. Gaussian. Classical approaches to design robust controllers in optimal control rely either on using probabilistic models for disturbances or adopting a (worst-case) 'minimax' perspective [7].

A few recent works adopt an online learning approach for *distributed* control: [13, 14] study a distributed online control problem over a multi-agent network of m identical linear systems, where each agent seeks to compete with the best centralized control policy in hindsight. This is fundamentally different from our setting, where we consider *selfish strategic* agents influencing a *single* linear dynamical system, and where each agent attempts to minimize their own individual cost. [28] propose a reduction from any standard regret minimizing control method to a distributed algorithm implemented by several controllers, which is distinct from our setting of multiple strategically competing agents. Recently, [29] proposed an online control approach for population dynamics where states are distributions in the simplex. We rather focus on the case of a finite and discrete large number of agents and discuss the influence of the total number of agents on individual regret.

Online learning with memory and in time-varying games. Our regret analysis uses tools from online learning with memory [5, 41]. Some of our results relate to the active research area of online learning in time-varying games [12, 20, 51, 23, 67, 4, 22, 62, 50, 58, 25, 26, 18]. However, these works do not address our online multi-agent control setting where time-varying costs depend on an underlying (LDS) with coupled state dynamics subject to adversarial disturbances.

2 Problem Formulation: Online Learning in Linear Dynamical Games

In this section, we formally introduce the online learning problem in linear dynamical games over a finite time horizon T. The state process evolves as a linear dynamical system:

$$x_{t+1} = Ax_t + \sum\nolimits_{i=1}^{N} B_i u_t^i + w_t, \quad t \geq 0\,, \tag{LDS}$$

where $x_t \in \mathcal{X} := \mathbb{R}^d$ is the state of the system initialized at a given (possibly random) state x_0, $u_t^i \in \mathcal{U}_i := \mathbb{R}^{k_i}$ is the strategy of agent $i \in [N] := \{1, \cdots, N\}$, $w_t \in \mathbb{R}^d$ is an arbitrary system disturbance and $A \in \mathbb{R}^{d\times d}, B_i \in \mathbb{R}^{d\times k_i}$ are the transition matrices defining the linear dynamical system. We assume $\|x_0\|_2 \le X_0$ almost surely for a finite constant X_0. All counterfactual trajectories start from the same realized initial state x_0. The bounds established in this paper may depend on X_0.

2.1 Online learning setting and feedback models

We consider the following online setting: at each time step t, all N agents observe the state x_t of the system. Then, each agent $i \in [N]$ selects a control input $u_t^i \in \mathbb{R}^{k_i}$ and incurs a loss $c_t^i(x_t, u_t^i)$, where $c_t^i : \mathbb{R}^d \times \mathbb{R}^{k_i} \to \mathbb{R}$ is an adversarially chosen cost function. Finally, the system transitions to the next state according to the dynamics (LDS). The goal of each agent i is to minimize their own cumulative cost over T rounds. We assume that each agent $i \in [N]$ knows the dynamics (A, B_i). For each $i \in [N]$, the cost function c_t^i is only accessible to agent i. The perturbation sequence $\{w_t\}$ is a priori unknown to agents. We study two *information settings* that differ in how much of the other agents' contribution to the shared dynamics is observable.

Information Setting 1 (State-only). *At each time step t, agent $i \in [N]$ observes the state x_t (fully observable setting) and then their own induced cost. Agent i has no access to the control inputs of other agents $j \neq i$.*

We also consider a second setting where agents have access to more information about the other interacting agents in the (LDS) system at each time step.

Information Setting 2 (Aggregate-Input). *At each time step t, agent i observes the state x_t, and then their own induced cost and the aggregated feedback $\sum_{j\neq i} B_j u_t^j$ that encodes information about other agents' control inputs. Each agent i knows the total number of agents N.*

This stronger information setting is analogous to the standard setting of full-information feedback (hindsight observability) in the literature of online learning in games. This setting allows a player to evaluate their loss for any counterfactual action. Similarly, in our setting, observing the state and aggregated control lets each agent reconstruct the disturbance and thus compute their counterfactual loss for any alternative control they could have individually chosen, given others' actions.

2.2 Regret framework for online learning in linear dynamical games

In this section, we give a formal definition of our performance metric for online learning in linear dynamical games, inspired from both single-agent online control and learning in games.

Individual policy regret. Since the system dynamics depend on unknown costs and possibly adversarial perturbations, determining an optimal controller a priori is not possible in general. Therefore, in contrast to classical and robust optimal control, we consider *regret* as a performance measure, following the recent line of works on (single-agent) online non-stochastic control [37]. For each agent i, let Π_i be a benchmark class of deterministic causal policies on its observation history. Writing h_t^i for the information available before the round-t control and ξ_i for internal randomness, the learner chooses $u_t^i = \mathcal{A}_i(\mathsf{h}_t^i, \xi_i)$ using algorithm $\mathcal{A}_i$. For any $T \ge H \ge 1$, we define the regret of agent i w.r.t. policy class Π_i when agent i runs algorithm $\mathcal{A}_i$ and other agents use controls $\{u_t^{-i}\}$ as:

$$\mathrm{Reg}_i^{H:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i) = \sum\nolimits_{t=H}^{T} c_t^i(x_t, u_t^i) - \min_{\pi^i\in\Pi_i} \sum\nolimits_{t=H}^{T} c_t^i(x_t^{\pi^i}, u_t^{\pi^i}), \tag{1}$$

where $x_t^{\pi^i}, u_t^{\pi^i}$ are the *counterfactual state and controls* under the policy π^i for agent i:

$$x_{t+1}^{\pi^i} = Ax_t^{\pi^i} + B_i u_t^{\pi^i} + \sum\nolimits_{j\neq i} B_j u_t^j + w_t\,, \quad x_0^{\pi^i} = x_0, \qquad u_t^{\pi^i} = \pi^i(\mathsf{h}_t^{i,\pi^i}).$$

Here h_t^{i,π^i} is the policy's counterfactual observation history. The counterfactual state sequence corresponds to the state sequence that would be observed if agent i were to unilaterally deviate to using policy π^i, instead of their online learning algorithm $\mathcal{A}_i$ (and where all other agents stick to their online control input sequence). Note that when $N = 1$, expression (1) recovers the regret definition for single-agent online control.

In this work, we consider two natural policy comparator classes, which we introduce as follows:

Comparator policy class 1: Strongly stable linear controllers (Π_i^{lin}). For agent i, a *linear controller* is defined by a matrix $K_i \in \mathbb{R}^{k_i \times d}$ s.t. $u_t^i = -K_i x_t$. We say that a linear policy K_i is *stable* if $\rho(A - B_i K_i) < 1$ (where $\rho(\cdot)$ denotes the spectral radius), in which case the closed-loop state-feedback linear dynamical system is globally asymptotically stable. *Strong stability* of a controller is a quantitative version of stability which allows for deriving non-asymptotic guarantees.

Definition 1 (Strong stability, e.g. [17]). *A linear policy K is (κ, γ)-strongly stable (for $\kappa > 0$ and $0 < \gamma < 1$) for a linear dynamical system specified by (A, B) if $\|K\| \le \kappa$ and if there exists matrices L, Q s.t. $A - BK = QLQ^{-1}$ with $\|L\| \le 1 - \gamma$, and $\|Q\| \cdot \|Q^{-1}\| \le \kappa$.*

Note that strong stability implies stability, and any stable policy is strongly-stable for some (κ, γ). A natural policy comparator class is that of *strongly stable linear controllers* Π_i^{lin}, parameterized by:

$$\mathcal{K}_i := \left\{ K_i \in \mathbb{R}^{k_i \times d} : K_i \text{ is } (\kappa_i, \gamma_i)\text{-strongly stable} \right\}, \text{ for some } \kappa_i > 0, \gamma_i \in (0, 1).$$

Comparator policy class 2: Disturbance Action Controller (DAC) policies (Π_i^{DAC}). The state sequence induced by a linear controller is not a linear function of its parameters. As a consequence, the induced cost is non-convex in the control parameters in general, even if the cost function is convex in both the state and the control input (see, e.g., [21]). Following prior work in single-agent control, we consider *Disturbance Action Controller* (DAC) policies. The system state induced by such policies is *affine* in the policy parameters and one can invoke tools from online convex optimization when the cost functions are convex in the state and control input. For a sequence of perturbations $\{w_t\}$, a DAC policy $\pi^i(M_i, K_i)$ for agent $i \in [N]$ is then specified by learnable matrix parameters $M_i = [M_i^{[0]}, M_i^{[1]}, \cdots, M_i^{[H-1]}]$ for a memory length $H \ge 1$, with a fixed given stabilizing controller K_i. The policy $\pi^i(M_i, K_i)$ selects action u_t^i at a state x_t as:

$$u_t^i = -K_i x_t + \sum_{p=1}^{H} M_i^{[p-1]} w_{t-p}\,. \tag{DAC-i}$$

Note that for $p < 0$ we let $w_p = 0$, and moreover, the perturbations w_t are not observed by the learners but rather computed online using the structure of (LDS) and the state observations (we discuss these points later). The policy can thus be implemented in an online fashion by agent i, and we henceforth use the notation $M_{i,t} = [M_{i,t}^{[p]}]_{0 \le p \le H-1}$ to reference the parameters of player i at time t. For a fixed H and stabilizing controller K_i, let $\mathcal{M}_i = \left\{ M_i = \{M_i^{[0]}, \cdots, M_i^{[H-1]}\} : \|M_i^{[p]}\|_{\text{op}} \le 2\kappa^2(1-\gamma)^p, p = 0, \cdots, H-1 \right\}$ denote the set of all DAC policy parameters for agent i, where $(\kappa, \gamma) = (\kappa_i, \gamma_i)$ under Setting 1 and $(\kappa, \gamma) = (\bar{\kappa}, \bar{\gamma})$ under Setting 2.

Remark 1. *The stacked representation $x_{t+1} = Ax_t + Bu_t + w_t$, with $B = [B_1, \ldots, B_N]$ and $u_t = (u_t^1, \ldots, u_t^N)$, does not by itself reduce our individual regret problem to centralized online control [2]. Player i optimizes its own cost, chooses only u_t^i, and uses local model knowledge (A, B_i) and the prescribed feedback. Its comparator changes only its own policy while fixing opponents' realized controls $u_{1:T}^{-i}$. A centralized controller instead optimizes a specified objective over joint policies, generally inducing different counterfactual trajectories. Therefore, its joint regret guarantee does not automatically imply our individual unilateral guarantees.*

3 Individual Regret Guarantees

In this section, we present our results on individual regret guarantees. We analyze an online gradient descent with memory algorithm, where each agent independently updates its DAC policy parameters via online gradient descent (Algorithm 1). In the single-agent setting ($N = 1$), this algorithm was introduced and analyzed by [2]. In our multi-agent setting, the coupling of state dynamics across all agents induces new obstacles to implementing and analyzing this gradient-based approach. We elaborate first on the computational challenge.

Memory. The cost $c_t^i(x_t, u_t^i)$ incurred by agent $i \in [N]$ at time step t depends on the state x_t of the system, which itself depends on all past states and control inputs from $t = 0$. However, to run the online gradient descent subroutine of Algorithm 1, agent i must be able to evaluate its cost function c_t^i on counterfactual state-action pairs. Unlike the single-agent case, counterfactual evaluation here depends not only on the agent's own past controls but also on the entire joint sequence of other agents' controls. This dependence breaks the straightforward counterfactual construction of the single-agent setting and requires a new memory-based approximation tailored to the multi-agent coupling.

Fix a player i and the realized controls of the other agents. Set $A_i = A - B_i K_i$ and $d_s = w_s + \sum_{j \neq i} B_j u_s^j$. The DAC disturbance sequence is $r_s = d_s$ under Setting 1 and $r_s = w_s$ under Setting 2. Both sequences are held fixed in unilateral counterfactuals, and $d_s = r_s = 0$ for $s < 0$. For a fixed candidate parameter $Q \in \mathcal{M}_i$, the true counterfactual trajectory is generated by executing the corresponding DAC policy from time zero, starting from x_0, and therefore depends on the entire past input sequence. Recomputing this trajectory from time zero for each candidate would require work growing with t. To obtain an efficiently evaluable loss with bounded memory, we truncate the counterfactual rollout and initialize it at zero at a recent truncation boundary. Individual strong stability of A_i makes the influence of the state at that boundary decay geometrically with the rollout length. Together with the state bounds established below, this allows us to control the cumulative approximation error using a memory length H logarithmic in T. Specifically, let $h = H + 1$. Starting from zero at time $t - h$, the truncated state and action under Q are given as follows:

$$y_t^i(Q) = \sum_{a=1}^{h} A_i^{a-1} \left(d_{t-a} + B_i \sum_{p=1}^{H} Q^{[p-1]} r_{t-a-p} \right), v_t^i(Q) = -K_i y_t^i(Q) + \sum_{p=1}^{H} Q^{[p-1]} r_{t-p}\,, \quad (2)$$

and $f_t^i(Q) = c_t^i(y_t^i(Q), v_t^i(Q))$. These finite sums define the surrogate for every $t \geq 1$, including early rounds through zero padding, and involve an input history of length $\mathcal{O}(H)$. The map f_t^i is convex and can be evaluated using the player's local model, its cost feedback, and the reconstructed inputs. Appendix D.1 defines the true unilateral counterfactual and the corresponding memory loss $\ell_t^i(M_{i,t-h:t})$, whose diagonal is f_t^i. The single-argument notation $\ell_t^i(Q)$ also denotes $f_t^i(Q)$.

Algorithm variants. Depending on the information setting (Settings 1 and 2), we define two variants of Algorithm 1, each described from the perspective of a fixed agent $i \in [N]$. These variants capture different levels of feedback and are essential for obtaining our regret guarantees.

Algorithm 1 Online Gradient Descent with Memory for agent $i \in [N]$

Input: memory H, step size $\eta > 0$, $M_{i,1} \in \mathcal{M}_i$, given stabilizing baseline linear controller K_i.
Observe x_0, apply $u_0^i = -K_i x_0$, and set $M_{i,s} = M_{i,1}$ for $s \leq 0$ and $r_s = d_s = 0$ for $s < 0$.
for $t = 1, \ldots, T$ **do**
  Observe x_t and compute $d_{t-1} = x_t - A x_{t-1} - B_i u_{t-1}^i$.
  if Information Setting 1 **then**
    Set $r_{t-1} = d_{t-1}$.
  else
    Observe $\sum_{j \neq i} B_j u_{t-1}^j$ and set $r_{t-1} = w_{t-1} = d_{t-1} - \sum_{j \neq i} B_j u_{t-1}^j$.
  end if
  Apply $u_t^i = -K_i x_t + \sum_{p=1}^{H} M_{i,t}^{[p-1]} r_{t-p}$.
  Incur $c_t^i(x_t, u_t^i)$ and receive the post-action cost feedback.
  Construct f_t^i using (2) and update $M_{i,t+1} = \Pi_{\mathcal{M}_i}\big(M_{i,t} - \eta \nabla f_t^i(M_{i,t})\big)$.
end for

Remark 2. *Gradients and projections use the block Frobenius inner product, with* $\|M_i\|_F^2 = \sum_{p=0}^{H-1} \|M_i^{[p]}\|_F^2$. *Algorithm 1 assumes that each agent can evaluate the gradients* $\nabla \ell_t^i$ *of their own cost function evaluated at counterfactual state-action pairs.*

Standing Assumptions. Finally, before presenting our regret guarantees, we introduce our standing assumptions, all standard in the recent literature on online control (see e.g., [2, 38, 37, 30, 52]).

Assumption 1 (Cost functions)**.** *The following assumptions hold for every* $i \in [N]$*:*

(i) The cost function $c_t^i : \mathcal{X} \times \mathcal{U}_i \to \mathbb{R}$ *is jointly convex and differentiable.*

(ii) There exist $\beta, G > 0$ *s.t. for any* $D > 0$ *and every* $(x, u^i) \in \mathcal{X} \times \mathcal{U}_i$ *s.t.* $\|x\| \leq D, \|u^i\| \leq D$*, we have* $|c_t^i(x, u)| \leq \beta D^2$ *and* $\|\nabla_x c_t^i(x, u^i)\|, \|\nabla_u c_t^i(x, u^i)\| \leq GD$ *.*

Lemma 1. *Under Assumption 1, the loss function* ℓ_t^i *is convex w.r.t.* M_i *for all* $i \in [N]$*.*

Assumption 2 (Bounded disturbances)**.** *There exists* $W > 0$ *s.t. for all* $t \geq 0$*,* $\|w_t\| \leq W$ *.*[1]

[1]This assumption is standard in the online control literature. See [40] for recent work relaxing it.

3.1 Information setting 1: State-only Feedback

Under Information Setting 1, agents do not have access to other agents' control inputs. However, from the viewpoint of a given agent i, we observe that (LDS) can be rewritten:

$$x_{t+1} = Ax_t + B_i u_t^i + \widetilde{w}_t\,, \quad \widetilde{w}_t = \sum\nolimits_{j\neq i} B_j u_t^j + w_t\,. \tag{3}$$

In this view, in Algorithm 1, we naturally propose that agent i executes a (DAC-i) policy with disturbance sequence $\widetilde{w}_t$. Given expression (3), note that $\widetilde{w}_t$ (unlike w_t) can be calculated by agent i at each time step since $\widetilde{w}_t = x_{t+1} - Ax_t - B_i u_t^i$, and this computation only involves information observed under the information setting (state observations and the agent's control input). Under this strategy, agent i thus faces a LDS (3) controlled by its own control inputs, and for this we make a strong stability assumption adapted to the multi-agent setting.

Assumption 3 (Agent-wise strong stability)**.** *Each learner $i \in [N]$ knows a (κ_i, γ_i)-strongly stable linear controller K_i for the LDS specified by (A, B_i).*

Under this assumption, we present our first regret bound.

Theorem 1 (Individual regret under state-only feedback). *Let Assumptions 1, 2 and 3 hold. Suppose there exists $U \geq 1$ s.t. for all $t \geq 0, j \in [N], \|u_t^j\| \leq U$. If agent $i \in [N]$ runs Algorithm 1 under Setting 1 with (DAC-i) policy on perturbation sequence $\{\widetilde{w}_t\}$ and step size $\eta = \Theta(1/(G\widetilde{W}^2\sqrt{T}))$, where $\widetilde{W} = W + (N-1)U(\max_j \|B_j\|)$, and with $H = \max\{1, \lceil \log(\kappa_i T)/\gamma_i \rceil\}$, then for $T \geq H+1$, $\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{lin}) = \widetilde{\mathcal{O}}(U^2 N^2 \sqrt{T})$.*[1]

The proof of Theorem 1 consists of applying the single-agent regret guarantee for gradient perturbation controllers [2, Theorem 5.1] for each agent i on the new perturbation sequence $\{\widetilde{w}_t\}$ in (3), and we give the full details in Section E. While the theorem highlights the robustness of gradient perturbation controllers to adversarial disturbances in this setting, the regret bound grows quadratically with both the number of agents N and the magnitude U of the control inputs. This dependence arises from bounding the other agents' aggregate input as part of the effective disturbance. The resulting upper bound deteriorates with N and U. Whether its dependence on N is unavoidable remains open.

Remark 3. *The boundedness of control inputs in Theorem 1 is a restrictive condition on the realized opponent inputs. Individual stability alone does not imply it, see Section C for further discussion.*

3.2 Information setting 2: Aggregate-Input Feedback

The regret bound in Theorem 1 scales quadratically with the number of agents. We now consider Information Setting 2 and analyze the case in which all agents use DAC policies. Under a global strong stability assumption, we obtain an individual regret bound with improved dependence on N.

Assumption 4 (Global strong stability)**.** *Each learner $i \in [N]$ knows a linear controller K_i such that $(K_1, \cdots, K_N)^T$ is $(\bar{\kappa}, \bar{\gamma})$-strongly stable for the LDS $(A, [B_1, \cdots, B_N])$.*

Assumption 4 is a global assumption which is relevant when each agent i executes a (DAC-i) policy (with matrix K_i). Indeed, the system state evolution of (LDS) in the absence of disturbances, and when all players use linear controllers, can be written as $x_{t+1} = Ax_t - [B_1, \cdots, B_N](K_1, \cdots, K_N)^T x_t$. Without global stability, if the spectral radius of $A - \sum_{i=1}^N B_i K_i$ is not smaller than 1, the states can become unbounded. Each agent i has access to the global parameters $\bar{\kappa}, \bar{\gamma}$ which can be precomputed before each agent runs their own Algorithm 1. Recall that the matrices K_i are not learning parameters and are precomputed even in the independent learning setting. Only the matrix parameters M_i of (DAC-i) policies are learned by the agents.

Under Setting 2, all agents can compute the original disturbance w_t at each time step (instead of $\tilde{w}_t$ as in Theorem 1). However, at every timestep t, each agent updates their own policy parameters without access to other agent's policy parameters at that round. After acting, each agent first incurs the loss according to their individual cost function, and *then* observes the aggregated feedback. This feedback is used to inform their next policy update.

[1] For readability, here and throughout, we use $\widetilde{\mathcal{O}}$ to hide polynomial factors in problem parameters and (poly)logarithmic factors in T and N. We state the exact dependencies in the proofs.

Our next result shows that when agent i runs Algorithm 1 with (a) a conservative stepsize scaled by N and (b) a larger memory which depends logarithmically on N (compared to Theorem 1), they guarantee a regret w.r.t. the DAC policy class scaling only *linearly* in N. This result is also robust to other agents' strategies (as they can execute arbitrary (DAC-i) policies). Theorem 2 compares with memory-H DAC policies as other players also use DAC policies to be robust to the disturbance sequence (w_t), whereas Theorem 1 compares with strongly stable linear controllers.

Theorem 2 (Individual regret under aggregate-input feedback). *Let Assumptions 1 to 4 hold. Then if agent $i \in [N]$ runs Algorithm 1 under Setting 2 with a (DAC-i) policy on perturbation sequence $\{w_t\}$, step size $\eta = 1/(N\sqrt{T})$, and with $H = \lceil \log(2\kappa_* N^2\sqrt{T})/\gamma_* \rceil$ where κ_*, γ_* are the stability bounds in (47), and when **all** other agents use (DAC-i) policies with perturbation sequence (w_t) and feasible parameters $M_{j,t} \in \mathcal{M}_j$, then for any $T \geq H+1$,*

$$\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{DAC}) = \widetilde{\mathcal{O}}(N\sqrt{T}).$$

Proof Overview. Our analysis relies on a regret decomposition with two main terms: a counterfactual state-control error due to the use of a loss with limited memory H, and a regret term induced by the online gradient descent component of Algorithm 1. The main technical challenges we overcome are two-fold: first, states may grow unbounded with an undesirable scaling in N, and thus we control their magnitude by studying the state evolution when all agents use DAC policies (using Assumption 4), and while tracking the dependence on N. Second, we control both terms of the regret decomposition by selecting H and η to balance the terms in our upper bound. See Appendix F.

We remark that the linear dependence on N in the regret bound is enabled by global stability (Assumption 4). By contrast, if only individual stability (Assumption 3) holds, even when agents can access aggregated input information, the dependence on N deteriorates (see Appendix C.3 for a discussion). Moreover, under a stronger assumption on the cost functions (compared to Assumption 1-(ii)), we further prove a sublinear regret for agent i that scales only *polylogarithmically* in N.

Assumption 5 (Lipschitz costs)**.** *There exists $\bar{L} > 0$ s.t. for any agent $i \in [N]$ and for all state-control pairs $(x, u^i), (\tilde{x}, \tilde{u}^i) \in \mathcal{X} \times \mathcal{U}_i$, $|c_t^i(x, u^i) - c_t^i(\tilde{x}, \tilde{u}^i)| \leq \bar{L}(\|x - \tilde{x}\| + \|u^i - \tilde{u}^i\|)$.*

Note that the Lipschitz constant does not scale with the state and control input magnitude. Under this assumption, we obtain an improved regret guarantee (proven in Appendix G):

Theorem 3. *Under the setting of Theorem 2, replace gradient boundedness in Assumption 1 -(ii) by Assumption 5. Set instead $\eta = 1/\sqrt{T}$ and choose $H = \lceil \log(2\kappa_* N\sqrt{T})/\gamma_* \rceil$ with κ_*, γ_* as in (47). Then we have for any $T \geq H+1$: $\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{DAC}) = \widetilde{\mathcal{O}}(\sqrt{T})$.*

Note that using Assumption 5 in Theorem 1 does not result in the same improved dependence on N as the regret will still scale with the magnitude of the modified disturbance $\widetilde{w}_t$, which is of order N.

3.3 Regret lower bounds

We next show that the $\sqrt{T}$ horizon dependence of our regret guarantees is optimal up to logarithmic factors. The following lower bound applies under both information settings, against both strongly stable linear controllers and fixed-baseline DAC policies, even when the costs satisfy Assumption 5. For a nonempty interval $I \subseteq \{1, \ldots, T\}$, let Reg_i^I be the regret in (1) with sums taken over $t \in I$.

Theorem 4 (Horizon lower bound for linear and DAC comparators). *Let $N \geq 1$ and $i \in [N]$. There exist scalar dynamics, deterministic disturbances, and supplied baseline controllers satisfying Assumptions 2–4 with $W = 1, \kappa_j = \bar{\kappa} = 1, \gamma_j = \bar{\gamma} = \frac{1}{2}$, such that the following holds. For every $T \geq 1$, $H \geq 1$, prescribed nonempty interval $I \subseteq \{1, \ldots, T\}$, and causal, possibly randomized learner $\mathcal{A}_i$ under either information setting, there exists a deterministic oblivious cost sequence satisfying Assumption 1 with $\beta = G = 4$ and Assumption 5 with $\bar{L} = 1$ for which $\mathbb{E}_{\mathcal{A}_i}[\mathrm{Reg}_i^I(\mathcal{A}_i, \{u_t^{-i}\}, \Pi)] \geq \frac{\sqrt{|I|}}{2\sqrt{3}}$ simultaneously for the $(1, \frac{1}{2})$-strongly stable linear class $\Pi = \Pi_i^{\mathrm{lin}}$ and the memory-H DAC class $\Pi = \Pi_i^{\mathrm{DAC}}$ with fixed supplied baseline $K_i = 0$. The conclusion holds for every sequence of opponents' controls.*

Remark 4. *The expectation in Theorem 4 is over the learner's internal randomization. The dynamics, disturbances, and supplied baseline controllers are independent of $T, H, I, \mathcal{A}_i$. For each fixed $T, H, I, \mathcal{A}_i$, the cost sequence is selected before interaction. It may depend on the learner's decision rule and randomization law, but not on its realized random seed. The same cost sequence establishes the bound for both comparator classes.*

Proof idea and relation to online learning. The proof embeds the classical random sign lower bound for online learning [6, 36] into an admissible scalar instance of our model. We choose $A = B_j = 0$ for every player j, $w_t = 1$, and $x_0 = 0$, so the learner and every comparator have the same state $x_t = 1$ for all $t \geq 1$. At this common state, the constructed costs satisfy $c_t^i(1, u) = 3/4 + \sigma_t u$ for $|u| \leq 1/2$, where the signs σ_t are independent and uniform. The proof verifies that these losses extend to jointly convex, continuously differentiable costs satisfying Assumptions 1 and 5 globally, and that the comparison actions $u = \pm 1/2$ are realized by both admissible linear controllers and memory-H DAC policies with the fixed supplied baseline $K_i = 0$. Convexity also gives the required conditional expected-cost bound for arbitrary real learner controls. These verifications establish the embedding into our cost and policy classes while the statistical lower bound mechanism is the classical random sign argument. We provide a complete and self-contained proof in Appendix H.

Implications for the upper bounds. Taking $I = \{1, \ldots, T\}$ gives an $\Omega(\sqrt{T})$ full-horizon lower bound. For the evaluation interval in Theorems 1–3, taking $I = \{H + 1, \ldots, T\}$ gives $\sqrt{T - H}/(2\sqrt{3})$, which is $\Omega(\sqrt{T})$ whenever $H \leq T/2$. Thus, with the remaining problem parameters fixed and logarithmic memory, our upper bounds have optimal horizon dependence up to logarithmic factors. The constructed costs also satisfy Assumption 5, so this conclusion covers Theorem 3. The lower bound already holds for a single agent and does not establish the necessity of the factors depending on N in our upper bounds.

4 Equilibrium Tracking in the Common Interest Setting

In the previous section, we developed individual regret guarantees when other agents execute linear or DAC policies with possibly misaligned or adversarially-chosen cost functions. In this section, we focus on the *common interest setting*, where all agents incur the same state-only cost [2]: $c_t^i(x, u_i) = c_t(x)$ for every $i \in [N], x \in \mathcal{X}, u_i \in \mathcal{U}_i$. We retain the bounded disturbance and joint strong-stability assumptions, and the convexity and growth conditions of Assumption 1, specialized to $c_t(x)$.

Fix a memory length $H \geq 1$ and write $\bar{A} = A - \sum_{i=1}^N B_i K_i$. Set $w_s = 0$ for $s < 0$. Let $\mathcal{M} = \prod_{i=1}^N \mathcal{M}_i$ be the product of the DAC parameter domains. For a fixed joint profile $M = (M_1, \ldots, M_N)$, define the zero-initialized truncated rollout:

$$y_t(M) = \sum_{r=1}^{H+1} \bar{A}^{r-1} \left(w_{t-r} + \sum_{i=1}^N B_i \sum_{p=1}^H M_i^{[p-1]} w_{t-r-p} \right), F_t(M) := \bar{\ell}_t(M) := c_t(y_t(M)). \quad (4)$$

Since the cost functions are time-varying (not only via the strategies of the different players), our linear dynamical game can be seen as a *time-varying game*. There have been considerable efforts endeavoring to extend the scope of traditional game-theoretic results to the time-varying setting and this is an active research area (see related work in Section 1). In particular, our results in this section are inspired from recent developments for time-varying, normal-form, finite potential games in [4]. In such games, agents participate in a potential game at each time step. We observe that a common interest linear dynamical game can be seen as a stateful, time-varying potential *continuous* convex game where costs are functions of states driven by an underlying (LDS) influenced by multiple agents. At each time step, the utility of each player is given by their cost function, and their strategy is defined by their DAC policy parameters.

Since our setting involves adversarial, time-varying costs depending on state dynamics influenced by adversarial (time-varying) disturbances, convergence to (static) Nash equilibria is irrelevant in general. Nevertheless, we establish equilibrium gap *tracking* guarantees for our dynamic setting. To

[2] Our analysis also extends to a common joint cost $C_t(x, u^1, \ldots, u^N)$ under analogous joint convexity, growth, and smoothness assumptions and access to the corresponding coordinate gradients. Using a joint state–action surrogate gives an analogue of Theorem 5, with modified constants and cost-variation budget. We restrict to state-only costs for simplicity.

state our result, we introduce notations for time-varying best responses and equilibrium gaps:

$$g_{i,t}(M) := F_t(M) - \min_{Q_i \in \mathcal{M}_i} F_t(Q_i, M_{-i}), \qquad g_t(M) := \max_{i \in [N]} g_{i,t}(M). \tag{5}$$

After choosing its round-t action, each player i has access to $\nabla_i F_t(M_t)$ and performs the update:

$$M_{i,t+1} = \Pi_{\mathcal{M}_i}\left(M_{i,t} - L^{-1}\nabla_i F_t(M_t)\right), \tag{6}$$

where $L > 0$ is the uniform smoothness bound for F_t specified in Appendix I. Projections and parameter gradients use the block Frobenius inner product. All players use the same current profile M_t in this update. We additionally assume access to coordinate gradients in this section. This access is guaranteed for instance under access to the common cost function, the relevant joint closed-loop response, the disturbance history, and the current joint DAC profile.

We now make a smoothness assumption on the common cost function c_t that is standard in the analysis of gradient methods in both optimization and learning in games.

Assumption 6 (Smoothness of the common state cost)**.** *There exists $\zeta \geq 0$ such that, for every t and all $x, x' \in \mathbb{R}^d$, $\|\nabla c_t(x) - \nabla c_t(x')\|_2 \leq \zeta \|x - x'\|_2$.*

Under these assumptions, when all agents use the updates (6), we bound the average equilibrium gap by the variation of both cost functions and disturbances defined as follows for any $T \geq 1$:

$$V_c(T) := \sum_{t=1}^{T-1} \sup_{\|x\|_2 \leq D_x} |c_{t+1}(x) - c_t(x)|, \quad V_w^+(T) := \|w_0\|_2 + \sum_{s=0}^{T-2} \|w_{s+1} - w_s\|_2, \tag{7}$$

where D_x uniformly bounds $\|y_t(M)\|_2$ and G_x uniformly bounds $\|\nabla c_t(x)\|_2$ on $\|x\|_2 \leq D_x$. An empty sum is equal to zero.

Theorem 5 (**Equilibrium tracking**)**.** *Let Assumptions 1, 2, and 4 hold for the common-state-cost model specified above, together with the stated feedback and coordinate-gradient access and Assumption 6. Then, for every $M_1 \in \mathcal{M}$, the simultaneous updates (6) satisfy, for every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(M_t)^2 \leq \frac{2C_M}{LT}\left[F_1(M_1) + V_c(T) + G_x Q\, V_w^+(T)\right], \tag{8}$$

where the constants D_x, G_x, Q, L, C_M, independent of T for fixed H, are given in Appendix I.

For fixed memory and fixed problem parameters, sublinear cost and disturbance variation imply a vanishing average squared equilibrium gap. If both variation budgets are uniformly bounded, the average squared gap is $O(T^{-1})$ and some iterate has gap $O(T^{-1/2})$. Notice that these guarantees concern the diagonal truncated DAC stage games defined in (4)-(5).

Proof overview. To prove the theorem, we use a time-varying potential-based argument tailored to our dynamic stateful setting, independent from our individual regret guarantees. We extend the approach of [4] for time-varying *(finite) normal-form* potential games to (a) cover *(continuous) convex games* and (b) account for state dynamics and adversarial disturbances in addition to the time-varying costs in our linear dynamic game setting. See Appendix I for an overview and a detailed proof.

5 Conclusion and Future Work

This work initiates and makes progress on online learning in linear dynamic games in strategic environments subject to adversarial disturbances, taking a first step toward bridging online control with learning in games. In particular, we proved individual regret guarantees and equilibrium tracking guarantees for the induced truncated DAC stage games for online learning with adversarial disturbances and time-varying costs. Our results also open several directions for future research. On the technical side, it is interesting to investigate whether tighter regret bounds can be obtained with respect to the number of agents or under structural assumptions such as time-invariant costs. On the modeling side, important challenges include extending our analysis to settings with a network of communicating agents with local state observations, unknown or time-varying dynamics [37, 52, 31] and to feedback-limited regimes [63], where learners can only access partially observed states and partially informed bandit costs. A broader challenge is to go beyond *linear* state dynamics.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their valuable feedback, which helped improve this manuscript. The initial version of this work was completed without generative AI assistance. In preparing the revised version, we used ChatGPT to assist with improving the presentation of the results, revising the proofs, and conducting the simulations reported in the appendix. The authors take full responsibility for the final content.

This work is supported by the MOE Tier 2 Grant (MOE-T2EP20223-0018), the CQT++ Core Research Funding Grant (SUTD) (RS-NRCQT-00002), the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (Award Number: AISG2-RP-2020-016), and partially by Project MIS 5154714 of the National Recovery and Resilience Plan, Greece 2.0, funded by the European Union under the NextGenerationEU Program.

References

[1] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[2] Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

[3] Amin Aghajani and Ali Doustmohammadi. Formation control of multi-vehicle systems using cooperative game theory. In *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, pages 704–709. IEEE, 2015.

[4] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On the convergence of no-regret learning dynamics in time-varying games. *Advances in Neural Information Processing Systems*, 36:16367–16405, 2023.

[5] Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. *Advances in Neural Information Processing Systems*, 28, 2015.

[6] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[7] Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[8] Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

[9] Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

[10] Songfu Cai, Fei Han, and Xuanyu Cao. Performative control for linear dynamical systems. *Advances in Neural Information Processing Systems*, 2024.

[11] Dan Calderone and Meeko Oishi. The impact of social value orientation on nash equilibria of two player quadratic games. *arXiv preprint arXiv:2411.08809*, 2024.

[12] Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against nash equilibria in adversarially changing zero-sum games. In *International Conference on Machine Learning*, pages 921–930. PMLR, 2019.

[13] Ting-Jui Chang and Shahin Shahrampour. Regret analysis of distributed online control for lti systems with adversarial disturbances. *arXiv preprint arXiv:2310.03206*, 2023.

[14] Ting-Jui Chang and Shahin Shahrampour. Regret analysis of distributed online lqr control for unknown lti systems. *IEEE Transactions on Automatic Control*, 69(1):667–673, 2023.

[15] Fei Chen and Wei Ren. On the control of multi-agent systems: A survey. *Foundations and Trends® in Systems and Control*, 6(4):339–499, 2019.

[16] Chih-Yuan Chiu, Jingqi Li, Maulik Bhatt, and Negar Mehr. To what extent do open-loop and feedback nash equilibria diverge in general-sum linear quadratic dynamic games? *IEEE Control Systems Letters*, 8:2583–2588, 2024.

[17] Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pages 1029–1038. PMLR, 2018.

[18] Ludovico Crippa, Yonatan Gur, and Bar Light. Equilibria under dynamic benchmark consistency in non-stationary multi-agent systems. *arXiv preprint arXiv:2501.11897*, 2025.

[19] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

[20] Benoit Duvocelle, Panayotis Mertikopoulos, Mathias Staudigl, and Dries Vermeulen. Multiagent online learning in time-varying games. *Mathematics of Operations Research*, 48(2):914–941, 2023.

[21] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

[22] Yi Feng, Hu Fu, Qun Hu, Ping Li, Ioannis Panageas, Xiao Wang, et al. On the last-iterate convergence in time-varying zero-sum games: Extra gradient succeeds where optimism fails. *Advances in Neural Information Processing Systems*, 36:21933–21944, 2023.

[23] Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, and Lillian Ratliff. Online learning in periodic zero-sum games. *Advances in Neural Information Processing Systems*, 34:10313–10325, 2021.

[24] Dylan Foster and Max Simchowitz. Logarithmic regret for adversarial online control. In *International Conference on Machine Learning*, pages 3211–3221. PMLR, 2020.

[25] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. Time-varyingness in auction breaks revenue equivalence. *arXiv preprint arXiv:2410.12306*, 2024.

[26] Yuma Fujimoto, Kaito Ariu, and Kenshi Abe. Synchronization in learning in periodic zero-sum games triggers divergence from nash equilibrium. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23194–23202, 2025.

[27] Udaya Ghai, Arushi Gupta, Wenhan Xia, Karan Singh, and Elad Hazan. Online nonstochastic model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 36:23362–23388, 2023.

[28] Udaya Ghai, Udari Madhushani, Naomi Leonard, and Elad Hazan. A regret minimization approach to multi-agent control. In *International Conference on Machine Learning*, pages 7422–7434. PMLR, 2022.

[29] Noah Golowich, Elad Hazan, Zhou Lu, Dhruv Rohatgi, and Y. Jennifer Sun. Online control in population dynamics. In *Advances in Neural Information Processing Systems*, volume 37, pages 111571–111613, 2024.

[30] Paula Gradu, John Hallman, and Elad Hazan. Non-stochastic control with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 10764–10774, 2020.

[31] Paula Gradu, Elad Hazan, and Edgar Minasyan. Adaptive regret for control of time-varying dynamics. In *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, 2023.

[32] Yuxiang Guan, Giulio Salizzoni, Maryam Kamgarpour, and Tyler H Summers. A policy iteration algorithm for n-player general-sum linear quadratic dynamic games. *IEEE CDC*, 2024.

[33] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods find the nash equilibrium in n-player general-sum linear-quadratic games. *Journal of Machine Learning Research*, 24(139):1–56, 2023.

[34] Le Han, Yeming Lin, Kun Liu, Bing Cui, and Yuanqing Xia. Formation control via linear quadratic differential game with model uncertainties and unknown disturbances. In *2019 Chinese Control Conference (CCC)*, pages 6200–6205. IEEE, 2019.

[35] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[36] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[37] Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pages 408–421. PMLR, 2020.

[38] Elad Hazan and Karan Singh. *Introduction to Online Control*. Cambridge University Press, 2025.

[39] Sara Hosseinirad, Giulio Salizzoni, Alireza Alian Porzani, and Maryam Kamgarpour. On linear quadratic potential games. *arXiv preprint arXiv:2305.13476*, 2023.

[40] Kaito Ito and Taira Tsuchiya. Online control of linear systems under unbounded noise. *arXiv preprint arXiv:2402.10252*, 2024.

[41] Raunak Kumar, Sarah Dean, and Robert Kleinberg. Online convex optimization with unbounded memory. *Advances in Neural Information Processing Systems*, 36:26229–26270, 2023.

[42] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022.

[43] Zhou Lu, Y Jennifer Sun, and Zhiyu Zhang. Population dynamics control with partial observations. *arXiv preprint arXiv:2502.14079*, 2025.

[44] Jason R Marden and Jeff S Shamma. Game theory and distributed control. In *Handbook of game theory with economic applications*, volume 4, pages 861–899. Elsevier, 2015.

[45] Jason R Marden and Jeff S Shamma. Game theory and control. *Annual review of control, robotics, and autonomous systems*, 1(1):105–134, 2018.

[46] Andrea Martin, Luca Furieri, Florian Dörfler, John Lygeros, and Giancarlo Ferrari-Trecate. Regret optimal control for uncertain stochastic systems. *European Journal of Control*, 80:101051, 2024.

[47] Daniele Martinelli, Andrea Martin, Giancarlo Ferrari-Trecate, and Luca Furieri. Closing the gap to quadratic invariance: A regret minimization approach to optimal distributed control. In *2024 European Control Conference (ECC)*, pages 756–761, 2024.

[48] Vladimir V Mazalov, Anna N Rettieva, and Konstantin E Avrachenkov. Linear-quadratic discrete-time dynamic potential games. *Automation and Remote Control*, 78:1537–1544, 2017.

[49] Eric Mazumdar, Lillian J. Ratliff, Michael I. Jordan, and S. Shankar Sastry. Policy-gradient algorithms have no guarantees of convergence in linear quadratic games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020.

[50] Qing-xin Meng and Jian-wei Liu. Proximal point method for online saddle point problem. *arXiv preprint arXiv:2407.04591*, 2024.

[51] Panayotis Mertikopoulos and Mathias Staudigl. Equilibrium tracking and convergence in dynamic games. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 930–935. IEEE, 2021.

[52] Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. Online control of unknown time-varying dynamical systems. *Advances in Neural Information Processing Systems*, 34:15934–15945, 2021.

[53] Michael Muehlebach, Zhiyu He, and Michael I Jordan. The sample complexity of online reinforcement learning: A multi-model perspective. *arXiv preprint arXiv:2501.15910*, 2025.

[54] Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. *Advances in Neural Information Processing Systems*, 33:18318–18329, 2020.

[55] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

[56] Y. Jennifer Sun and Zhou Lu. Tight rates for bandit control beyond quadratics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[57] Youbang Sun, Tao Liu, Ruida Zhou, PR Kumar, and Shahin Shahrampour. Provably fast convergence of independent natural policy gradient for markov potential games. *Advances in neural information processing systems*, 36:43951–43971, 2023.

[58] Feras Al Taha, Kiran Rokade, and Francesca Parise. Learning in time-varying monotone network games with dynamic populations. *arXiv preprint arXiv:2408.06253*, 2024.

[59] Anastasios Tsiamis, Aren Karapetyan, Yueshan Li, Efe C. Balta, and John Lygeros. Predictive linear online tracking for unknown targets. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[60] Muhammad Aneeq uz Zaman, Shubham Aggarwal, Melih Bastopcu, and Tamer Basar. Policy optimization finds nash equilibrium in regularized general-sum lq games. *IEEE CDC*, 2024.

[61] Jiduan Wu, Anas Barakat, Ilyas Fatkhullin, and Niao He. Learning zero-sum linear quadratic games with improved sample complexity. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 2602–2609, 2023.

[62] Yu-Hu Yan, Peng Zhao, and Zhi-Hua Zhou. Fast rates in time-varying strongly monotone games. In *International Conference on Machine Learning*, pages 39138–39164. PMLR, 2023.

[63] Yu-Hu Yan, Peng Zhao, and Zhi-Hua Zhou. Online non-stochastic control with partial feedback. *Journal of Machine Learning Research*, 24(273):1–50, 2023.

[64] Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. *Advances in Neural Information Processing Systems*, 32, 2019.

[65] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International conference on machine learning*, pages 5872–5881. PMLR, 2018.

[66] Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in neural information processing systems*, 34:2949–2964, 2021.

[67] Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. In *International Conference on Machine Learning*, pages 26772–26808. PMLR, 2022.

[68] Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. *Advances in Neural Information Processing Systems*, 35:1923–1935, 2022.

[69] Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *IEEE Transactions on Automatic Control*, 69(10):6499–6514, 2024.

[70] Hongyu Zhou and Vasileios Tzoumas. Simultaneous system identification and model predictive control with no dynamic regret. *arXiv preprint arXiv:2407.04143*, 2024.

Contents

1 Introduction **1**

2 Problem Formulation: Online Learning in Linear Dynamical Games **3**

2.1 Online learning setting and feedback models 4

2.2 Regret framework for online learning in linear dynamical games 4

3 Individual Regret Guarantees **5**

3.1 Information setting 1: State-only Feedback 7

3.2 Information setting 2: Aggregate-Input Feedback 7

3.3 Regret lower bounds 8

4 Equilibrium Tracking in the Common Interest Setting **9**

5 Conclusion and Future Work **10**

A Extended Related Work Discussion **17**

B Examples **18**

B.1 Description 18

B.2 About Adversarial Disturbances 18

C Further Discussion of Assumptions **19**

C.1 Assumption 2 19

C.2 Assumption 3 19

C.3 Assumption 4 19

D Preparatory Results for the Main Proofs **19**

D.1 Notation: counterfactual and idealized states and actions 19

D.2 State evolution 20

D.3 Transfer matrix bound 23

D.4 State, action and difference of state and action bounds 24

E Proof of Theorem 1 **28**

F Proof of Theorem 2 **28**

F.1 Regret decomposition and proof overview 29

F.2 Counterfactual state and action deviation error 30

F.3 Online gradient descent with memory regret bound 31

F.3.1 Coordinate-wise loss Lipschitzness 32

F.3.2 Gradient loss boundedness 32

G Proof of Theorem 3 **32**

H Proof of Theorem 4 **35**

I Proof of Theorem 5 **40**

I.1 Proof of Proposition 3 41

I.2 Proof of Proposition 4 42

I.3 Proof of Proposition 5 43

I.4 Proof of Proposition 6 45

I.5 Proof of Lemma 1 45

J Tools from Online Convex Optimization **46**

J.1 Online convex optimization with memory 46

J.2 Time regret decomposition 47

K Simulations **48**

K.1 Setting 48

K.2 Comments 48

K.3 Additional experiments regarding stability 51

A Extended Related Work Discussion

Online non-stochastic control. Our work builds on a recent and growing line of research focusing on the use of online learning techniques to address control problems with adversarially perturbed dynamical systems [35, 1, 2, 37, 24, 55, 54, 30, 27, 46]. We refer the reader to a nice introduction to the topic in the recent monograph of [38] and the references therein for a survey. Recent follow-up works include studies on dynamic regret for online tracking [59], performative control [10], online control in population dynamics [29, 43], simultaneous system identification and MPC with regret guarantees [70], online RL [53, 27], partial feedback settings [63] and bandit settings [56] to name a few. Most of the works in this line of research are devoted to the control of linear dynamical systems influenced by a *single* controller. We discuss a few exceptions in the next section.

Multi-agent control. The interface between game theory and control has given rise to a large body of work over the last decades to study settings involving multiple interacting controllers, see e.g.[44, 45, 15] for relevant surveys. Within the game-theoretic control literature, linear-quadratic (LQ) games is one of the canonical benchmark problems which has been studied in a variety of settings including LQ differential games [8, Chap. 6], LQ potential games [48, 39], zero-sum LQ games [64, 9, 66, 61], static two-player quadratic games [11] and general-sum LQ games [60, 49, 33, 16, 32]. Some of these works typically consider the same (LDS) and assume quadratic costs for systems which are either deterministic ($w_t = 0$) or perturbed by a noise sequence $\{w_t\}$ which is i.i.d. Gaussian. In particular they do not adopt the online learning perspective and do not address the case of arbitrary disturbances. Classical approaches to design robust controllers in the optimal control literature rely either on using statistical and probabilistic models for disturbances such as for linear quadratic Gaussian design, or adopting a (worst-case) game theoretic perspective via designing 'minimax' controllers like in H_∞ control [7]. Only few recent works adopt an online learning perspective for *distributed* control [28, 13, 14, 47]. [13, 14] studied a distributed online control problem over a multi-agent network of m identical linear time-invariant systems in the presence of adversarial perturbations. Each agent seeks to generate a control sequence that can compete with the best centralized control policy in hindsight. In contrast, we address a multi-agent setting involving *strategic* agents influencing a *single* linear dynamical system. Our state dynamics are not separable and are influenced by all the agents. The cost of each agent in our model is influenced by the (shared) observed state which is governed by all the agents' control inputs and the goal of each agent is to minimize their own individual cost.

Markov Games. Regret bounds have been previously established for discrete finite Markov games. Our multi-agent linear control setting can be seen as a continuous analog to Markov games. However, note that our linear dynamical system is fundamentally different from the usual Markov game (or stochastic game) setting involving an unknown state transition kernel outputting the next state probability distribution as a function of the current state and the (joint) actions of all players. When considering multi-agent potential games, there are three important distinctions with existing works on Markov potential games (e.g. [42, 69, 19, 68, 57]):

- In our work, the state and action spaces are continuous and are not mixed extensions of finite sets of states and actions. Most of the bounds scale with the cardinality of the action spaces of the players and are therefore vacuous in our continuous action space setting. In addition, our results use a suitable control policy for the linear dynamical system setting. The finite-action softmax parameterizations in these works [68, 57] do not directly apply to our continuous control spaces.
- Our results consider adversarial disturbances, and hence the state transitions of the underlying system may not even be Markovian or stochastic, the disturbances can be chosen adversarially depending on the far past.
- Our work considers cost functions that are time-varying, which is in contrast with the standard fixed reward setting in the mentioned Markov potential games works. We also do not consider discounted rewards, and the potential assumption we use is with respect to the cost function itself, and not on the aggregate cost over a time horizon.

Decentralized Multi-Agent RL. Our work is also complementary to decentralized Multi-Agent RL (e.g. [65]), but our setting is different from several aspects. First, our setting is adversarial/nonstochastic: disturbances (and hence state transitions) need not be stochastic or Markovian and may depend adversarially on the history. Second, our model uses continuous state and action spaces, whereas much of the finite-time MARL theory is developed for finite-state/finite-action Markov

games, with bounds that scale in the cardinality of the state/action spaces. Third, the objectives are different: we study a game-theoretic online control problem with agent-specific time-varying general convex costs, derive individual regret guarantees, and in the common-interest case prove equilibrium-tracking guarantees. By contrast, [65] focuses on convergence for a single globally averaged return objective across agents.

B Examples

B.1 Description

We provide a few concrete examples to illustrate our linear dynamical game setting.

(a) Smart grid markets. In modern power grids, electricity is generated and distributed by a mix of independent energy producers such as traditional plants and renewable energy providers. These actors act selfishly and adapt to market conditions while they also jointly influence the grid. Let x_t be the grid state defined by characteristics such as line loads and aggregate reserves, let u_t^i be generator i's power output decision (i.e. their control input) and let the sequence w_t capture the demand fluctuation, the system noise and/or renewable energy shocks. Then, the system dynamics may evolve according to (LDS) (e.g. by linearization around an operating point). Each generator i has their local cost function which accounts for the cost of production including e.g. fuel and a penalty for deviating from a target grid state.

(b) Formation control. Consider a multi-agent system consisting of N vehicles or robots. The state (position, velocity) and control input of each agent i at each time step t are respectively given by x_t^i and u_t^i . Suppose the (joint) state of the multi-agent system evolves according to (LDS). The formation of the multi-agent system is defined by specifying a desired distance to be maintained over time between the states of agents that are adjacent. The goal of each agent is to minimize their own formation error and energy consumption. A similar formation control problem has been studied in the control literature *in the absence of adversarial perturbations* ($w_t = 0$) using differential games (see e.g. [3, 34]) and discrete linear quadratic games [39].

(c) Bioresource management. A set of firms (or countries) exploit a set of renewable resources (e.g. a fish population) whose evolution is driven by (LDS) where $x_t \in \mathbb{R}^d$ denotes the vector of quantities of d distinct resources, the matrix A encodes their natural growth rate, the control u_t^i models the exploitation rate of the firm i and w_t refers to perturbations due to exogenous factors such as weather conditions. Each firm i has the goal to maximize their profit while minimizing their exploitation cost. See e.g. [48] in the noiseless setting ($w_t = 0$).

B.2 About Adversarial Disturbances

In multi-agent systems, considering adversarial disturbances allows us to model a wide range of realistic, worst-case, or strategically motivated perturbations ranging from strategic behavior in energy markets to adversarial environments in robotics and ecological shocks in resource management, ensuring system robustness even under hostile or extreme scenarios.

We provide below examples of adversarial disturbances in each of the examples described in section B.1 above and comment on their importance:

- **Smart grid markets:** An adversarial disturbance could model sudden demand spikes, strategic demand manipulation by large consumers (i.e. major electricity buyers who have significant influence over the overall demand on the grid), malicious data injection attacks that falsify renewable generation forecasts or misreporting. For instance, an actor might manipulate demand predictions to influence market prices or grid loads in their favor.
- **Formation control:** Adversarial disturbances capture environmental disturbances with structured worst-case behavior, such as wind gusts or magnetic interference that affect formations in potentially harmful ways. It can also capture adversarial agents or spoofed sensor data to destabilize the formation. In hostile or uncertain environments (e.g., surveillance drones in contested airspace), agents must maintain formation despite external influences that could intentionally disrupt coordination.
- **Bioresource management:** Adversarial disturbances may reflect deliberate misinformation about resource levels, illegal over-harvesting by untracked actors, or policy shocks (e.g., sudden trade

bans) that drastically affect the resource dynamics in a harmful way. Robust resource management must consider these disturbances to avoid collapse or irreversible damage.

C Further Discussion of Assumptions

C.1 Assumption 2

Bounded disturbances are standard in nonstochastic online control. See [40] for a recent work allowing unbounded noise. As for the boundedness of the control inputs, the projection of policy parameters in the algorithm bounds the learned DAC matrices but it does not by itself bound the complete control input u_t^j. In particular, agent-wise strong stability of each $A - B_i K_i$ does not imply stability of the joint closed-loop matrix $A - \sum_{i=1}^{N} B_i K_i$. Relaxing this restrictive boundedness condition would require for instance an additional joint closed-loop boundedness or small-gain condition.

C.2 Assumption 3

This stability assumption requires each agent to have access to an agent-wise strongly stable linear controller. We make three important remarks regarding this assumption:

- This is a multi-agent analogue of the standard stabilizing-controller assumption used throughout the online/nonstochastic control literature.
- As is standard in prior work on single-agent online control, we assume that agents have initial access to a stabilizing controller. Note that such controllers can be obtained offline using an SDP relaxation (e.g., using the method of [17]). Our main focus is on the challenging task of learning DAC policy parameters under adversarial disturbances.
- Strong stability gives the uniform trajectory bounds used in our analysis. Without suitable stability, state norms and cumulative costs can grow exponentially with the horizon, so the analysis used here is unavailable in general.

C.3 Assumption 4

Individual and joint strong stability are distinct requirements. For example, with scalar matrices $A = 0$, $B_1 = B_2 = 1$ and $K_1 = K_2 = 3/4$, both individual closed-loop matrices equal $-3/4$, whereas the joint closed-loop matrix equals $-3/2$ and is unstable. Conversely, $A = 2$, $B_1 = B_2 = 1$, $K_1 = 0$ and $K_2 = 2$ give a stable joint closed-loop matrix but an unstable individual matrix $A - B_1 K_1 = 2$.

Theorems 2 and 3 use joint strong stability to bound the realized trajectory and individual strong stability of the same supplied baselines to control unilateral counterfactuals with opponents' realized controls held fixed. Theorem 5 uses joint strong stability for its joint truncated rollouts. Its proof does not establish the same result from individual stability alone. Theorem 1 instead assumes a bound on the realized opponent inputs.

D Preparatory Results for the Main Proofs

D.1 Notation: counterfactual and idealized states and actions

Fix a player $i \in [N]$ and the realized control sequences of its opponents. Define

$$A_i := A - B_i K_i, \qquad d_s := w_s + \sum_{j \neq i} B_j u_s^j, \quad \forall s \geq 0.$$

The DAC policy uses the effective disturbance sequence $r_s = d_s$ under Information Setting 1 and $r_s = w_s$ under Information Setting 2. Set $d_s = r_s = 0$ for $s < 0$. Both sequences are held fixed when varying the unilateral comparator. In particular, opponents' realized controls are not recomputed on the counterfactual state.

For a feasible parameter sequence $Q_{0:T}$, the true unilateral counterfactual is defined by

$$\begin{aligned} x_0^Q &= x_0, \\ u_s^Q &= -K_i x_s^Q + \sum_{p=1}^{H} Q_s^{[p-1]} r_{s-p}, \\ x_{s+1}^Q &= A_i x_s^Q + B_i \sum_{p=1}^{H} Q_s^{[p-1]} r_{s-p} + d_s. \end{aligned} \tag{9}$$

For a fixed $Q \in \mathcal{M}_i$, the same notation refers to the stationary sequence $Q_s = Q$. Equivalently, these stationary trajectories can be denoted by $x_t^{K_i}(Q), u_t^{i,K_i}(Q)$. The time-zero DAC correction is zero, so every DAC comparator with this fixed baseline applies $u_0^Q = -K_i x_0$. A linear comparator with a different baseline K_i^* instead applies $-K_i^* x_0$.

Write $h = H + 1$. For every $t \geq 1$, define the zero-initialized finite-memory state, action, and loss by

$$y_t^i(Q_{t-h:t-1}) = \sum_{a=1}^{h} A_i^{a-1} \left(d_{t-a} + B_i \sum_{p=1}^{H} Q_{t-a}^{[p-1]} r_{t-a-p} \right), \tag{10}$$

$$v_t^i(Q_{t-h:t}) = -K_i y_t^i(Q_{t-h:t-1}) + \sum_{p=1}^{H} Q_t^{[p-1]} r_{t-p}, \tag{11}$$

$$\ell_t^i(Q_{t-h:t}) = c_t^i\big(y_t^i(Q_{t-h:t-1}), v_t^i(Q_{t-h:t})\big), \tag{12}$$

$$f_t^i(Q) := \ell_t^i(Q, \ldots, Q). \tag{13}$$

The single-argument notation $\ell_t^i(Q)$ is also used interchangeably with the diagonal function $f_t^i(Q)$. Similarly, the notation y_t^{i,K_i}, v_t^{i,K_i} denotes y_t^i, v_t^i when the baseline needs to be displayed explicitly.

Negative-time parameter values can be set to the initial feasible parameter: they multiply zero-padded disturbance values. These definitions therefore also apply to early rounds $1 \leq t < h$. For $t \geq h$, the truncation starts at time $t - h$ with state zero and includes the h transitions through time $t - 1$; the ideal action additionally uses the current parameter Q_t. The memory loss consequently has $h + 1 = H + 2$ parameter arguments. The ideal state and action are affine in these arguments. Thus ℓ_t^i is jointly convex and its diagonal f_t^i is convex under Assumption 1. Algorithm 1 uses ∇f_t^i.

D.2 State evolution

We record both individual and joint state expansions. In this subsection m denotes a general unrolling index, so an expansion with index m contains $m + 1$ transitions. For integers $t \geq 0$, $0 \leq m \leq t$, and $0 \leq l \leq H + m$, define

$$\widetilde{A}_{K_i} := A - B_i K_i, \qquad \Psi_{t,l}^{i,m}(M_{i,t-m:t}) := \widetilde{A}_{K_i}^l \mathbf{1}_{\{l \leq m\}} + \sum_{k=0}^{m} \widetilde{A}_{K_i}^k B_i M_{i,t-k}^{[l-k-1]} \mathbf{1}_{\{1 \leq l-k \leq H\}}, \tag{14}$$

$$\bar{A}_K := A - \sum_{i=1}^{N} B_i K_i, \qquad \bar{\Psi}_{t,l}^{m}(M_{t-m:t}) := \bar{A}_K^l \mathbf{1}_{\{l \leq m\}} + \sum_{k=0}^{m} \bar{A}_K^k \sum_{i=1}^{N} B_i M_{i,t-k}^{[l-k-1]} \mathbf{1}_{\{1 \leq l-k \leq H\}}. \tag{15}$$

In these transfer-matrix formulas only, blocks outside $\{0, \ldots, H-1\}$ are interpreted as zero. The matrices $\Psi_{t,l}^{i,m}$ and $\bar{\Psi}_{t,l}^{m}$ are affine in their DAC parameter arguments.

Proposition 1 (State evolution)**.** *For integers $t \geq 0$ and $0 \leq m \leq t$, the following hold.*

(i) Suppose player i uses a nonstationary DAC policy with $\widetilde{w}_s := w_s + \sum_{j \neq i} B_j u_s^j$, while opponents' realized controls are fixed. Then

$$x_{t+1} = \widetilde{A}_{K_i}^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \Psi_{t,l}^{i,m}(M_{i,t-m:t}) \widetilde{w}_{t-l}.$$

(ii) Suppose all agents use nonstationary DAC policies driven by w_s. Then the actual joint state satisfies

$$x_{t+1} = \bar{A}_K^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \bar{\Psi}_{t,l}^{m}(M_{t-m:t}) w_{t-l}. \tag{16}$$

Proof. We prove the two claims separately. Recall that disturbances are extended by zero to negative indices: $w_s = \widetilde{w}_s = 0$ for $s < 0$. All empty sums below are interpreted as zero.

We first record the elementary unrolling identity used in both claims. For a recursion of the form

$$z_{s+1} = C z_s + b_s,$$

we have, for every $0 \le m \le t$,

$$z_{t+1} = C^{m+1} z_{t-m} + \sum_{k=0}^{m} C^k b_{t-k}. \tag{17}$$

Indeed, the case $m = 0$ is the defining recursion. If the identity holds with $m = r < t$, substituting $z_{t-r} = C z_{t-r-1} + b_{t-r-1}$ gives

$$z_{t+1} = C^{r+1}\big(C z_{t-r-1} + b_{t-r-1}\big) + \sum_{k=0}^{r} C^k b_{t-k} = C^{r+2} z_{t-r-1} + \sum_{k=0}^{r+1} C^k b_{t-k},$$

which proves (17) by induction.

Proof of Claim (i). For $s \ge 0$, player i uses the DAC policy:

$$u_s^i = -K_i x_s + \sum_{p=1}^{H} M_{i,s}^{[p-1]} \widetilde{w}_{s-p}.$$

Substituting this expression into the system dynamics and using $\widetilde{w}_s = w_s + \sum_{j \ne i} B_j u_s^j$ yields:

$$\begin{aligned} x_{s+1} &= A x_s + B_i u_s^i + \sum_{j \ne i} B_j u_s^j + w_s \\ &= A x_s + B_i \left(-K_i x_s + \sum_{p=1}^{H} M_{i,s}^{[p-1]} \widetilde{w}_{s-p} \right) + \widetilde{w}_s \\ &= \widetilde{A}_{K_i} x_s + \widetilde{w}_s + B_i \sum_{p=1}^{H} M_{i,s}^{[p-1]} \widetilde{w}_{s-p}, \end{aligned}$$

where $\widetilde{A}_{K_i} = A - B_i K_i$.

Applying (17) with $C = \widetilde{A}_{K_i}$ and $b_s = \widetilde{w}_s + B_i \sum_{p=1}^{H} M_{i,s}^{[p-1]} \widetilde{w}_{s-p}$, we obtain:

$$\begin{aligned} x_{t+1} =& \widetilde{A}_{K_i}^{m+1} x_{t-m} + \sum_{k=0}^{m} \widetilde{A}_{K_i}^{k} \left(\widetilde{w}_{t-k} + B_i \sum_{p=1}^{H} M_{i,t-k}^{[p-1]} \widetilde{w}_{t-k-p} \right) \\ =& \widetilde{A}_{K_i}^{m+1} x_{t-m} + \sum_{k=0}^{m} \widetilde{A}_{K_i}^{k} \widetilde{w}_{t-k} + \sum_{k=0}^{m} \sum_{p=1}^{H} \widetilde{A}_{K_i}^{k} B_i M_{i,t-k}^{[p-1]} \widetilde{w}_{t-k-p}. \end{aligned}$$

We now collect the terms involving the same disturbance. For the direct disturbance contribution, simply rename k as l:

$$\sum_{k=0}^{m} \widetilde{A}_{K_i}^{k} \widetilde{w}_{t-k} = \sum_{l=0}^{H+m} \widetilde{A}_{K_i}^{l} \mathbf{1}_{\{l \le m\}} \widetilde{w}_{t-l}.$$

For the DAC contribution, set $l = k + p$. The conditions $0 \le k \le m$ and $1 \le p \le H$ are equivalent to the following conditions:

$$1 \le l \le H + m, \qquad \max\{0, l - H\} \le k \le \min\{m, l - 1\}.$$

Moreover, $p-1 = l-k-1$. Therefore,

$$\sum_{k=0}^{m}\sum_{p=1}^{H} \widetilde{A}_{K_i}^{k} B_i M_{i,t-k}^{[p-1]} \widetilde{w}_{t-k-p} = \sum_{l=1}^{H+m} \left(\sum_{k=\max\{0,l-H\}}^{\min\{m,l-1\}} \widetilde{A}_{K_i}^{k} B_i M_{i,t-k}^{[l-k-1]} \right) \widetilde{w}_{t-l}.$$

There is no DAC contribution at $l=0$, since $k+p \geq 1$.

Combining the direct and DAC contributions gives:

$$\begin{aligned} x_{t+1} &= \widetilde{A}_{K_i}^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \left(\widetilde{A}_{K_i}^{l} \mathbf{1}_{\{l\leq m\}} + \sum_{\substack{0\leq k\leq m \\ 1\leq l-k\leq H}} \widetilde{A}_{K_i}^{k} B_i M_{i,t-k}^{[l-k-1]} \right) \widetilde{w}_{t-l} \\ &= \widetilde{A}_{K_i}^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \Psi_{t,l}^{i,m}(M_{i,t-m:t}) \widetilde{w}_{t-l}, \end{aligned}$$

where the last equality is the definition of $\Psi_{t,l}^{i,m}$. This proves Claim (i).

Proof of Claim (ii). Each player $i \in [N]$ now uses the disturbance sequence $\{w_s\}$ in its DAC policy:

$$u_s^i = -K_i x_s + \sum_{p=1}^{H} M_{i,s}^{[p-1]} w_{s-p}.$$

Substituting all players' policies into the dynamics yields:

$$\begin{aligned} x_{s+1} &= A x_s + \sum_{i=1}^{N} B_i u_s^i + w_s \\ &= A x_s + \sum_{i=1}^{N} B_i \left(-K_i x_s + \sum_{p=1}^{H} M_{i,s}^{[p-1]} w_{s-p} \right) + w_s \\ &= \left(A - \sum_{i=1}^{N} B_i K_i \right) x_s + w_s + \sum_{i=1}^{N} B_i \sum_{p=1}^{H} M_{i,s}^{[p-1]} w_{s-p} \\ &= \bar{A}_K x_s + w_s + \sum_{i=1}^{N} B_i \sum_{p=1}^{H} M_{i,s}^{[p-1]} w_{s-p}. \end{aligned}$$

In particular, the direct disturbance w_s appears once, outside the sum over players.

Applying (17) with $C = \bar{A}_K$ and $b_s = w_s + \sum_{i=1}^{N} B_i \sum_{p=1}^{H} M_{i,s}^{[p-1]} w_{s-p}$ gives:

$$\begin{aligned} x_{t+1} &= \bar{A}_K^{m+1} x_{t-m} + \sum_{k=0}^{m} \bar{A}_K^{k} \left(w_{t-k} + \sum_{i=1}^{N} B_i \sum_{p=1}^{H} M_{i,t-k}^{[p-1]} w_{t-k-p} \right) \\ &= \bar{A}_K^{m+1} x_{t-m} + \sum_{k=0}^{m} \bar{A}_K^{k} w_{t-k} + \sum_{k=0}^{m}\sum_{i=1}^{N}\sum_{p=1}^{H} \bar{A}_K^{k} B_i M_{i,t-k}^{[p-1]} w_{t-k-p}. \end{aligned}$$

As in the first claim, the direct disturbance contribution is

$$\sum_{k=0}^{m} \bar{A}_K^{k} w_{t-k} = \sum_{l=0}^{H+m} \bar{A}_K^{l} \mathbf{1}_{\{l\leq m\}} w_{t-l}.$$

For the DAC contributions, use the same change of indices $l = k+p$. The finite sum over players does not change the admissible range of k, so

$$\sum_{k=0}^{m}\sum_{i=1}^{N}\sum_{p=1}^{H} \bar{A}_K^{k} B_i M_{i,t-k}^{[p-1]} w_{t-k-p} = \sum_{l=1}^{H+m} \left(\sum_{k=\max\{0,l-H\}}^{\min\{m,l-1\}} \bar{A}_K^{k} \sum_{i=1}^{N} B_i M_{i,t-k}^{[l-k-1]} \right) w_{t-l}.$$

Combining these expressions, and extending the DAC sum to $l = 0$ using the empty-sum convention, we obtain:

$$x_{t+1} = \bar{A}_K^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \left(\bar{A}_K^l \mathbf{1}_{\{l \le m\}} + \sum_{\substack{0 \le k \le m \\ 1 \le l-k \le H}} \bar{A}_K^k \sum_{i=1}^{N} B_i M_{i,t-k}^{[l-k-1]} \right) w_{t-l}$$
$$= \bar{A}_K^{m+1} x_{t-m} + \sum_{l=0}^{H+m} \bar{\Psi}_{t,l}^m(M_{t-m:t}) w_{t-l}.$$

The last equality follows from the definition of $\bar{\Psi}_{t,l}^m$ and proves (16). □

D.3 Transfer matrix bound

Lemma 2. *Let Assumption 4 hold and put $q = 1 - \bar{\gamma}$ and $B_\Sigma = \sum_{i=1}^N \|B_i\|_{\mathrm{op}}$. Suppose that $\|M_{i,s}^{[p]}\|_{\mathrm{op}} \le \tau q^p$ for some $\tau > 0$, every $s \ge 0$, and $0 \le p < H$. For integers $t \ge 0$, $0 \le m \le t$, and $0 \le l \le H + m$,*

$$\|\bar{\Psi}_{t,l}^m(M_{t-m:t})\|_{\mathrm{op}} \le \bar{\kappa} q^l \mathbf{1}_{\{l \le m\}} + \mathbf{1}_{\{l \ge 1\}} H \bar{\kappa} \tau B_\Sigma q^{l-1}.$$

Proof. Recall from (15) that the state transfer matrix is

$$\bar{\Psi}_{t,l}^m(M_{t-m:t}) = \bar{A}_K^l \mathbf{1}_{\{l \le m\}} + \sum_{\substack{0 \le k \le m \\ 1 \le l-k \le H}} \bar{A}_K^k \sum_{i=1}^{N} B_i M_{i,t-k}^{[l-k-1]}, \tag{18}$$

where $\bar{A}_K = A - \sum_{i=1}^N B_i K_i$. We have written the second term as a restricted sum so that every DAC block appearing in the expression has an index in $\{0, \dots, H-1\}$. This is equivalent to the indicator notation in (15). Throughout the proof, an empty sum is equal to zero.

By Assumption 4, there exist an invertible matrix Q and a matrix L such that:

$$\bar{A}_K = QLQ^{-1}, \qquad \|Q\|_{\mathrm{op}} \|Q^{-1}\|_{\mathrm{op}} \le \bar{\kappa}, \qquad \|L\|_{\mathrm{op}} \le 1 - \bar{\gamma} = q.$$

Consequently, for every integer $a \ge 0$, submultiplicativity of the operator norm gives

$$\|\bar{A}_K^a\|_{\mathrm{op}} = \|QL^aQ^{-1}\|_{\mathrm{op}} \le \|Q\|_{\mathrm{op}} \|L^a\|_{\mathrm{op}} \|Q^{-1}\|_{\mathrm{op}} \le \|Q\|_{\mathrm{op}} \|Q^{-1}\|_{\mathrm{op}} \|L\|_{\mathrm{op}}^a \le \bar{\kappa} q^a. \tag{19}$$

This bound also holds for $a = 0$, since $1 = \|I_d\|_{\mathrm{op}} \le \|Q\|_{\mathrm{op}} \|Q^{-1}\|_{\mathrm{op}} \le \bar{\kappa}$.

If $l = 0$, there is no integer $k \ge 0$ satisfying $1 \le l - k \le H$. Thus the DAC contribution in (18) is zero, and $\bar{\Psi}_{t,0}^m(M_{t-m:t}) = \bar{A}_K^0 = I_d$. It follows that $\|\bar{\Psi}_{t,0}^m(M_{t-m:t})\|_{\mathrm{op}} = 1 \le \bar{\kappa}$, which is exactly the claimed bound when $l = 0$.

We are now ready to bound the transfer matrix for $l \ge 1$. Fix now an integer $1 \le l \le H + m$. For every index k included in the restricted sum,

$$t - k \ge t - m \ge 0, \qquad 0 \le l - k - 1 \le H - 1.$$

The assumed DAC block bound therefore applies and gives $\|M_{i,t-k}^{[l-k-1]}\|_{\mathrm{op}} \le \tau q^{l-k-1}$. Using the triangle inequality and submultiplicativity in (18), we obtain:

$$\|\bar{\Psi}_{t,l}^m(M_{t-m:t})\|_{\mathrm{op}} \le \|\bar{A}_K^l\|_{\mathrm{op}} \mathbf{1}_{\{l \le m\}} + \sum_{\substack{0 \le k \le m \\ 1 \le l-k \le H}} \|\bar{A}_K^k\|_{\mathrm{op}} \sum_{i=1}^{N} \|B_i\|_{\mathrm{op}} \|M_{i,t-k}^{[l-k-1]}\|_{\mathrm{op}}$$
$$\le \bar{\kappa} q^l \mathbf{1}_{\{l \le m\}} + \bar{\kappa}\tau \sum_{i=1}^{N} \|B_i\|_{\mathrm{op}} \sum_{\substack{0 \le k \le m \\ 1 \le l-k \le H}} q^k q^{l-k-1}$$
$$= \bar{\kappa} q^l \mathbf{1}_{\{l \le m\}} + \bar{\kappa}\tau B_\Sigma q^{l-1} \sum_{\substack{0 \le k \le m \\ 1 \le l-k \le H}} 1.$$

Here, the second inequality uses (19) and the DAC block bound. In particular, after these bounds are applied, each admissible index k contributes the same upper bound.

The conditions $0 \le k \le m$ and $1 \le l-k \le H$ are equivalent to $\max\{0, l-H\} \le k \le \min\{m, l-1\}$. Every such index therefore belongs to $\{l-H, l-H+1, \dots, l-1\}$, which contains exactly H integers. Intersecting this set with $\{0, \dots, m\}$ cannot increase its cardinality. Hence $\sum_{\substack{0\le k\le m \\ 1\le l-k\le H}} 1 \le H$. Substituting this estimate into the preceding bound gives, for every $l \ge 1$ in the stated range,

$$\|\bar{\Psi}^m_{t,l}(M_{t-m:t})\|_{\mathrm{op}} \le \bar{\kappa} q^l \mathbf{1}_{\{l\le m\}} + H\bar{\kappa}\tau B_\Sigma q^{l-1}.$$

Combining this inequality with the case $l = 0$ proves:

$$\|\bar{\Psi}^m_{t,l}(M_{t-m:t})\|_{\mathrm{op}} \le \bar{\kappa} q^l \mathbf{1}_{\{l\le m\}} + \mathbf{1}_{\{l\ge 1\}} H\bar{\kappa}\tau B_\Sigma q^{l-1},$$

as claimed. □

D.4 State, action and difference of state and action bounds

The goal of the next proposition is to control the norms of states, actions and differences of states and actions. Note that we pay particular attention to the problem constants involved to elucidate the dependence of our bounds on the number of agents N and the magnitude of the control inputs of all the agents. The result is a more refined version of [2, Lemma 5.5] which is adapted to our multi-agent control setting when each agent executes a (DAC-i) policy.

Proposition 2. *Let Assumptions 2 and 4 hold, and suppose $\|x_0\|_2 \le X_0$. Let all agents use DAC policies with $\|M^{[p]}_{i,s}\|_{\mathrm{op}} \le \tau q^p$, where $q = 1-\bar{\gamma}$, $\tau > 0$, and $0 \le p < H$. Set*

$$m_H := \tau \sum_{p=0}^{H-1} q^p, \qquad B_\Sigma := \sum_{i=1}^{N} \|B_i\|_{\mathrm{op}}, \qquad Y := \frac{\bar{\kappa} W}{\bar{\gamma}}(1 + B_\Sigma m_H), \qquad X := \bar{\kappa} X_0 + Y.$$

With negative-time disturbances set to zero, define the joint ideal state and action by

$$y^K_t := \sum_{a=1}^{H+1} \bar{A}_K^{a-1} \left(w_{t-a} + \sum_{j=1}^{N} B_j \sum_{p=1}^{H} M^{[p-1]}_{j,t-a} w_{t-a-p} \right), \tag{20}$$

$$v^{i,K}_t := -K_i y^K_t + \sum_{p=1}^{H} M^{[p-1]}_{i,t} w_{t-p}. \tag{21}$$

Negative-time parameter values are arbitrary feasible values, since their contributions are zero. Then, we have the following bounds:

(i) For every $t \ge 0$, the actual joint state satisfies

$$\|x_t\|_2 \le X. \tag{22}$$

(ii) For every $t \ge 1$, $\|y^K_t\|_2 \le Y$.

(iii) If K^* is another jointly $(\bar{\kappa}, \bar{\gamma})$-strongly stable linear controller, then

$$\|x^{K^*}_t(0)\|_2 \le \bar{\kappa} X_0 + \bar{\kappa} W/\bar{\gamma} \qquad (t \ge 0).$$

(iv) For every $t \ge 0$, $\|u^i_t\|_2 \le \bar{\kappa} X + W m_H$.

(v) For every $t \ge 1$, $\|v^{i,K}_t\|_2 \le \bar{\kappa} Y + W m_H$.

(vi) For every $t \ge H+1$,

$$\|x_t - y^K_t\|_2 \le \bar{\kappa} q^{H+1} X. \tag{23}$$

(vii) For every $t \ge H+1$, $\|u^i_t - v^{i,K}_t\|_2 \le \bar{\kappa}^2 q^{H+1} X$.

(viii) A common bound for all the displayed states and actions is

$$D := \bar{\kappa}^2 X + W m_H. \tag{24}$$

Moreover, for $t \geq H+1$,

$$\max\{\|x_t - y_t^K\|_2, \|u_t^i - v_t^{i,K}\|_2\} \leq q^{H+1} D. \tag{25}$$

In our setting, the DAC domain uses $\tau = 2\bar{\kappa}^2$. If $X_0, W, \bar{\kappa}, \bar{\gamma}^{-1}$ and $\max_{i\in[N]} \|B_i\|_{\mathrm{op}}$ are uniformly bounded in N, then $X, Y, D = \mathcal{O}(N)$, uniformly in H, since $m_H \leq \tau/\bar{\gamma}$.

Proof. We prove each statement of the proposition separately. Throughout the proof, we use the consequence of joint strong stability established in (19): $\|\bar{A}_K^a\|_{\mathrm{op}} \leq \bar{\kappa} q^a$ for every integer $a \geq 0$. We also use $\|K_i\|_{\mathrm{op}} \leq \bar{\kappa}$ for every $i \in [N]$.

Indeed, each K_i is a block row of the stacked controller $K = (K_1^\top, \ldots, K_N^\top)^\top$, and joint strong stability includes the bound $\|K\|_{\mathrm{op}} \leq \bar{\kappa}$.

Proof of Claim (i). Substituting the DAC policies into the state dynamics gives, for every $s \geq 0$,

$$\begin{aligned} x_{s+1} &= Ax_s + \sum_{j=1}^N B_j \left(-K_j x_s + \sum_{p=1}^H M_{j,s}^{[p-1]} w_{s-p} \right) + w_s \\ &= \bar{A}_K x_s + w_s + \sum_{j=1}^N B_j \sum_{p=1}^H M_{j,s}^{[p-1]} w_{s-p}. \end{aligned}$$

For convenience, define:

$$b_s := w_s + \sum_{j=1}^N B_j \sum_{p=1}^H M_{j,s}^{[p-1]} w_{s-p}.$$

Thus the actual joint state satisfies $x_{s+1} = \bar{A}_K x_s + b_s$.

We first bound b_s. For every player j, the block constraints and the disturbance bound give:

$$\left\| \sum_{p=1}^H M_{j,s}^{[p-1]} w_{s-p} \right\|_2 \leq \sum_{p=1}^H \|M_{j,s}^{[p-1]}\|_{\mathrm{op}} \|w_{s-p}\|_2 \leq W\tau \sum_{p=1}^H q^{p-1} = W m_H.$$

The estimate remains valid when some indices $s-p$ are negative, since the corresponding disturbances are zero. Consequently,

$$\|b_s\|_2 \leq \|w_s\|_2 + \sum_{j=1}^N \|B_j\|_{\mathrm{op}} \left\| \sum_{p=1}^H M_{j,s}^{[p-1]} w_{s-p} \right\|_2 \leq W + B_\Sigma W m_H = W(1 + B_\Sigma m_H). \tag{26}$$

For $s < 0$, the zero-padding convention gives $b_s = 0$, so the same bound also holds for negative indices.

Unrolling the actual recurrence from time zero yields:

$$x_t = \bar{A}_K^t x_0 + \sum_{s=0}^{t-1} \bar{A}_K^{t-1-s} b_s.$$

For $t = 0$, the sum is empty and this identity reduces to $x_0 = x_0$. Using (19), (26), and $\|x_0\|_2 \leq X_0$, we obtain:

$$\begin{aligned} \|x_t\|_2 &\leq \|\bar{A}_K^t\|_{\mathrm{op}} \|x_0\|_2 + \sum_{s=0}^{t-1} \|\bar{A}_K^{t-1-s}\|_{\mathrm{op}} \|b_s\|_2 \\ &\leq \bar{\kappa} q^t X_0 + \bar{\kappa} W (1 + B_\Sigma m_H) \sum_{s=0}^{t-1} q^{t-1-s} \\ &= \bar{\kappa} q^t X_0 + \bar{\kappa} W (1 + B_\Sigma m_H) \frac{1 - q^t}{1 - q} \\ &\leq \bar{\kappa} X_0 + \frac{\bar{\kappa} W}{\bar{\gamma}} (1 + B_\Sigma m_H) \\ &= \bar{\kappa} X_0 + Y = X. \end{aligned}$$

Here we used $1 - q = \bar{\gamma}$ and $0 < q < 1$. This proves (22) for every $t \geq 0$.

Proof of Claim (ii). By the definition of the joint ideal state in (20),

$$y_t^K = \sum_{a=1}^{H+1} \bar{A}_K^{a-1} b_{t-a}.$$

There is no initial-state contribution in this expression because the ideal rollout starts from zero at its truncation boundary. Applying the power bound and (26), we obtain:

$$\begin{aligned}
\|y_t^K\|_2 &\leq \sum_{a=1}^{H+1} \|\bar{A}_K^{a-1}\|_{\mathrm{op}} \, \|b_{t-a}\|_2 \\
&\leq \bar{\kappa} W (1 + B_\Sigma m_H) \sum_{a=1}^{H+1} q^{a-1} \\
&= \bar{\kappa} W (1 + B_\Sigma m_H) \frac{1 - q^{H+1}}{1 - q} \\
&\leq \frac{\bar{\kappa} W}{\bar{\gamma}} (1 + B_\Sigma m_H) = Y.
\end{aligned}$$

This argument applies for every $t \geq 1$. In particular, during early rounds, terms with $t - a < 0$ vanish because $b_{t-a} = 0$.

Proof of Claim (iii). Let

$$\bar{A}_{K^*} := A - \sum_{j=1}^{N} B_j K_j^*.$$

When all players use the linear controllers K_j^*, the state satisfies

$$x_{s+1}^{K^*}(0) = \bar{A}_{K^*} x_s^{K^*}(0) + w_s, \qquad x_0^{K^*}(0) = x_0.$$

Here the argument 0 denotes zero DAC correction; the initial state remains x_0.

Unrolling this recurrence gives:

$$x_t^{K^*}(0) = \bar{A}_{K^*}^t x_0 + \sum_{s=0}^{t-1} \bar{A}_{K^*}^{t-1-s} w_s.$$

Since K^* is jointly $(\bar{\kappa}, \bar{\gamma})$-strongly stable, its closed-loop matrix satisfies $\|\bar{A}_{K^*}^a\|_{\mathrm{op}} \leq \bar{\kappa} q^a$ for all $a \geq 0$. Therefore,

$$\begin{aligned}
\|x_t^{K^*}(0)\|_2 &\leq \|\bar{A}_{K^*}^t\|_{\mathrm{op}} \cdot \|x_0\|_2 + \sum_{s=0}^{t-1} \|\bar{A}_{K^*}^{t-1-s}\|_{\mathrm{op}} \cdot \|w_s\|_2 \\
&\leq \bar{\kappa} q^t X_0 + \bar{\kappa} W \sum_{s=0}^{t-1} q^{t-1-s} \\
&\leq \bar{\kappa} X_0 + \frac{\bar{\kappa} W}{\bar{\gamma}}.
\end{aligned}$$

The same calculation includes $t = 0$ through the empty sum convention.

Proof of Claim (iv). The actual action of player i is:

$$u_t^i = -K_i x_t + \sum_{p=1}^{H} M_{i,t}^{[p-1]} w_{t-p}. \tag{27}$$

By the triangle inequality, the controller norm bound, and Claim (i),

$$\|u_t^i\|_2 \leq \|K_i\|_{\mathrm{op}} \|x_t\|_2 + \sum_{p=1}^{H} \|M_{i,t}^{[p-1]}\|_{\mathrm{op}} \|w_{t-p}\|_2 \leq \bar{\kappa} X + W \tau \sum_{p=1}^{H} q^{p-1} = \bar{\kappa} X + W m_H.$$

This holds for every $t \geq 0$. At $t = 0$, all disturbances in the DAC correction are zero, so the action is simply $u_0^i = -K_i x_0$.

Proof of Claim (v). By (21), the ideal action is given by:

$$v_t^{i,K} = -K_i y_t^K + \sum_{p=1}^{H} M_{i,t}^{[p-1]} w_{t-p}.$$

Using Claim (ii) in place of the actual-state bound, we obtain for every $t \geq 1$,

$$\begin{aligned}\|v_t^{i,K}\|_2 &\leq \|K_i\|_{\mathrm{op}} \cdot \|y_t^K\|_2 + \sum_{p=1}^{H} \|M_{i,t}^{[p-1]}\|_{\mathrm{op}} \cdot \|w_{t-p}\|_2 \\ &\leq \bar{\kappa} Y + W\tau \sum_{p=1}^{H} q^{p-1} \\ &= \bar{\kappa} Y + W m_H \,.\end{aligned}$$

Proof of Claim (vi). Fix $t \geq H + 1$. Unrolling the actual recurrence $x_{s+1} = \bar{A}_K x_s + b_s$ over its last $H + 1$ transitions gives:

$$x_t = \bar{A}_K^{H+1} x_{t-H-1} + \sum_{a=1}^{H+1} \bar{A}_K^{a-1} b_{t-a}. \tag{28}$$

The sum in this expression is exactly the joint ideal state y_t^K defined in (20). In particular, the actual and ideal rollouts use the same joint DAC parameter sequence and the same disturbance values over the truncation window. Subtracting the two expressions therefore yields the exact identity:

$$x_t - y_t^K = \bar{A}_K^{H+1} x_{t-H-1}.$$

It follows that:

$$\|x_t - y_t^K\|_2 \leq \|\bar{A}_K^{H+1}\|_{\mathrm{op}} \cdot \|x_{t-H-1}\|_2 \leq \bar{\kappa} q^{H+1} \|x_{t-H-1}\|_2 \leq \bar{\kappa} q^{H+1} X.$$

The last inequality follows from Claim (i), which applies because $t - H - 1 \geq 0$. This includes the boundary case $t = H + 1$, where the omitted state contribution is $\bar{A}_K^{H+1} x_0$. Thus (23) holds.

Proof of Claim (vii). The actual and ideal actions use the same time-t DAC parameters and disturbance history. Consequently, subtracting (21) from (27) cancels their DAC corrections:

$$u_t^i - v_t^{i,K} = -K_i x_t + \sum_{p=1}^{H} M_{i,t}^{[p-1]} w_{t-p} + K_i y_t^K - \sum_{p=1}^{H} M_{i,t}^{[p-1]} w_{t-p} = -K_i (x_t - y_t^K).$$

Using $\|K_i\|_{\mathrm{op}} \leq \bar{\kappa}$ and Claim (vi), we obtain, for $t \geq H + 1$,

$$\|u_t^i - v_t^{i,K}\|_2 \leq \|K_i\|_{\mathrm{op}} \|x_t - y_t^K\|_2 \leq \bar{\kappa} \|x_t - y_t^K\|_2 \leq \bar{\kappa}^2 q^{H+1} X.$$

Proof of Claim (viii). First, joint strong stability implies $\bar{\kappa} \geq 1$. Indeed, for its similarity matrix Q, $1 = \|I_d\|_{\mathrm{op}} = \|QQ^{-1}\|_{\mathrm{op}} \leq \|Q\|_{\mathrm{op}} \|Q^{-1}\|_{\mathrm{op}} \leq \bar{\kappa}$. Recall that $D = \bar{\kappa}^2 X + W m_H$.

We first check the state bounds. Since $X = \bar{\kappa} X_0 + Y$, we have $Y \leq X$. Moreover,

$$\bar{\kappa} X_0 + \frac{\bar{\kappa} W}{\bar{\gamma}} \leq \bar{\kappa} X_0 + \frac{\bar{\kappa} W}{\bar{\gamma}} (1 + B_\Sigma m_H) = X.$$

Thus every state bound in Claims (i)–(iii) is at most X, and $X \leq \bar{\kappa}^2 X \leq D$.

Next, the action bounds in Claims (iv) and (v) satisfy

$$\bar{\kappa} Y + W m_H \leq \bar{\kappa} X + W m_H \leq \bar{\kappa}^2 X + W m_H = D.$$

Hence D bounds every displayed state and action, within the time ranges specified in the proposition.

Finally, Claims (vi) and (vii), together with $\bar{\kappa} \geq 1$, give:

$$\max\{\|x_t - y_t^K\|_2,\, \|u_t^i - v_t^{i,K}\|_2\} \leq q^{H+1} \max\{\bar{\kappa} X, \bar{\kappa}^2 X\} = q^{H+1}\bar{\kappa}^2 X \leq q^{H+1} D,$$

for every $t \geq H+1$. This proves (25).

To verify the stated population dependence, observe that

$$m_H = \tau \frac{1-q^H}{1-q} \leq \frac{\tau}{\bar{\gamma}}, \qquad B_\Sigma \leq N \max_{j\in[N]} \|B_j\|_{\mathrm{op}}.$$

Therefore, we obtain:

$$Y \leq \frac{\bar{\kappa} W}{\bar{\gamma}} \left(1 + N \max_{j\in[N]} \|B_j\|_{\mathrm{op}} \frac{\tau}{\bar{\gamma}}\right).$$

If $X_0, W, \bar{\kappa}, \tau, \bar{\gamma}^{-1}$ and $\max_{j\in[N]} \|B_j\|_{\mathrm{op}}$ are uniformly bounded in N, this gives $Y = \mathcal{O}(N)$. The definitions of X and D then imply $X, D = \mathcal{O}(N)$ as well and these estimates are uniform in H. □

E Proof of Theorem 1

Here, we give the proof of Theorem 1, which we restate here:

Theorem 1 (Individual regret under state-only feedback). *Let Assumptions 1, 2 and 3 hold. Suppose there exists $U \geq 1$ s.t. for all $t \geq 0, j \in [N], \|u_t^j\| \leq U$. If agent $i \in [N]$ runs Algorithm 1 under Setting 1 with (DAC-i) policy on perturbation sequence $\{\widetilde{w}_t\}$ and step size $\eta = \Theta(1/(G\widetilde{W}^2\sqrt{T}))$, where $\widetilde{W} = W + (N-1)U(\max_j \|B_j\|)$, and with $H = \max\{1, \lceil \log(\kappa_i T)/\gamma_i \rceil\}$, then for $T \geq H+1$, $\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{lin}) = \widetilde{\mathcal{O}}(U^2 N^2 \sqrt{T})$.*[1]

Remark 5. *The notation $\widetilde{\mathcal{O}}$ in Theorem 1 hides polynomial factors in $\gamma_i^{-1}, \kappa_i, \|B_i\|, G, d$ and logarithmic factors in T.*

Proof. Under Assumptions 1, 2 and 3, we apply [2, Theorem 5.1] for each agent $i \in [N]$. It remains to ensure that the considered perturbation sequence $\{\widetilde{w}_t\}$ in (3) also satisfies the boundedness condition of Assumption 2 using the boundedness of control inputs by U as follows:

$$\|\widetilde{w}_t\| = \left\| \sum\nolimits_{j\neq i} B_j u_t^j + w_t \right\| \leq \|w_t\| + \sum\nolimits_{j\neq i} \|B_j\| \cdot \|u_t^j\| \leq W + (N-1)U(\max_j \|B_j\|)\,, \quad (29)$$

where the last inequality follows from using boundedness of the control inputs of all the agents together with the bounded disturbances assumption (Assumption 2).

Selecting a step size $\eta = \Theta(1/(G\widetilde{W}^2\sqrt{T}))$, where $\widetilde{W} = W + (N-1)U(\max_j \|B_j\|)$, and a (per-agent) memory length $H \geq \log(\kappa_i T)/\gamma_i$, we obtain the desired regret for any $T \geq H+1$,

$$\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{\mathrm{lin}}) = \widetilde{\mathcal{O}}(U^2 N^2 \sqrt{T})\,. \quad (30)$$

This concludes the proof. □

F Proof of Theorem 2

We prove Theorem 2, restated here.

Theorem 2 (Individual regret under aggregate-input feedback). *Let Assumptions 1 to 4 hold. Then if agent $i \in [N]$ runs Algorithm 1 under Setting 2 with a (DAC-i) policy on perturbation sequence $\{w_t\}$, step size $\eta = 1/(N\sqrt{T})$, and with $H = \lceil \log(2\kappa_* N^2 \sqrt{T})/\gamma_* \rceil$ where κ_*, γ_* are the stability bounds in (47), and when **all** other agents use (DAC-i) policies with perturbation sequence (w_t) and feasible parameters $M_{j,t} \in \mathcal{M}_j$, then for any $T \geq H+1$,*

$$\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{DAC}) = \widetilde{\mathcal{O}}(N\sqrt{T}).$$

Fix a player i and a realized trajectory. Opponents' realized controls, costs, and disturbances are held fixed in every unilateral counterfactual below. In particular, opponents' feedback policies are not rerun on the counterfactual state. Recall the following useful notation:

$$A_i = A - B_i K_i, \quad q_i = 1 - \gamma_i, \quad b_i = \|B_i\|_{\mathrm{op}}, \quad \nu_i = \|K_i\|_{\mathrm{op}}, \quad h = H + 1,$$
$$\bar{q} = 1 - \bar{\gamma}, \quad a_p = 2\bar{\kappa}^2 \bar{q}^p, \quad m_H = \sum_{p=0}^{H-1} a_p, \quad B_\Sigma = \sum_{j=1}^{N} \|B_j\|_{\mathrm{op}}. \tag{31}$$

Define the following constants, whose bounds are proved below:

$$X = \bar{\kappa} X_0 + \frac{\bar{\kappa} W}{\bar{\gamma}}(1 + B_\Sigma m_H), \qquad R_i = \frac{2\kappa_i b_i W m_H}{\gamma_i}, \qquad X_i = X + R_i,$$
$$D_i^{\mathrm{sa}} = (1+\nu_i)(1+\kappa_i) X_i + W m_H, \qquad E_i = (1+\nu_i)\kappa_i X_i. \tag{32}$$

The proof uses a constant Λ_i such that

$$|c_t^i(x,u) - c_t^i(x',u')| \le \Lambda_i\big(\|x - x'\|_2 + \|u - u'\|_2\big) \tag{33}$$

whenever all four state/action vectors have norm at most D_i^{sa}. For Theorem 2, Assumption 1 gives $\Lambda_i = G D_i^{\mathrm{sa}}$. We keep Λ_i explicit so the same argument also applies under Assumption 5 with $\Lambda_i = \bar{L}$.

Remark 6. *The population rates below keep the dimensions, $X_0, W, G, \bar{\kappa}, \bar{\gamma}^{-1}$, $\max_j \|B_j\|_{\mathrm{op}}$, the individual strong-stability certificates, and κ_*, γ_*^{-1} uniformly bounded as N varies. Factors depending polynomially on these quantities and polylogarithmically on N, T are suppressed in $\widetilde{\mathcal{O}}$.*

Remark 7 (The initial evaluation interval)**.** *For every fixed DAC comparator Q, the trajectory comparison proved below gives:*

$$\|x_t^Q - x_t\|_2 + \|u_t^Q - u_t^i\|_2 \le P_i := (1+\nu_i) R_i + 2 W m_H.$$

Consequently, $\mathrm{Reg}_i^{1:H}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{\mathrm{DAC}}) \le H \Lambda_i P_i$. *Here $P_i = O(1)$, so the prefix is $O(NH)$ for Theorem 2 and $O(H)$ under Assumption 5. Together with Lemma 6, these estimates also preserve the corresponding full-horizon rates.*

F.1 Regret decomposition and proof overview

Let x_t^Q, u_t^Q denote the unilateral counterfactual obtained by using the fixed parameter $Q \in \mathcal{M}_i$ from time zero, with the same initial state and the opponents' realized controls held fixed. Let $C_t(Q) := c_t^i(x_t^Q, u_t^Q)$. For the interval $h = H+1, \dots, T$, define the shorthand:

$$\mathrm{Reg}_i^{h:T} := \mathrm{Reg}_i^{h:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{\mathrm{DAC}}) = \sum_{t=h}^{T} c_t^i(x_t, u_t^i) - \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} C_t(Q). \tag{34}$$

Use the unilateral memory loss from Section D.1, with external input $d_s = w_s + \sum_{j \ne i} B_j u_s^j$, DAC perturbation sequence $r_s = w_s$, and diagonal $f_t^i(Q) = \ell_t^i(Q, \dots, Q)$. Adding and subtracting the surrogate costs gives:

$$\begin{aligned}
\mathrm{Reg}_i^{h:T} = {} & \underbrace{\sum_{t=h}^{T} \big[c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t})\big]}_{\text{actual trajectory approximation error}} \\
& + \underbrace{\sum_{t=h}^{T} \ell_t^i(M_{i,t-h:t}) - \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q)}_{\text{online gradient descent with memory regret}} \\
& + \underbrace{\min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q) - \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} C_t(Q)}_{\text{comparator approximation error}}.
\end{aligned} \tag{35}$$

The minima exist because the parameter domain is compact and the costs are continuous functions of the parameter. Section F.2 bounds the first and third terms using individual stability of A_i. Section F.3 bounds the middle term and combines the estimates. Joint stability is used to bound the actual trajectory.

F.2 Counterfactual state and action deviation error

For any feasible parameter sequence Q_s, the unilateral dynamics are:

$$x_{s+1}^Q = A_i x_s^Q + B_i \sum_{p=1}^{H} Q_s^{[p-1]} w_{s-p} + d_s, \qquad x_0^Q = x_0. \tag{36}$$

Under the simultaneous DAC policies of the actual agents, Proposition 2 gives $\|x_t\|_2 \le X$ with X as in (32). Subtracting the actual recurrence from (36) cancels d_s:

$$x_{s+1}^Q - x_{s+1} = A_i(x_s^Q - x_s) + B_i \sum_{p=1}^{H} (Q_s^{[p-1]} - M_{i,s}^{[p-1]}) w_{s-p}.$$

The correction difference has norm at most $2Wm_H$. Since the initial difference is zero, individual strong stability therefore yields:

$$\|x_t^Q - x_t\|_2 \le 2\kappa_i b_i W m_H \sum_{r\ge 0} q_i^r = R_i, \qquad \|x_t^Q\|_2 \le X_i. \tag{37}$$

This holds for every feasible sequence Q_s, including any extension of a truncated parameter window to a full sequence.

The correct ideal state and action are as follows:

$$y_t^i(Q_{t-h:t-1}) = \sum_{r=1}^{h} A_i^{r-1} \left[d_{t-r} + B_i \sum_{p=1}^{H} Q_{t-r}^{[p-1]} w_{t-r-p} \right],$$
$$v_t^i(Q_{t-h:t}) = -K_i y_t^i(Q_{t-h:t-1}) + \sum_{p=1}^{H} Q_t^{[p-1]} w_{t-p}. \tag{38}$$

Set $d_s = w_s = 0$ for negative indices. For $t \ge h$, unrolling (36) gives

$$x_t^Q - y_t^i(Q_{t-h:t-1}) = A_i^h x_{t-h}^Q.$$

For $1 \le t < h$, zero padding instead gives $y_t^i(Q_{t-h:t-1}) = x_t^Q - A_i^t x_0$. Thus all true and ideal states have norm at most $(1+\kappa_i)X_i$, and their actions have norm at most $\nu_i(1+\kappa_i)X_i + Wm_H$. All these quantities are bounded by D_i^{sa}, including early rounds and mixed feasible parameter histories. For the scored rounds $t \ge h$,

$$\|x_t^Q - y_t^i(Q_{t-h:t-1})\|_2 + \|u_t^Q - v_t^i(Q_{t-h:t})\|_2 \le E_i q_i^h. \tag{39}$$

Combining (33) and (39) proves

$$\left| c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t}) \right| \le \Lambda_i E_i q_i^h. \tag{40}$$

Writing $n = T - H$, summing the previous inequality gives:

$$\sum_{t=h}^{T} \left[c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t}) \right] \le \Lambda_i E_i n q_i^h. \tag{41}$$

The same estimate holds uniformly for every fixed comparator:

$$\left| \sum_{t=h}^{T} f_t^i(Q) - \sum_{t=h}^{T} C_t(Q) \right| \le \Lambda_i E_i n q_i^h.$$

Taking a minimizer of the true comparator cost, or directly comparing the two minima using this uniform inequality, gives

$$\sum_{t=h}^{T} \left[c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t}) \right] + \min_{Q\in\mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q) - \min_{Q\in\mathcal{M}_i} \sum_{t=h}^{T} C_t(Q) \le 2\Lambda_i E_i n q_i^h. \tag{42}$$

Finally, subtracting the DAC action formulas and using (37) yields the bound P_i in the prefix remark.

F.3 Online gradient descent with memory regret bound

All projections and gradients use block Frobenius geometry. The DAC domain has diameter at most:

$$D_{0,i} = 2\left(\min(k_i, d)\sum_{p=0}^{H-1} a_p^2\right)^{1/2}. \tag{43}$$

Indeed, a block difference has operator norm at most $2a_p$, and its Frobenius norm is at most $2\sqrt{\min(k_i, d)}a_p$. Squaring these bounds and summing over the blocks proves the claim.

The following two subsections establish the constants

$$\begin{aligned} L_i^{\mathrm{mem}} &= \Lambda_i W\sqrt{H}\max\{1, (1+\nu_i)\kappa_i b_i\}, \\ G_{0,i} &= \Lambda_i W\sqrt{H}\left(1 + \frac{(1+\nu_i)\kappa_i b_i}{\gamma_i}\right) \end{aligned} \tag{44}$$

for coordinate-wise Lipschitzness of ℓ_t^i and the gradient bound for f_t^i. These bounds hold for every $t \geq 1$. Convexity follows from the affine unilateral state and action maps. Theorem 6, applied on the interval $h, \ldots, T$, therefore gives:

$$\sum_{t=h}^{T} \ell_t^i(M_{i,t-h:t}) - \min_{Q\in\mathcal{M}_i}\sum_{t=h}^{T} f_t^i(Q) \leq \frac{D_{0,i}^2}{2\eta} + \frac{\eta n}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i} h(h+1)\right]. \tag{45}$$

For completeness, projection nonexpansiveness implies $\|M_{i,s+1} - M_{i,s}\|_F \leq \eta G_{0,i}$ for $s \geq 1$. The convention $M_{i,0} = M_{i,1}$ gives the same bound at $s = 0$. Thus the discrepancy between a memory loss and its diagonal is at most:

$$L_i^{\mathrm{mem}}\sum_{k=1}^{h}\|M_{i,t-k} - M_{i,t}\|_F \leq \eta L_i^{\mathrm{mem}} G_{0,i}\frac{h(h+1)}{2}.$$

Adding this to the ordinary OGD bound $D_{0,i}^2/(2\eta) + \eta n G_{0,i}^2/2$ also proves (45) directly. In particular, the updates before the evaluation interval are accounted for and the parameter history is not reset at h.

Combining (35), (42), and (45) gives the finite-time estimate:

$$\mathrm{Reg}_i^{H+1:T} \leq 2\Lambda_i E_i (T-H) q_i^h + \frac{D_{0,i}^2}{2\eta} + \frac{\eta(T-H)}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i} h(h+1)\right]. \tag{46}$$

End of Proof of Theorem 2. Set $\Lambda_i = GD_i^{\mathrm{sa}}$ and use:

$$\eta = \frac{1}{N\sqrt{T}}, \qquad H = \max\left\{1, \left\lceil\frac{\log(2\kappa_* N^2\sqrt{T})}{\gamma_*}\right\rceil\right\},$$

where the given stability bounds satisfy:

$$\kappa_* \geq \max\{\bar{\kappa}, \max_{j\in[N]}\kappa_j\}, \qquad 0 < \gamma_* \leq \min\{\bar{\gamma}, \min_{j\in[N]}\gamma_j\}. \tag{47}$$

Under the uniform parameter convention stated above, $m_H = O(1)$, $X_i, D_i^{\mathrm{sa}}, E_i = O(N)$, $D_{0,i} = O(1)$, and $L_i^{\mathrm{mem}}, G_{0,i} = O(N\sqrt{H})$. Moreover,

$$q_i^h \leq e^{-\gamma_* H} \leq \frac{1}{2\kappa_* N^2\sqrt{T}}.$$

The truncation term in (46) is therefore $O(\sqrt{T})$, and the memory-OGD term is bounded by

$$O\big(N\sqrt{T}(1+H^3)\big) = \widetilde{\mathcal{O}}(N\sqrt{T}). \tag{48}$$

This proves the claimed suffix regret bound for $T \geq H+1$. The prefix remark gives the corresponding full-horizon rate.

F.3.1 Coordinate-wise loss Lipschitzness

Lemma 3 (Coordinate-wise loss Lipschitzness). *Let $t \geq 1$ and $k \in \{0, \ldots, h\}$. Consider two feasible parameter histories that differ only in the entry at time $t-k$, replacing Q by $\widetilde{Q}$. With the opponents' realized controls held fixed,*

$$\left|\ell_t^i(\ldots, Q, \ldots) - \ell_t^i(\ldots, \widetilde{Q}, \ldots)\right| \leq L_i^{\mathrm{mem}} \|Q - \widetilde{Q}\|_F,$$

where L_i^{mem} is given in (44).

Proof. Let $\Delta Q = Q - \widetilde{Q}$, and let Δy_t^i and Δv_t^i be the corresponding ideal state and action differences. The unilateral rollout (38) gives:

$$\Delta y_t^i = \begin{cases} 0, & k = 0, \\ A_i^{k-1} B_i \sum_{p=1}^{H} \Delta Q^{[p-1]} w_{t-k-p}, & 1 \leq k \leq h. \end{cases} \tag{49}$$

Moreover, $\Delta v_t^i = -K_i \Delta y_t^i + \mathbf{1}_{\{k=0\}} \sum_{p=1}^{H} \Delta Q^{[p-1]} w_{t-p}$. The block norm inequality

$$\sum_{p=1}^{H} \|\Delta Q^{[p-1]}\|_{\mathrm{op}} \leq \sum_{p=1}^{H} \|\Delta Q^{[p-1]}\|_F \leq \sqrt{H} \|\Delta Q\|_F$$

implies

$$\|\Delta y_t^i\|_2 + \|\Delta v_t^i\|_2 \leq W\sqrt{H} \max\{1, (1+\nu_i)\kappa_i b_i\} \|\Delta Q\|_F.$$

All the states and actions involved lie in the region bounded by D_i^{sa}. Applying (33) completes the proof. This argument includes the oldest coordinate $k = h$ and remains valid when $B_i = 0$. □

F.3.2 Gradient loss boundedness

Lemma 4 (Diagonal parameter-gradient bound). *For every $t \geq 1$ and every $Q \in \mathcal{M}_i$,*

$$\|\nabla f_t^i(Q)\|_F \leq G_{0,i},$$

where $G_{0,i}$ is defined in (44).

Proof. The derivative of the diagonal state map in direction ΔQ is given by:

$$\mathcal{J}_{i,t}[\Delta Q] = \sum_{r=1}^{h} A_i^{r-1} B_i \sum_{p=1}^{H} \Delta Q^{[p-1]} w_{t-r-p}.$$

Its Euclidean norm is at most $W\sqrt{H}\kappa_i b_i \|\Delta Q\|_F / \gamma_i$. The diagonal action derivative is $-K_i \mathcal{J}_{i,t}[\Delta Q] + \sum_{p=1}^{H} \Delta Q^{[p-1]} w_{t-p}$. The differentiable cost satisfying (33) has partial gradients bounded by Λ_i on the relevant region. For $\Lambda_i = G D_i^{\mathrm{sa}}$ this follows directly from Assumption 1, and for $\Lambda_i = L$ it follows from Assumption 5. The chain rule therefore bounds every directional derivative of f_t^i by:

$$\Lambda_i W \sqrt{H} \left(1 + \frac{(1+\nu_i)\kappa_i b_i}{\gamma_i}\right) \|\Delta Q\|_F.$$

Taking the supremum over directions of unit block Frobenius norm proves the claim. □

G Proof of Theorem 3

Here, we develop the proof of Theorem 3, restated here:

Theorem 3. *Under the setting of Theorem 2, replace gradient boundedness in Assumption 1 -(ii) by Assumption 5. Set instead $\eta = 1/\sqrt{T}$ and choose $H = \lceil \log(2\kappa_* N\sqrt{T})/\gamma_* \rceil$ with κ_*, γ_* as in (47). Then we have for any $T \geq H + 1$: $\mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{DAC}) = \widetilde{\mathcal{O}}(\sqrt{T})$.*

Proof. The proof follows the same strategy as the proof of Theorem 2. We explain the modifications needed when the cost functions satisfy the uniform Lipschitz assumption 5. The state and action bounds established in Section F depend only on the dynamics, the feasible DAC parameters, the initial-state bound, and the disturbance bound. Consequently, these trajectory bounds remain unchanged. The modification concerns how trajectory differences and parameter changes are converted into cost differences.

Fix a player $i \in [N]$. We use the unilateral counterfactuals and surrogates from Section F, with the opponents' realized controls held fixed. Recall the notation:

$$h = H + 1, \qquad q_i = 1 - \gamma_i, \qquad b_i = \|B_i\|_{\mathrm{op}}, \qquad \nu_i = \|K_i\|_{\mathrm{op}},$$

and the constants $X_i, D_i^{\mathrm{sa}}, E_i, D_{0,i}$ defined there. Recall also the notation:

$$C_t(Q) := c_t^i(x_t^Q, u_t^Q), \qquad f_t^i(Q) := \ell_t^i(Q, \dots, Q), \qquad Q \in \mathcal{M}_i,$$

for the true unilateral comparator cost and the diagonal surrogate cost, respectively.

Under Assumption 5, we have

$$|c_t^i(x, u) - c_t^i(x', u')| \le \bar{L}\big(\|x - x'\|_2 + \|u - u'\|_2\big).$$

Thus the local cost bound (33) holds with $\Lambda_i = \bar{L}$, replacing the radius-dependent factor $\Lambda_i = GD_i^{\mathrm{sa}}$ used in the preceding theorem.

Recall the regret decomposition (35) in Section F.1. We adapt its approximation terms and its online gradient descent regret term as follows.

- **Counterfactual state and action deviation errors.** We first bound the two approximation terms in the regret decomposition: the discrepancy between the actual cost and its memory surrogate, and the discrepancy between the surrogate comparator cost and the true comparator cost.

  The trajectory estimate (39) bounds the sum of the state and action truncation errors by $E_i q_i^h$ for every $t \ge h$. Applying the uniform Lipschitz assumption therefore gives

  $$\big|c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t})\big| \le \bar{L} E_i q_i^h.$$

  The same argument applies to every fixed feasible comparator $Q \in \mathcal{M}_i$:

  $$|f_t^i(Q) - C_t(Q)| \le \bar{L} E_i q_i^h.$$

  In both inequalities, the trajectory estimate is unchanged; only the cost factor Λ_i is replaced by $\bar{L}$.

  There are $T - H$ rounds in the evaluation interval $\{H + 1, \dots, T\}$. Summing the first inequality therefore bounds the actual trajectory approximation term by $\bar{L} E_i (T - H) q_i^h$. Moreover, the second inequality is uniform over Q, so

  $$\min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q) - \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} C_t(Q) \le \bar{L} E_i (T - H) q_i^h.$$

  Indeed, this follows by evaluating the surrogate sum at a minimizer of the true comparator sum and applying the uniform approximation bound.

  Consequently, the sum of the first and third terms in (35) satisfies

  $$\sum_{t=h}^{T} \big[c_t^i(x_t, u_t^i) - \ell_t^i(M_{i,t-h:t})\big] + \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q) - \min_{Q \in \mathcal{M}_i} \sum_{t=h}^{T} C_t(Q) \le 2\bar{L} E_i (T - H)(1 - \gamma_i)^{H+1}. \quad (50)$$

  Under the uniform problem parameter bounds stated in Section F, we have $E_i = \mathcal{O}(N)$. Thus the approximation bound has a coefficient growing linearly in N. The additional state-radius factor present when $\Lambda_i = GD_i^{\mathrm{sa}}$ has been removed.

- **Online gradient descent with memory regret bound.** The parameter domain, the block Frobenius geometry, and the projected gradient update are unchanged. In particular, the same diameter bound $D_{0,i}$ applies.

  Recall from (45) that the online-gradient contribution is bounded by:

$$\sum_{t=h}^{T} \ell_t^i(M_{i,t-h:t}) - \min_{Q\in\mathcal{M}_i} \sum_{t=h}^{T} f_t^i(Q) \le \frac{D_{0,i}^2}{2\eta} + \frac{\eta(T-H)}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i}(H+1)(H+2)\right]. \tag{51}$$

  The factor $(H+1)(H+2)$ is $h(h+1)$, because the memory loss depends on the parameter history $M_{i,t-h:t}$ with $h = H+1$.

  It remains to specify the coordinate-wise Lipschitz constant L_i^{mem} and the diagonal gradient bound $G_{0,i}$ under Assumption 5. The following two modifications suffice.

  (i) *Coordinate-wise Lipschitzness.* In the proof of Lemma 3, the effect of changing a single parameter in the history is first bounded at the level of the ideal state and action. These bounds depend on the individual closed-loop matrix $A_i = A - B_i K_i$, the disturbance bound, and the DAC parameter geometry.
  The cost difference is then bounded using (33). Replacing Λ_i by $\bar{L}$ gives:

$$L_i^{\mathrm{mem}} = \bar{L} W \sqrt{H}\, \max\{1, (1+\nu_i)\kappa_i b_i\}.$$

  Thus, when two feasible parameter histories differ only in one coordinate, their memory losses differ by at most L_i^{mem} times the block Frobenius distance between those two parameter values. In particular, this Lipschitz constant no longer contains the state-radius factor D_i^{sa}.

  (ii) *Diagonal gradient boundedness.* Similarly, the proof of Lemma 4 bounds the sensitivity of the diagonal ideal state and action to the fixed parameter Q. Applying the same argument with $\Lambda_i = \bar{L}$ gives:

$$\|\nabla f_t^i(Q)\|_F \le G_{0,i} := \bar{L} W \sqrt{H}\left(1 + \frac{(1+\nu_i)\kappa_i b_i}{\gamma_i}\right), \qquad Q \in \mathcal{M}_i.$$

  This gradient bound also contains no state-radius factor.

  Under the same uniform problem-parameter bounds, these constants satisfy

$$D_{0,i} = \mathcal{O}(1), \qquad L_i^{\mathrm{mem}} = \mathcal{O}(\sqrt{H}), \qquad G_{0,i} = \mathcal{O}(\sqrt{H}).$$

  Therefore, the online-gradient bound has no explicit polynomial dependence on N. Its dependence on the population size through the chosen memory length H will be only polylogarithmic.

Choice of memory and step size. Combining the two parts above with (35) gives:

$$\mathrm{Reg}_i^{H+1:T} \le 2\bar{L}E_i(T-H)q_i^{H+1} + \frac{D_{0,i}^2}{2\eta} + \frac{\eta(T-H)}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i}(H+1)(H+2)\right].$$

This is the estimate (46) with $\Lambda_i = \bar{L}$. Then, choose

$$\eta = \frac{1}{\sqrt{T}}, \qquad H = \max\left\{1, \left\lceil \frac{\log(2\kappa_* N\sqrt{T})}{\gamma_*} \right\rceil\right\},$$

where κ_*, γ_* are the stability bounds in (47). In particular, $\gamma_* \le \gamma_i$. Since $1 - z \le e^{-z}$ for $z \ge 0$, this choice gives:

$$q_i^{H+1} = (1-\gamma_i)^{H+1} \le e^{-\gamma_i(H+1)} \le e^{-\gamma_* H} \le \frac{1}{2\kappa_* N\sqrt{T}}.$$

For the approximation term, using $E_i = \mathcal{O}(N)$ and $T - H \le T$, we obtain:

$$2\bar{L}E_i(T-H)q_i^{H+1} \le \frac{\bar{L}E_i(T-H)}{\kappa_* N\sqrt{T}} = \mathcal{O}(\sqrt{T}).$$

For the online gradient descent regret term, the bounds on its constants give:

$$G_{0,i}^2 = \mathcal{O}(H), \qquad L_i^{\mathrm{mem}} G_{0,i}(H+1)(H+2) = \mathcal{O}(H^3).$$

Consequently, with $\eta = 1/\sqrt{T}$,

$$\begin{aligned}
&\frac{D_{0,i}^2}{2\eta} + \frac{\eta(T-H)}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i}(H+1)(H+2)\right] \\
&\quad \le \frac{D_{0,i}^2\sqrt{T}}{2} + \frac{\sqrt{T}}{2}\left[G_{0,i}^2 + L_i^{\mathrm{mem}} G_{0,i}(H+1)(H+2)\right] \\
&\quad = \mathcal{O}\big(\sqrt{T}(1+H^3)\big).
\end{aligned}$$

For fixed stability envelopes, the selected H is logarithmic in N and T. Combining the preceding estimates proves, for $T \ge H+1$,

$$\mathrm{Reg}_i^{H+1:T}\big(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{\mathrm{DAC}}\big) \le \widetilde{\mathcal{O}}(\sqrt{T}).$$

The initial evaluation interval. The same full-horizon rate follows by including the initial rounds. Recall from Section F that, for every fixed feasible comparator Q,

$$\|x_t^Q - x_t\|_2 \le R_i, \qquad \|u_t^Q - u_t^i\|_2 \le \nu_i R_i + 2W m_H,$$

where $R_i = \frac{2\kappa_i b_i W m_H}{\gamma_i}$. Thus, it follows that:

$$\|x_t^Q - x_t\|_2 + \|u_t^Q - u_t^i\|_2 \le P_i := (1+\nu_i)R_i + 2W m_H.$$

Then, Assumption 5 implies:

$$|c_t^i(x_t, u_t^i) - C_t(Q)| \le \bar{L} P_i$$

uniformly over Q. Summing over the first H rounds and taking the best fixed comparator gives

$$\mathrm{Reg}_i^{1:H}\big(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i^{\mathrm{DAC}}\big) \le H\bar{L}P_i.$$

Since m_H and R_i are uniformly bounded under the stated parameter convention, $P_i = \mathcal{O}(1)$. The prefix contribution is therefore $\mathcal{O}(H)$. By Lemma 6,

$$\mathrm{Reg}_i^{1:T} \le \mathrm{Reg}_i^{1:H} + \mathrm{Reg}_i^{H+1:T} \le \widetilde{\mathcal{O}}(\sqrt{T}).$$

The hidden factors depend polynomially on $X_0, W, \bar{L}, \bar{\kappa}, \bar{\gamma}^{-1}, \kappa_i, \gamma_i^{-1}, \kappa_*, \gamma_*^{-1}$, the input-matrix bounds, and the dimensions, and only polylogarithmically on N and T. □

H Proof of Theorem 4

In this section, we develop the proof of Theorem 4, which we restate here:

Theorem 4 (Horizon lower bound for linear and DAC comparators). *Let $N \ge 1$ and $i \in [N]$. There exist scalar dynamics, deterministic disturbances, and supplied baseline controllers satisfying Assumptions 2–4 with $W = 1, \kappa_j = \bar{\kappa} = 1, \gamma_j = \bar{\gamma} = \frac{1}{2}$, such that the following holds. For every $T \ge 1$, $H \ge 1$, prescribed nonempty interval $I \subseteq \{1, \dots, T\}$, and causal, possibly randomized learner $\mathcal{A}_i$ under either information setting, there exists a deterministic oblivious cost sequence satisfying Assumption 1 with $\beta = G = 4$ and Assumption 5 with $\bar{L} = 1$ for which $\mathbb{E}_{\mathcal{A}_i}[\mathrm{Reg}_i^I(\mathcal{A}_i, \{u_t^{-i}\}, \Pi)] \ge \frac{\sqrt{|I|}}{2\sqrt{3}}$ simultaneously for the $(1, \frac{1}{2})$-strongly stable linear class $\Pi = \Pi_i^{\mathrm{lin}}$ and the memory-H DAC class $\Pi = \Pi_i^{\mathrm{DAC}}$ with fixed supplied baseline $K_i = 0$. The conclusion holds for every sequence of opponents' controls.*

The horizon lower bound is inherited from the classical random sign argument for online learning; see [6, Theorem 4.1] and [36, Section 3.2, Theorem 3.2]. We give a self-contained embedding that verifies its compatibility with our cost assumptions, controller benchmarks, and unrestricted learner controls. The statistical lower bound mechanism is standard.

Why the embedding needs a proof. While the lower bound relies on the same core argument as in standard online learning, we cannot immediately apply the classical result since the following properties need to be verified:

1. **Controller realization and counterfactual states.** Fixed comparison actions must be represented by admissible controllers. Moreover, the learner and comparator generally induce different states. We must construct an instance where the desired action correspondence and counterfactual cost comparison are valid.
2. **Admissible cost functions.** Nonzero affine cost functions fail the origin-centered growth and gradient conditions in Assumption 1. We must preserve the classical affine losses on the comparison slice while extending them to jointly convex, continuously differentiable costs satisfying the global assumptions.
3. **All allowed learner controls.** The comparison interval does not restrict the learner's control space. After making the costs nonlinear outside that interval, we must verify that every real control still has the required conditional expected cost lower bound. While the standard example relies on linearity of the loss functions and sign cancellation, we extend the argument here to our nonlinear cost setting using convexity.
4. **DAC feasibility.** For the DAC conclusion, the same comparison trajectories must be realized by feasible memory-H parameters with the fixed supplied baseline.

Notation and timing. Dynamics begin at time zero, and costs are evaluated on a prescribed nonempty interval $I \subseteq \{1, \dots, T\}$. The current cost function is revealed only after the current control input is chosen. The lower bound remains valid even if the learner knows the dynamics, the entire deterministic disturbance sequence, and all past cost functions. For a fixed disturbance sequence, recall that the regret incurred by player i over the time interval I is given by:

$$R_i^I(\Pi) := \sum_{t \in I} c_t^i(x_t, u_t^i) - \min_{\pi \in \Pi} \sum_{t \in I} c_t^i(x_t^\pi, u_t^\pi). \tag{52}$$

All counterfactual trajectories begin at the same initial state at time zero, and opponents' realized control sequences are held fixed.

Proof. Fix $T, H, I, \mathcal{A}_i$, let $n = |I|$, and let ξ denote the learner's random seed.

Step 0. Classical proof template and required embedding

For independent uniform signs $\sigma_t \in \{-1, 1\}$, the classical argument uses $f_t(u) = \sigma_t u$ and the comparison interval $[-1/2, 1/2]$. A causal learner restricted to this interval has expected cumulative loss zero since $\mathbb{E}[\sigma_t] = 0$, whereas, writing $S_I = \sum_{t \in I} \sigma_t$,

$$\min_{u \in [-1/2, 1/2]} \sum_{t \in I} \sigma_t u = -\tfrac{1}{2}|S_I|.$$

Thus the expected regret is $\frac{1}{2}\mathbb{E}|S_I|$.

We reproduce this calculation within the dynamical model, up to a common per-round constant $b = 3/4$. Specifically, we will prove

$$\mathbb{E}C_I^{\mathcal{A}} \geq \tfrac{3}{4}n, \qquad \min_{\pi \in \Pi} C_I^\pi \leq \tfrac{3}{4}n - \tfrac{1}{2}|S_I|, \tag{53}$$

where $C_I^{\mathcal{A}}$ and C_I^π are the learner's and comparator's cumulative costs. Steps 1–4 establish these bounds with admissible dynamics, costs, and controllers. Combining the above inequalities implies:

$$\mathbb{E}_{\sigma,\xi}[R_i^I(\Pi)] \geq \frac{1}{2}\mathbb{E}_\sigma[|S_I|], \quad \Pi \in \{\Pi_i^{\mathrm{lin}}, \Pi_i^{\mathrm{DAC}}\}. \tag{54}$$

Steps 5–7 complete the standard fluctuation argument showing that $\mathbb{E}|S_I| = \Omega(\sqrt{|I|})$ and deducing the existence of the cost function from the lower bound in expectation.

Step 1. Construction of the scalar dynamical system instance

Take scalar states and scalar controls for every player, and choose

$$A = 0, \qquad B_j = 0 \quad (j \in [N]), \qquad w_t = 1 \quad (t \geq 0), \qquad x_0 = 0. \tag{55}$$

As in the DAC definition, recall that $w_t = 0$ for $t < 0$. Then $x_{t+1} = 1$ for every $t \geq 0$, so

$$x_t = x_t^\pi = 1, \qquad \forall t \geq 1, \tag{56}$$

for the learner and every comparator, independently of all controls.

The choice $B_j = 0$ makes the counterfactual states coincide. The choice $w_t = 1$ ensures that homogeneous linear controllers can represent nonzero constant actions: $u_t = -Kx_t = -K$ for $t \geq 1$. In contrast, zero disturbances and $x_0 = 0$ would make all such controllers produce zero inputs.

Let $K_j = 0$ be the baseline controller for every player. Then $Q = 1$ and $L = 0$ guarantee strong stability with $\kappa_j = \bar{\kappa} = 1$ and $\gamma_j = \bar{\gamma} = \frac{1}{2}$. Moreover, the disturbance bound holds with $W = 1$ since $|w_t| = 1 \leq 1$ for all $t \geq 0$. Opponents' controls affect neither the state nor aggregate feedback, since $\sum_{j \neq i} B_j u_t^j = 0$. Their costs may be identically set to zero. The construction works under both information settings.

Step 2. Construction of the hard cost functions satisfying the assumptions

We now construct hard cost functions that satisfy our assumptions while retaining the main properties of the classical affine losses used in the standard online learning proof at the common state $x = 1$.

Consider first the Huber function with threshold $\frac{1}{2}$:

$$h(z) = \begin{cases} z^2, & |z| \leq \frac{1}{2}, \\ |z| - \frac{1}{4}, & |z| > \frac{1}{2}. \end{cases} \tag{57}$$

Sample independent uniform signs $\sigma_t \in \{-1, 1\}$ for $t \in I$ before play, independently of ξ, and set

$$c_t^i(x, u) = \begin{cases} h(x + \sigma_t u), & t \in I, \\ 0, & t \notin I. \end{cases} \tag{58}$$

Set $c_0^i = 0$ and the other players' costs to zero. This defines an oblivious distribution over cost sequences.

The derivative $h'(z)$ equals -1 for $z < -1/2$, $2z$ for $|z| \leq 1/2$, and 1 for $z > 1/2$. It is continuous and nondecreasing, so h is continuously differentiable and convex. Moreover,

$$0 \leq h(z) \leq z^2, \qquad |h'(z)| \leq 2|z|, \qquad |h'(z)| \leq 1. \tag{59}$$

Composition with the linear map $(x, u) \mapsto x + \sigma_t u$ makes c_t^i jointly convex and continuously differentiable. If $|x| \leq D$ and $|u| \leq D$, then $|x + \sigma_t u| \leq 2D$, and

$$0 \leq c_t^i(x, u) \leq 4D^2, \qquad |\partial_x c_t^i(x, u)|,\ |\partial_u c_t^i(x, u)| \leq 4D.$$

This verifies Assumption 1 with $\beta = G = 4$. Also,

$$|c_t^i(x, u) - c_t^i(x', u')| \leq |x - x'| + |u - u'|,$$

which verifies Assumption 5 with $\bar{L} = 1$.

Finally, we have the exact identity:

$$c_t^i(1, u) = \tfrac{3}{4} + \sigma_t u\,, \qquad \forall t \in I\,, |u| \leq \tfrac{1}{2}. \tag{60}$$

This is the standard loss plus a common constant. The learner is not restricted to this slice, so its expected cost still requires a bound valid for every real control.

Remark 8 (On the choice of the function h)**.** *For $|u| \leq 1/2$, the arguments $1 + \sigma_t u$ lie in $[1/2, 3/2]$. We therefore seek a convex function h that is affine with slope one on that interval, while satisfying $h(0) = h'(0) = 0$ to accommodate Assumption 1. Joining the quadratic z^2 at $z = 1/2$ to an affine branch of slope one gives $z - 1/4$: both its value and derivative agree with the quadratic at the joining point. Thus the quadratic core enforces the origin-centered conditions while the positive affine branch preserves the classical comparison losses.*

Step 3. Lower bound on the expected learner cost for arbitrary controls

Let ξ denote the learner's random seed, containing all its internal randomization and sampled independently of the sign vector $(\sigma_s)_{s \in I}$. For every $t \geq 1$, define the σ-algebra:

$$\mathcal{F}_t^- := \sigma(\xi, (\sigma_s)_{s \in I,\ s < t})\,.$$

In the constructed instance, the state and aggregate feedback are deterministic, and all past cost feedback is determined by the past signs and the learner's random seed. Consequently, causality implies that u_t^i is $\mathcal{F}_t^-$-measurable. For every $t \in I$, the current sign σ_t is independent of $\mathcal{F}_t^-$, so

$$\mathbb{P}(\sigma_t = 1 \mid \mathcal{F}_t^-) = \mathbb{P}(\sigma_t = -1 \mid \mathcal{F}_t^-) = \frac{1}{2}, \qquad \text{almost surely.}$$

Using the $\mathcal{F}_t^-$-measurability of u_t^i and the conditional distribution of σ_t, we obtain, for every $t \in I$,

$$\begin{aligned}\mathbb{E}[c_t^i(x_t, u_t^i) \mid \mathcal{F}_t^-] &= \tfrac{1}{2}\big[h(1+u_t^i) + h(1-u_t^i)\big] \\ &\geq h\left(\frac{(1+u_t^i)+(1-u_t^i)}{2}\right) = h(1) = \tfrac{3}{4}. \end{aligned} \tag{61}$$

For affine losses the classical proof uses exact sign cancellation; for the admissible convex extension, Jensen's inequality gives a lower bound for every real input.

Summing (61) gives

$$\mathbb{E}_{\sigma,\xi} C_I^{\mathcal{A}} \geq \tfrac{3}{4} n, \qquad C_I^{\mathcal{A}} := \sum_{t \in I} c_t^i(x_t, u_t^i). \tag{62}$$

The costs are nonnegative, so this remains valid if the expected learner cost is infinite.

Step 4. Common upper bound on the comparator term for linear and DAC policies

(a) Strongly stable linear comparators. For $s \in \{-1, 1\}$ choose

$$K_s = \frac{s}{2}.$$

Then $|K_s| \leq 1$ and $A - B_i K_s = 0$, so both controllers belong to the $(1, 1/2)$-strongly stable linear class. For $t = 0$, $u_0^{K_s} = -K_s x_0 = 0$ and for $t \geq 1$,

$$u_t^{K_s} = -K_s x_t^{K_s} = -\frac{s}{2}.$$

By (60),

$$c_t^i(x_t^{K_s}, u_t^{K_s}) = \tfrac{3}{4} - \tfrac{s\sigma_t}{2}, \qquad \forall t \in I.$$

Consequently, with $S_I = \sum_{t \in I} \sigma_t$,

$$C_I^{K_s} = \tfrac{3}{4} n - \tfrac{s}{2} S_I, \qquad \min_{s \in \{-1,1\}} C_I^{K_s} = \tfrac{3}{4} n - \tfrac{1}{2}|S_I|. \tag{63}$$

The minimum over the full linear class is no larger than the minimum over this pair.

(b) Fixed-baseline DAC comparators. For every $H \geq 1$, set again $K_i = 0$ and choose:

$$M_s^{[0]} = s/2, \qquad M_s^{[p]} = 0, \quad \forall 1 \leq p < H.$$

These parameters satisfy the DAC feasibility conditions: $|M^{[p]}| \leq 2(\frac{1}{2})^p$ for $0 \leq p < H$. Indeed, at $\kappa = 1$ and $\gamma = 1/2$, the first-block radius is at least 1, whereas $|M_s^{[0]}| = 1/2$, and every remaining block is zero. Since $w_t = 0$ for $t < 0$ and $x_0 = 0$, we also have $u_0^{M_s} = 0$. For $t \geq 1$,

$$u_t^{M_s} = -K_i x_t^{M_s} + \sum_{p=1}^{H} M_s^{[p-1]} w_{t-p} = \frac{s}{2} w_{t-1} = \frac{s}{2}.$$

Thus these DAC policies realize the same pair of comparison trajectories and costs as the two linear controllers, with the sign labels interchanged. In Setting 1, the effective disturbances also equal w_t on this instance, so the same calculation applies there.

It follows that, for either comparator class,

$$\min_{\pi \in \Pi} C_I^{\pi} \leq \tfrac{3}{4} n - \tfrac{1}{2}|S_I|, \qquad \Pi \in \{\Pi_i^{\mathrm{lin}}, \Pi_i^{\mathrm{DAC}}\}. \tag{64}$$

Step 5. Combining Step 3 (cumulative cost lower bound) and Step 4 (comparator upper bound)

Define the common lower bound

$$D_I(\sigma,\xi) := C_I^{\mathcal{A}}(\sigma,\xi) - \tfrac{3}{4}n + \tfrac{1}{2}|S_I(\sigma)|\,, \tag{65}$$

where we make the dependence on randomness explicit using the notation $C_I^{\mathcal{A}}(\sigma,\xi) := C_I^{\mathcal{A}}$ and $S_I(\sigma) := S_I$. By (64), we obtain for every realization,

$$R_i^I(\Pi_i^{\mathrm{lin}}) \geq D_I(\sigma,\xi), \qquad R_i^I(\Pi_i^{\mathrm{DAC}}) \geq D_I(\sigma,\xi). \tag{66}$$

Combining with (62) yields:

$$\mathbb{E}_{\sigma,\xi} D_I(\sigma,\xi) \geq \tfrac{1}{2}\mathbb{E}_\sigma|S_I|. \tag{67}$$

The common baseline $3n/4$ cancels, just as the baseline cancels in the classical proof. What remains is the same random-sign fluctuation term $\frac{1}{2}\mathbb{E}_\sigma|S_I|$.

Step 6. Elementary lower bound on random sign fluctuation: $E_\sigma|S_I| \geq \sqrt{n/3}$.

This lower bound is standard in the literature and used as well in the classical online learning bound. We provide a full proof here for completeness.

First, by independence of the random variables σ_t which also has zero mean, and since $\sigma_t^2 = 1$, we have $\mathbb{E}S_I^2 = n$.

The fourth moment of S_I can be bounded as follows:

$$\mathbb{E}S_I^4 = \mathbb{E}\left[\sum_{i,j,k,l}\sigma_i\sigma_j\sigma_k\sigma_l\right] = \sum_{t\in I}\mathbb{E}[\sigma_t^4] + 6\sum_{i,j\in I, i<j}\mathbb{E}[\sigma_i^2\sigma_j^2] = n + 6\binom{n}{2} = 3n^2 - 2n \leq 3n^2\,, \tag{68}$$

where the first identity follows from expanding S_I^4, the second one follows from observing that any term with an index appearing an odd number of times has expectation zero, and the remaining terms are n terms with one index repeated four times and 6 $\binom{n}{2}$ terms corresponding to the terms where $\sigma_i\sigma_j\sigma_k\sigma_l = \sigma_i^2\sigma_j^2$ for $i \neq j$, i.e., for each of the $\binom{n}{2}$ unordered pairs of distinct indices, there are $\binom{4}{2} = 6$ arrangements with each index appearing twice. Every such term has expectation one.

Then, Hölder's inequality with exponents $3/2$ and 3 yields:

$$n = \mathbb{E}|S_I|^2 = \mathbb{E}\Big[|S_I|^{2/3}|S_I|^{4/3}\Big] \leq (\mathbb{E}|S_I|)^{2/3}(\mathbb{E}|S_I|^4)^{1/3}.$$

Rearranging and using (68), we obtain:

$$\mathbb{E}|S_I| \geq \frac{n^{3/2}}{\sqrt{\mathbb{E}S_I^4}} \geq \sqrt{n/3}. \tag{69}$$

Substitution into (67) gives

$$\mathbb{E}_{\sigma,\xi} D_I(\sigma,\xi) \geq \frac{\sqrt{n}}{2\sqrt{3}}. \tag{70}$$

Step 7. Fixing a deterministic oblivious cost sequence

The sign vector on I has finite support. By (70), at least one deterministic sign vector $\sigma^\star$ satisfies

$$\mathbb{E}_\xi D_I(\sigma^\star,\xi) \geq \frac{\sqrt{n}}{2\sqrt{3}}. \tag{71}$$

Otherwise, averaging these finitely many expectations would contradict (70). These expectations are well defined because $D_I(\sigma,\xi) \geq -3n/4$.

Fixing $\sigma^\star$ in (58) gives a deterministic oblivious cost sequence, chosen independently of the learner's realized random seed. Both inequalities in (66) hold for this same sequence, so (71) proves the claimed expected regret bound simultaneously for the linear and DAC comparator classes. The sequence works for every opponent control sequence, since opponents do not have any affect on costs and state dynamics for the chosen dynamical system instance. □

I Proof of Theorem 5

We first recall the theorem:

Theorem 5 (Equilibrium tracking). *Let Assumptions 1, 2, and 4 hold for the common-state-cost model specified above, together with the stated feedback and coordinate-gradient access and Assumption 6. Then, for every $M_1 \in \mathcal{M}$, the simultaneous updates (6) satisfy, for every $T \geq 1$,*

$$\frac{1}{T}\sum_{t=1}^{T} g_t(M_t)^2 \leq \frac{2C_M}{LT}\left[F_1(M_1) + V_c(T) + G_x Q\, V_w^+(T)\right], \tag{8}$$

where the constants D_x, G_x, Q, L, C_M, independent of T for fixed H, are given in Appendix I.

Outline of the proof. The proof of Theorem 5 can be divided into three main steps that are recorded in the following three propositions:

1. Proposition 3 upperbounds the sum of squared equilibrium gaps by the sum of policy parameter deviations across time and players.
2. Proposition 4 bounds these policy parameter deviations by the cumulative decrease of the common potential.
3. Finally, Proposition 5 bounds that decrease by the initial potential value and the cost and disturbance variation.

Proposition 3. *For the updates (6) and every $T \geq 1$,*

$$\sum_{t=1}^{T} g_t(M_t)^2 \leq C_M \sum_{t=1}^{T}\sum_{i=1}^{N} \|M_{i,t+1} - M_{i,t}\|_F^2, \tag{72}$$

where C_M is defined in (80).

Proposition 4. *For the updates (6), with the positive smoothness bound L in (80), and every $T \geq 1$,*

$$\sum_{t=1}^{T}\sum_{i=1}^{N} \|M_{i,t+1} - M_{i,t}\|_F^2 \leq \frac{2}{L}\sum_{t=1}^{T}\big[F_t(M_t) - F_t(M_{t+1})\big]. \tag{73}$$

Proposition 5. *For every $T \geq 1$ and every feasible sequence $M_1, \ldots, M_{T+1} \in \mathcal{M}$,*

$$\sum_{t=1}^{T}\big[F_t(M_t) - F_t(M_{t+1})\big] \leq F_1(M_1) + V_c(T) + G_x Q\, V_w^+(T). \tag{74}$$

This inequality does not require the sequence to follow the update rule.

Geometry and explicit constants. All matrix operator norms are denoted by $\|\cdot\|_{\mathrm{op}}$. On a player's DAC parameter space and on the joint product space, we use the block Frobenius norms

$$\|M_i\|_F^2 = \sum_{p=0}^{H-1} \|M_i^{[p]}\|_F^2, \qquad \|M\|_F^2 = \sum_{i=1}^{N} \|M_i\|_F^2.$$

Gradients, projections, and parameter inner products use this geometry. Let $h = H + 1$ and $q = 1 - \bar{\gamma} \in (0, 1)$. Joint strong stability gives:

$$\|\bar{A}^r\|_{\mathrm{op}} \leq \bar{\kappa} q^r, \quad \forall r \geq 0, \qquad \bar{A} = A - \sum_{i=1}^{N} B_i K_i. \tag{75}$$

Write the DAC block radii as a_p, so that

$$\mathcal{M}_i = \left\{(M_i^{[0]}, \ldots, M_i^{[H-1]}) : \|M_i^{[p]}\|_{\mathrm{op}} \leq a_p,\ 0 \leq p < H\right\}, \quad a_p = 2\bar{\kappa}^2 q^p. \tag{76}$$

The proof below only uses the actual radii a_p in (76). If $r_i = \min(k_i, d)$, define the diameter bound

$$D_i = 2\left(r_i \sum_{p=0}^{H-1} a_p^2\right)^{1/2}. \tag{77}$$

Then $\operatorname{diam}_F(\mathcal{M}_i) \le D_i$: each block difference has operator norm at most $2a_p$ and Frobenius norm at most $2\sqrt{r_i}\, a_p$. The domains are compact and convex.

Set

$$b_i = \|B_i\|_{\mathrm{op}}, \qquad B_\Sigma = \sum_{i=1}^{N} b_i, \qquad S_h = \bar{\kappa} \sum_{r=0}^{h-1} q^r, \qquad m_H = \sum_{p=0}^{H-1} a_p,$$
$$Q = S_h(1 + B_\Sigma m_H), \qquad D_x = WQ, \qquad G_x = GD_x. \tag{78}$$

In particular, $S_h \le \bar{\kappa}/\bar{\gamma}$ and $m_H \le 2\bar{\kappa}^2/\bar{\gamma}$. Define

$$J_i = W\sqrt{H}\, S_h b_i, \qquad J = \left(\sum_{i=1}^{N} J_i^2\right)^{1/2}, \tag{79}$$

and choose

$$L = \max\{1, \zeta J^2\}, \qquad G_{M,i} = G_x J_i, \qquad C_M = \sum_{i=1}^{N} (LD_i + G_{M,i})^2. \tag{80}$$

Lemma 5 proves that D_x bounds the rollout states, $G_{M,i}$ bounds the coordinate parameter gradients, and L is a common smoothness bound for F_t. The choice of L is positive even when $\zeta = 0$ or $J = 0$. One may also use any other fixed $L > 0$ with $L \ge \zeta J^2$ and define C_M using that same L. All these constants are independent of T for fixed memory and problem parameters.

Proof of Theorem 5 using the propositions. Combining the three propositions above gives:

$$\begin{aligned}
\sum_{t=1}^{T} g_t(M_t)^2 &\le C_M \sum_{t=1}^{T} \sum_{i=1}^{N} \|M_{i,t+1} - M_{i,t}\|_F^2 \\
&\le \frac{2C_M}{L} \sum_{t=1}^{T} \big[F_t(M_t) - F_t(M_{t+1})\big] \\
&\le \frac{2C_M}{L} \big[F_1(M_1) + V_c(T) + G_x Q\, V_w^+(T)\big].
\end{aligned}$$

Dividing by T proves (8). We now establish the propositions in the same order as above.

I.1 Proof of Proposition 3

First, recall that:

$$g_{i,t}(M_t) = F_t(M_t) - \min_{Q_i \in \mathcal{M}_i} F_t(Q_i, M_{-i,t}), \qquad g_t(M_t) = \max_{i \in [N]} g_{i,t}(M_t).$$

The minima exist because the domains are compact and F_t is continuous. Taking $Q_i = M_{i,t}$ shows that $g_{i,t}(M_t) \ge 0$. Consequently,

$$\sum_{t=1}^{T} g_t(M_t)^2 \le \sum_{t=1}^{T} \left(\sum_{i=1}^{N} g_{i,t}(M_t)\right)^2. \tag{81}$$

We relate the individual improvements to the parameter movements using the following proposition, proved in Section I.4.

Proposition 6. *For every $i \in [N]$, $t \ge 1$, and $Q_i \in \mathcal{M}_i$, the update (6) satisfies*

$$F_t(Q_i, M_{-i,t}) - F_t(M_t) \ge -(LD_i + G_{M,i})\|M_{i,t+1} - M_{i,t}\|_F, \tag{82}$$

with the constants in (77) and (80).

Taking the maximum improvement over Q_i yields

$$0 \le g_{i,t}(M_t) \le (LD_i + G_{M,i})\|M_{i,t+1} - M_{i,t}\|_F.$$

Summing over players and using Cauchy–Schwarz inequality gives

$$\begin{aligned}\left(\sum_{i=1}^{N} g_{i,t}(M_t)\right)^2 &\le \left(\sum_{i=1}^{N}(LD_i + G_{M,i})\|M_{i,t+1} - M_{i,t}\|_F\right)^2 \\ &\le \left(\sum_{i=1}^{N}(LD_i + G_{M,i})^2\right)\sum_{i=1}^{N}\|M_{i,t+1} - M_{i,t}\|_F^2 \\ &= C_M \sum_{i=1}^{N}\|M_{i,t+1} - M_{i,t}\|_F^2.\end{aligned}$$

Summing this inequality over t and using (81) proves the proposition.

I.2 Proof of Proposition 4

We first establish the regularity of the common diagonal potential, including the parameter-gradient bound used in Proposition 6.

Lemma 5 (State bounds, parameter gradients, and smoothness)**.** *For every $t \ge 1$ and every $M \in \mathcal{M}$, the rollout in (4) satisfies $\|y_t(M)\|_2 \le D_x$. The loss F_t is convex and differentiable, and*

$$\|\nabla_i F_t(M)\|_F \le G_{M,i}, \qquad \|\nabla F_t(M) - \nabla F_t(M')\|_F \le \zeta J^2 \|M - M'\|_F\,, \quad \forall M, M' \in \mathcal{M}.$$

In particular, F_t is L-smooth for the choice in (80).

Proof. Using the rollout formula (4), the disturbance bound, and joint strong stability, we obtain:

$$\begin{aligned}\|y_t(M)\|_2 &\le \sum_{r=1}^{h} \bar{\kappa} q^{r-1}\left(W + \sum_{i=1}^{N} b_i \sum_{p=1}^{H} a_{p-1} W\right) \\ &= W S_h (1 + B_\Sigma m_H) = D_x.\end{aligned}$$

Assumption 1, specialized to the state-only cost, therefore gives $\|\nabla c_t(x)\|_2 \le G D_x = G_x$ on $\|x\|_2 \le D_x$.

The map $M \mapsto y_t(M)$ is affine: its disturbance-only term is independent of M, and every remaining term is linear in the DAC blocks. Its derivative in coordinate i is the linear map

$$\mathcal{J}_{i,t}[\Delta M_i] = \sum_{r=1}^{h} \bar{A}^{r-1} B_i \sum_{p=1}^{H} \Delta M_i^{[p-1]} w_{t-r-p}. \tag{83}$$

Since

$$\sum_{p=0}^{H-1} \|\Delta M_i^{[p]}\|_{\mathrm{op}} \le \sum_{p=0}^{H-1} \|\Delta M_i^{[p]}\|_F \le \sqrt{H}\,\|\Delta M_i\|_F,$$

we obtain

$$\|\mathcal{J}_{i,t}[\Delta M_i]\|_2 \le W\sqrt{H}\, S_h b_i \|\Delta M_i\|_F = J_i \|\Delta M_i\|_F.$$

The full derivative $\mathcal{J}_t$ is the sum of the coordinate derivatives, so another application of Cauchy–Schwarz yields

$$\|\mathcal{J}_t[\Delta M]\|_2 \le \sum_{i=1}^{N} J_i \|\Delta M_i\|_F \le J \|\Delta M\|_F.$$

Because c_t is convex and differentiable, composition with the affine map y_t makes F_t convex and differentiable. The chain rule, with adjoints in the stated Euclidean and Frobenius inner products, gives

$$\nabla_i F_t(M) = \mathcal{J}_{i,t}^{*} \nabla c_t(y_t(M)).$$

Thus $\|\nabla_i F_t(M)\|_F \le J_i G_x = G_{M,i}$. Finally, $\mathcal{J}_t$ does not depend on M. Assumption 6 gives

$$\begin{aligned}\|\nabla F_t(M) - \nabla F_t(M')\|_F &\le J\,\|\nabla c_t(y_t(M)) - \nabla c_t(y_t(M'))\|_2 \\ &\le J\zeta\,\|y_t(M) - y_t(M')\|_2 \\ &\le \zeta J^2 \|M - M'\|_F.\end{aligned}$$

The choice $L \ge \zeta J^2$ proves the smoothness assertion. □

We now prove the descent inequality. The induced stage game has the common payoff F_t, so each player's coordinate gradient is $\nabla_i F_t$. Moreover, projection onto the product domain separates in the block Frobenius geometry:

$$\Pi_{\mathcal{M}}(Z) = \big(\Pi_{\mathcal{M}_1}(Z_1), \dots, \Pi_{\mathcal{M}_N}(Z_N)\big)$$

for every joint parameter vector Z. The simultaneous updates (6) are therefore exactly

$$M_{t+1} = \Pi_{\mathcal{M}}\left(M_t - L^{-1}\nabla F_t(M_t)\right). \tag{84}$$

By Lemma 5, uniform smoothness of F_t yields:

$$F_t(M_{t+1}) \le F_t(M_t) + \langle \nabla F_t(M_t), M_{t+1} - M_t\rangle + \frac{L}{2}\|M_{t+1} - M_t\|_F^2. \tag{85}$$

Using the characterization of the projection operator $\Pi_{\mathcal{M}}$, we have:

$$\forall M \in \mathcal{M}, \quad \langle M - M_{t+1}, M_t - \frac{1}{L}\nabla F_t(M_t) - M_{t+1}\rangle \le 0\,. \tag{86}$$

Then, using the above inequality with the feasible comparison point M_t gives:

$$\langle \nabla F_t(M_t), M_{t+1} - M_t\rangle \le -L\|M_{t+1} - M_t\|_F^2. \tag{87}$$

Combining (85) and (87),

$$F_t(M_{t+1}) \le F_t(M_t) - \frac{L}{2}\|M_{t+1} - M_t\|_F^2.$$

Rearranging and summing proves

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\|M_{i,t+1} - M_{i,t}\|_F^2 \le \frac{2}{L}\sum_{t=1}^{T}\big[F_t(M_t) - F_t(M_{t+1})\big],$$

as claimed. With $\eta = 1/L$, the coefficient is 2η.

I.3 Proof of Proposition 5

(i) Telescoping the potential decreases. First, the common cost is nonnegative under Assumption 1. Indeed, for every radius $D > 0$, that assumption gives

$$|c_t(0)| \le \beta D^2, \qquad \|\nabla c_t(0)\|_2 \le GD.$$

Letting D decrease to zero shows that $c_t(0) = 0$ and $\nabla c_t(0) = 0$. Convexity then implies

$$c_t(x) \ge c_t(0) + \langle \nabla c_t(0), x\rangle = 0.$$

In particular, $F_t(M) \ge 0$ for every feasible M. For any feasible sequence $M_1, \dots, M_{T+1}$, the exact telescoping identity therefore gives

$$\begin{aligned}\sum_{t=1}^{T}[F_t(M_t) - F_t(M_{t+1})] &= F_1(M_1) - F_T(M_{T+1}) \\ &\quad + \sum_{t=1}^{T-1}[F_{t+1}(M_{t+1}) - F_t(M_{t+1})] \\ &\le F_1(M_1) + \sum_{t=1}^{T-1}[F_{t+1}(M_{t+1}) - F_t(M_{t+1})].\end{aligned} \tag{88}$$

For $T = 1$, the temporal-variation sum is empty.

(ii) Separating cost variation and state variation. Let $M \in \mathcal{M}$ be a fixed joint profile. Then

$$F_{t+1}(M) - F_t(M) = c_{t+1}(y_{t+1}(M)) - c_t(y_{t+1}(M)) + c_t(y_{t+1}(M)) - c_t(y_t(M)).$$

The state bound in Lemma 5 gives

$$c_{t+1}(y_{t+1}(M)) - c_t(y_{t+1}(M)) \le \sup_{\|x\|_2 \le D_x} |c_{t+1}(x) - c_t(x)|. \tag{89}$$

Both states lie in the convex ball of radius D_x, where the gradient has norm at most G_x. Thus

$$c_t(y_{t+1}(M)) - c_t(y_t(M)) \le G_x \|y_{t+1}(M) - y_t(M)\|_2. \tag{90}$$

Define

$$\Delta^y_{t+1} := y_{t+1}(M_{t+1}) - y_t(M_{t+1}).$$

Both states in this difference use the same joint profile M_{t+1}. Inserting the preceding bounds into (88) yields

$$\sum_{t=1}^{T} [F_t(M_t) - F_t(M_{t+1})] \le F_1(M_1) + V_c(T) + G_x \sum_{t=1}^{T-1} \|\Delta^y_{t+1}\|_2. \tag{91}$$

(iii) Bounding the state variation. For every integer s, let $\delta_s = \|w_{s+1} - w_s\|_2$. The zero-padding convention gives

$$\delta_{-1} = \|w_0\|_2, \qquad \delta_s = 0\,, \forall s < -1, \qquad V^+_w(T) = \sum_{s=-1}^{T-2} \delta_s.$$

For the same fixed profile M, subtracting the two rollouts in (4) gives

$$\begin{aligned} y_{t+1}(M) - y_t(M) = &\sum_{r=1}^{h} \bar{A}^{r-1}(w_{t+1-r} - w_{t-r}) \\ &+ \sum_{r=1}^{h} \sum_{i=1}^{N} \sum_{p=1}^{H} \bar{A}^{r-1} B_i M_i^{[p-1]} (w_{t+1-r-p} - w_{t-r-p}). \end{aligned} \tag{92}$$

All DAC coefficients in this difference belong to that same profile. By joint strong stability and the block bounds,

$$\|y_{t+1}(M) - y_t(M)\|_2 \le \sum_{r=1}^{h} \bar{\kappa} q^{r-1} \delta_{t-r} + \sum_{r=1}^{h} \sum_{i=1}^{N} \sum_{p=1}^{H} \bar{\kappa} q^{r-1} b_i a_{p-1} \delta_{t-r-p}. \tag{93}$$

This bound is uniform over $M \in \mathcal{M}$. For each fixed r, p, every index in either shifted sum

$$\sum_{t=1}^{T-1} \delta_{t-r}, \qquad \sum_{t=1}^{T-1} \delta_{t-r-p}$$

is at most $T-2$, and terms with indices below -1 are zero. Each shifted sum is therefore at most $V^+_w(T)$. Taking the supremum over M in (93) and summing gives

$$\begin{aligned} \sum_{t=1}^{T-1} \|\Delta^y_{t+1}\|_2 &\le \sum_{t=1}^{T-1} \sup_{M \in \mathcal{M}} \|y_{t+1}(M) - y_t(M)\|_2 \\ &\le \left[\sum_{r=1}^{h} \bar{\kappa} q^{r-1} \left(1 + \sum_{i=1}^{N} b_i \sum_{p=1}^{H} a_{p-1} \right) \right] V^+_w(T) \\ &= S_h(1 + B_\Sigma m_H) V^+_w(T) = Q\, V^+_w(T). \end{aligned} \tag{94}$$

(iv) Combining the bounds. Substituting (94) into (91) gives

$$\sum_{t=1}^{T} [F_t(M_t) - F_t(M_{t+1})] \le F_1(M_1) + V_c(T) + G_x Q\, V^+_w(T),$$

which proves (74).

I.4 Proof of Proposition 6

Let $i \in [N]$, $t \geq 1$, and $Q_i \in \mathcal{M}_i$. The proof follows four steps.

(i) Convexity. By Lemma 5, the common diagonal loss is convex. In particular,

$$\begin{aligned} F_t(Q_i, M_{-i,t}) - F_t(M_t) &\geq \langle \nabla_i F_t(M_t), Q_i - M_{i,t} \rangle \\ &= \langle \nabla_i F_t(M_t), Q_i - M_{i,t+1} \rangle \\ &\quad + \langle \nabla_i F_t(M_t), M_{i,t+1} - M_{i,t} \rangle. \end{aligned} \tag{95}$$

(ii) Lower bound for the first inner product. The update (6) and projection optimality give:

$$\left\langle Q_i - M_{i,t+1}, M_{i,t} - L^{-1} \nabla_i F_t(M_t) - M_{i,t+1} \right\rangle \leq 0.$$

Rearranging, using Cauchy–Schwarz inequality and the diameter bound, we obtain:

$$\begin{aligned} \langle \nabla_i F_t(M_t), Q_i - M_{i,t+1} \rangle &\geq L \langle M_{i,t} - M_{i,t+1}, Q_i - M_{i,t+1} \rangle \\ &\geq -L \|M_{i,t+1} - M_{i,t}\|_F \cdot \|Q_i - M_{i,t+1}\|_F \\ &\geq -LD_i \|M_{i,t+1} - M_{i,t}\|_F. \end{aligned} \tag{96}$$

(iii) Lower bound for the second inner product. Lemma 5 gives the parameter gradient bound $\|\nabla_i F_t(M_t)\|_F \leq G_{M,i} = G_x J_i$. Therefore,

$$\begin{aligned} \langle \nabla_i F_t(M_t), M_{i,t+1} - M_{i,t} \rangle &\geq -\|\nabla_i F_t(M_t)\|_F \|M_{i,t+1} - M_{i,t}\|_F \\ &\geq -G_{M,i} \|M_{i,t+1} - M_{i,t}\|_F. \end{aligned} \tag{97}$$

(iv) Combining the bounds. Substituting (96) and (97) into (95) yields

$$F_t(Q_i, M_{-i,t}) - F_t(M_t) \geq -(LD_i + G_{M,i}) \|M_{i,t+1} - M_{i,t}\|_F.$$

This is the claimed inequality.

I.5 Proof of Lemma 1

We return here to the general individual cost setting. Recall that the loss function ℓ_t^i is defined for every $M_i \in \mathcal{M}_i$ by

$$\ell_t^i(M_i) = c_t^i(y_t^{i,K_i}(M_i), v_t^{i,K_i}(M_i)), \tag{98}$$

where the counterfactual idealized state $y_t^{i,K_i}(M_i)$ and action $v_t^{i,K_i}(M_i)$ are defined in section D.1.

The convexity argument uses only Assumption 1. Fix a player i, a time t, and the realized sequence of opponents' controls used in the unilateral comparison. Let $d_s = w_s + \sum_{j \neq i} B_j u_s^j$ for $s \geq 0$ denote the resulting external input. The DAC policy uses the effective disturbance sequence $r_s = d_s$ under state-only feedback and $r_s = w_s$ under aggregate-input feedback. Set $d_s = r_s = 0$ for $s < 0$. Both sequences are held fixed when varying the comparator parameter M_i.

Starting from the fixed zero state at the truncation boundary, the unilateral rollout gives:

$$z_{s+1}(M_i) = (A - B_i K_i) z_s(M_i) + B_i \sum_{p=1}^{H} M_i^{[p-1]} r_{s-p} + d_s.$$

Its dependence on M_i is affine by induction. Hence the terminal idealized state $y_t^{i,K_i}(M_i)$ and action

$$v_t^{i,K_i}(M_i) = -K_i y_t^{i,K_i}(M_i) + \sum_{p=1}^{H} M_i^{[p-1]} r_{t-p}$$

are affine in M_i. Assumption 1 makes c_t^i jointly convex, so

$$\ell_t^i(M_i) = c_t^i\big(y_t^{i,K_i}(M_i), v_t^{i,K_i}(M_i)\big)$$

is convex as the composition of a convex function and an affine map. This argument holds for any realized opponents' control sequence, including one generated by DAC policies.

J Tools from Online Convex Optimization

J.1 Online convex optimization with memory

We use the following explicit interval form of the online gradient descent with memory argument of [5]. The updates begin at round one, even when regret is evaluated only on a later interval.

Algorithm 2 Online Gradient Descent with Memory

1: **Input:** nonempty closed convex domain $\mathcal{K}$, memory $h \geq 1$, step size $\eta > 0$, and $z_{\text{init}} \in \mathcal{K}$.
2: Set $z_{1-h} = \cdots = z_1 = z_{\text{init}}$.
3: **for** $t = 1, \ldots, T$ **do**
4: Play z_t and incur $\ell_t(z_{t-h}, \ldots, z_t)$.
5: Observe the diagonal loss $f_t(z) := \ell_t(z, \ldots, z)$ or its required gradient.
6: Set $z_{t+1} = \Pi_{\mathcal{K}}(z_t - \eta \nabla f_t(z_t))$.
7: **end for**

Theorem 6 (Online gradient descent with memory)**.** *Let $\mathcal{K}$ be a nonempty closed convex subset of a finite-dimensional Euclidean space with diameter at most $D_0 > 0$. Let $\ell_t : \mathcal{K}^{h+1} \to \mathbb{R}$ be a sequence of losses such that:*

1. *There exists $L_0 > 0$ s.t. for any $z_0, \ldots, z_h, \widetilde{z}_j \in \mathcal{K}$, whenever two argument tuples differ only in coordinate $j \in \{0, \ldots, h\}$,*

$$\left|\ell_t(z_0, \ldots, z_j, \ldots, z_h) - \ell_t(z_0, \ldots, \widetilde{z}_j, \ldots, z_h)\right| \leq L_0 \|z_j - \widetilde{z}_j\|.$$

2. *The diagonal loss $f_t(z) := \ell_t(z, \ldots, z)$ is convex and differentiable, and there is $G_0 > 0$ such that $\|\nabla f_t(z)\| \leq G_0$ for every $z \in \mathcal{K}$ and every t.*

Then Algorithm 2 satisfies, for every $1 \leq a \leq b \leq T$,

$$\sum_{t=a}^{b} \ell_t(z_{t-h:t}) - \min_{z \in \mathcal{K}} \sum_{t=a}^{b} f_t(z) \leq \frac{D_0^2}{2\eta} + \frac{\eta(b-a+1)}{2}\left[G_0^2 + L_0 G_0 h(h+1)\right]. \tag{99}$$

In particular, choosing

$$\eta = \frac{D_0}{\sqrt{T[G_0^2 + L_0 G_0 h(h+1)]}}$$

bounds the left-hand side by $D_0\sqrt{T[G_0^2 + L_0 G_0 h(h+1)]}$.

Proof. For any fixed $z \in \mathcal{K}$, projection nonexpansiveness gives

$$\|z_{t+1} - z\|^2 \leq \|z_t - \eta \nabla f_t(z_t) - z\|^2.$$

Expanding, rearranging, and using convexity and the gradient bound, we obtain:

$$f_t(z_t) - f_t(z) \leq \frac{\|z_t - z\|^2 - \|z_{t+1} - z\|^2}{2\eta} + \frac{\eta G_0^2}{2}.$$

Summing from a to b and using the diameter bound yields:

$$\sum_{t=a}^{b} [f_t(z_t) - f_t(z)] \leq \frac{D_0^2}{2\eta} + \frac{\eta(b-a+1)G_0^2}{2}. \tag{100}$$

Also, since $\Pi_{\mathcal{K}}(z_t) = z_t$, and using nonexpansiveness of the projection, we have for all $t \geq 1$,

$$\|z_{t+1} - z_t\| = \|\Pi_{\mathcal{K}}(z_t - \eta \nabla f_t(z_t)) - \Pi_{\mathcal{K}}(z_t)\| \leq \eta \|\nabla f_t(z_t)\| \leq \eta G_0\,.$$

Consequently, for every $1 \leq k \leq h$,

$$\|z_t - z_{t-k}\| \leq k\eta G_0,$$

including when $t - k \leq 0$. Changing the arguments of the memory loss one at a time and using the coordinate-wise Lipschitzness assumption gives:

$$\ell_t(z_{t-h:t}) - f_t(z_t) \leq L_0 \sum_{k=1}^{h} \|z_{t-k} - z_t\| \leq \frac{\eta L_0 G_0 h(h+1)}{2}. \tag{101}$$

Combining (100) and (101), summing, and minimizing over the fixed comparator proves (99). The minimum exists because $\mathcal{K}$ is compact and the losses are continuous. The stated step-size conclusion uses $b - a + 1 \leq T$. □

In our control setting, the Euclidean geometry is the block Frobenius geometry, $h = H+1$, $a = H+1$, and $b = T$. The constants D_0, G_0, L_0 respectively denote parameter diameter, diagonal gradient bound, and memory-coordinate Lipschitz constant, and are distinct from the state bound D and the cost-gradient growth constant G.

J.2 Time regret decomposition

Lemma 6. *For every player $i \in [N]$, every fixed realized opponent control sequence $\{u_t^{-i}\}$, every comparator class Π_i for which the minima are defined, and integers $1 \leq H < T$,*

$$\mathrm{Reg}_i^{1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i) \leq \mathrm{Reg}_i^{1:H}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i) + \mathrm{Reg}_i^{H+1:T}(\mathcal{A}_i, \{u_t^{-i}\}, \Pi_i).$$

Every counterfactual trajectory in these interval regrets starts at time zero from the same original initial state; no state is reset at the interval boundary.

Proof. For any comparator policy π^i, write

$$A_{\pi^i} := \sum_{t=1}^{H} c_t^i(x_t^{\pi^i}, u_t^{\pi^i}), \qquad B_{\pi^i} := \sum_{t=H+1}^{T} c_t^i(x_t^{\pi^i}, u_t^{\pi^i}).$$

Then, we have:

$$\min_{\pi^i \in \Pi_i} (A_{\pi^i} + B_{\pi^i}) \geq \min_{\pi^i \in \Pi_i} A_{\pi^i} + \min_{\pi^i \in \Pi_i} B_{\pi^i}.$$

Subtracting this inequality from the learner's total cost and splitting that cost over $1{:}H$ and $H+1{:}T$ proves the result using (1). □

K Simulations

While our contributions are mainly theoretical, we also performed simple simulations to illustrate the algorithm's behavior.

We implemented Algorithm 1 in both settings (independent learning and aggregated control learning) for $N = 3$ agents with distinct quadratic cost functions, using a two-dimensional instance of (LDS). We considered three types of disturbance sequences: (1) a constant nonzero disturbance, (2) sinusoidal disturbances (with different frequencies for each coordinate), and (3) Gaussian noise. See sections below for details of the setup, vanishing per-agent time-average cost, control input curves, and 2D state trajectories. The simulations were run on Google Colab and only take a few seconds to minutes to be completed.

K.1 Setting

We consider a 2-dimensional ($d = 2$) LDS with $N = 3$ agents and scalar control inputs ($k_i = 1$ for $i \in \{1, 2, 3\}$) with:

$$A = \begin{bmatrix} 0.95 & 0.1 \\ 0 & 0.9 \end{bmatrix}, \quad B_1 = B_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_0 = \begin{bmatrix} 6 \\ 6 \end{bmatrix}. \tag{102}$$

Each agent $i \in \{1, 2, 3\}$ has their quadratic cost function $c_t^i(x, u^i) = x^\top Q_i x + r_i (u^i)^2$ where:

$$Q_i = \begin{bmatrix} 1 + 0.8i & 0 \\ 0 & 1 + 0.4(N - 1 - i) \end{bmatrix}, \quad r_i = 0.1(1 + 0.4i), \quad u^i \in \mathbb{R}. \tag{103}$$

The cost functions reflect different distances to the origin goal state $(0, 0)$, see figures 1, 2, 3 below (bottom right subplots).

We test Algorithm 1 with three different kinds of disturbances $w_t \in \mathbb{R}^2$:

(1) Constant disturbance: $w_t = 0.7$ (see Fig. 1),

(2) Sinusoidal disturbance: $w_{t,1} = \sin(0.1t), w_{t,2} = \sin(0.1t)$ where $w_{t,1}, w_{t,2}$ are the coordinates of w_t (see Fig. 2),

(3) Independent and identically distributed Gaussian: $w_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$ (see Fig. 3).[3]

For the hyperparameters of Algorithm 1, we set $T = 500$ for the time horizon, $\eta = 10^{-4}$ for the step size, $H = 5$ for the memory parameter and DAC policy parameters are initialized with zero values.

K.2 Comments

We make a few remarks on the results of the simulations (see Figs. 1, 2, 3 above):

- Starting from the initial state x_0 (see bottom right subplots in all the figures), the state evolves quickly towards the goal origin state minimizing the costs by strong stability of the controllers. In particular, this quick phase corresponds to an application of higher control inputs by all agents in all three disturbance scenarios (see Figs. 1 to 3, top right subplots). Then the control inputs have a similar shape to the disturbances themselves to stabilize the system (almost constant in the first case, sinusoidal in the second and random Gaussian in the third case).
- The plots show a reduction in per-agent time-average costs over the simulated horizon.
- It can be seen in all the figures that there is a slight advantage to the ACL setting (which can infer the disturbance values) compared to the independent learning setting in our simple simulation setting. Compare for instance the dotted per-agent time-average cost curves to the plain ones in bottom left subplots of all three figures.
- We can also observe from all the state trajectories (bottom right subplots) that Algorithm 1 in the ACL setting is more stable than in the IL setting. For instance, in the sinusoidal case (see Fig. 2,

[3]The untruncated Gaussian experiments are illustrative stress tests outside the deterministic bounded disturbance assumption of our regret theorems.

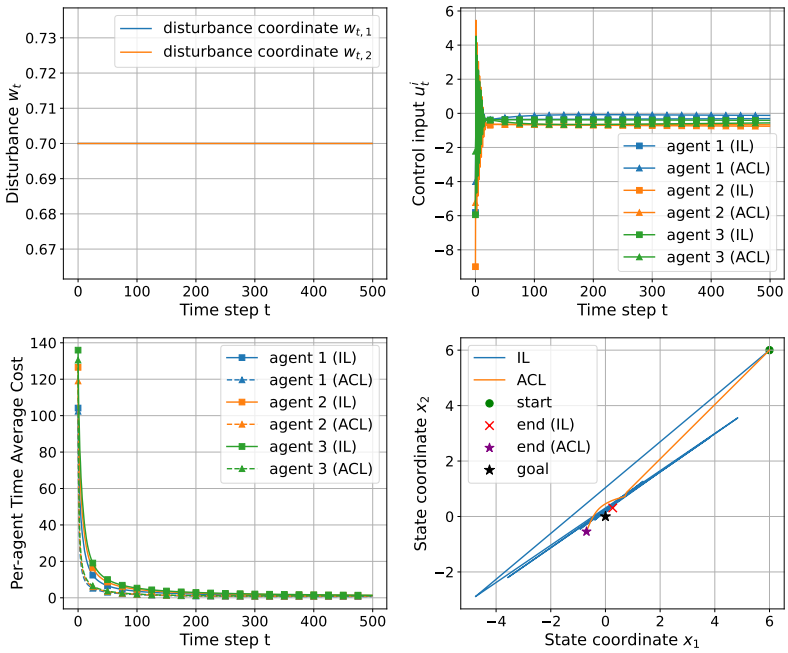

Figure 1: Illustration of the performance of Algorithm 1 on a simple multi-agent LDS with constant disturbance sequence. 'IL' stands for Independent Learning (see Information Setting 1), 'ACL' for Aggregated Control Learning (see Information Setting 2), 'start' refers to the initial state x_0, 'end' to the state at the last time step for $t = T$.

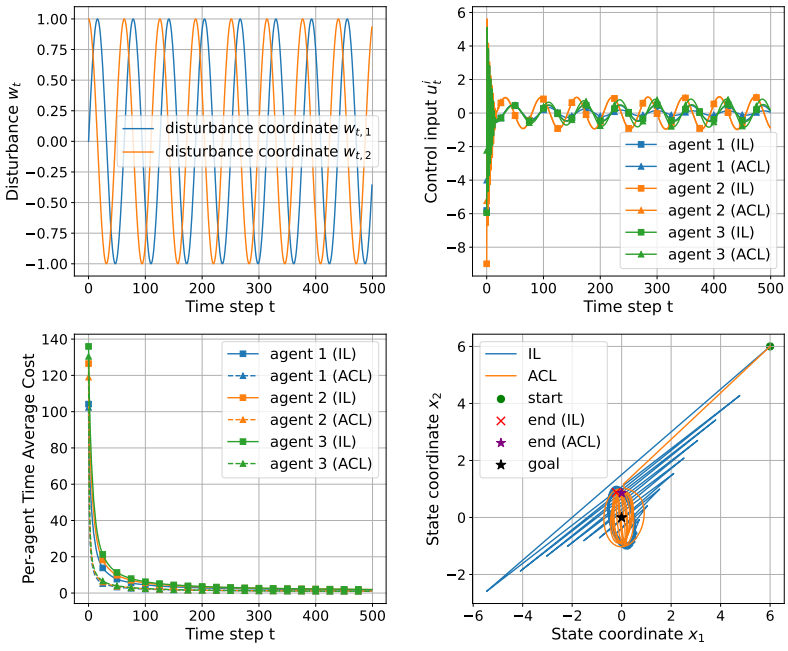

Figure 2: Illustration of the performance of Algorithm 1 on a simple multi-agent LDS with sinusoidal disturbance sequence. 'IL' stands for Independent Learning (see Information Setting 1, 'ACL' for Aggregated Control Learning (see Information Setting 2), 'start' refers to the initial state x_0, 'end' to the state at the last time step for $t = T$.

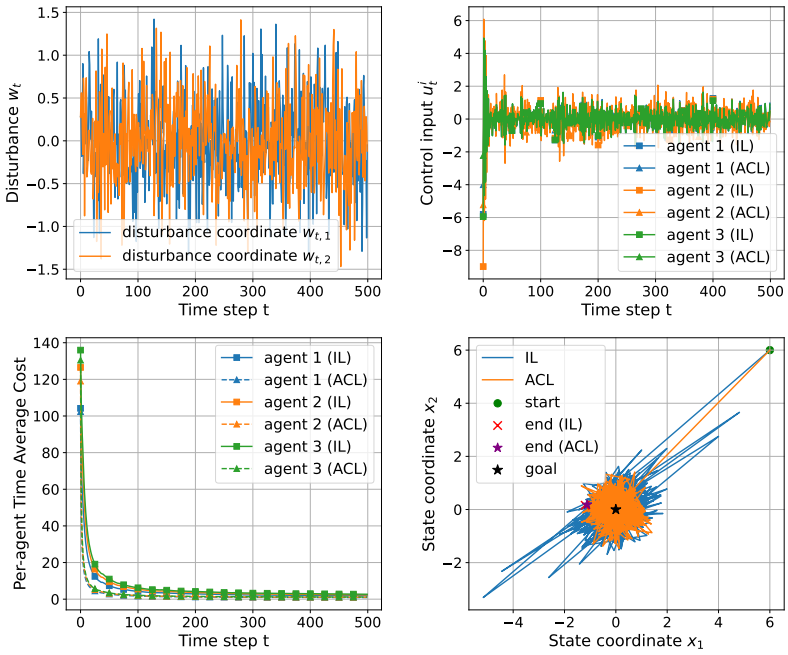

Figure 3: Illustration of the performance of Algorithm 1 on a simple multi-agent LDS with Gaussian disturbance sequence. 'IL' stands for Independent Learning (see Information Setting 1, 'ACL' for Aggregated Control Learning (see Information Setting 2), 'start' refers to the initial state x_0, 'end' to the state at the last time step for $t = T$.

bottom right subplot), there are less oscillations and their magnitude is smaller in the ACL setting as expected from our theory. In particular, the state trajectory converges to a neighborhood of the goal state defined by the amplitude of the disturbance sequence. The same observation can be made in the case of the Gaussian noise disturbance where the state trajectory in the ACL setting concentrates more around the origin than in the IL setting as expected. The region of concentration is controlled by the standard deviation of the Gaussian noise disturbance sequence.

K.3 Additional experiments regarding stability

We run additional simulations in dimension 8 with 5 agents under adversarial disturbances, and a sensitivity analysis over several values of a stability parameter ρ. These track both the time-averaged total cost and the state norm over 200 iterations; see Figures 4, 5 and 6. The main takeaway is that performance degrades smoothly as the stability margin decreases and deteriorates near instability, consistent with the role of strong stability in our analysis.

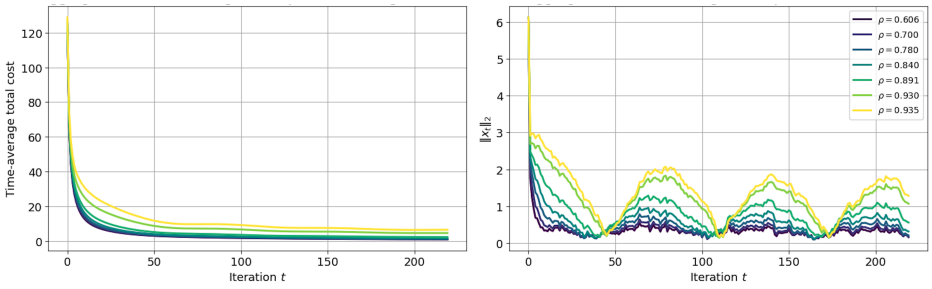

Figure 4: Illustration of the performance of Algorithm 1 on a multi-agent LDS with $N = 5, d = 8$ and composite adversarial disturbance (randomly additively mixing sin, cosine and Gaussian perturbations with different amplitudes, phases and variance respectively).

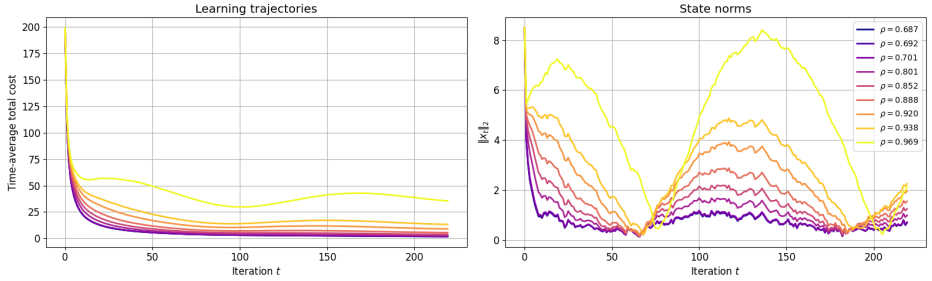

Figure 5: Same setting with only cosine disturbance.

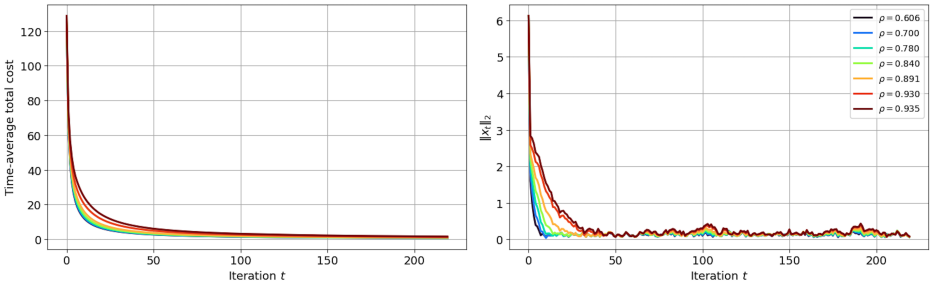

Figure 6: Same setting with only Gaussian disturbances.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Contributions are stated in both the abstract and introduction consistently.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please see conclusion for a discussion to extend our model.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: Assumptions are specified for each result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: This is a theoretical paper. We provide a few simulations in the appendix for illustration, the code is entirely reproducible by running the provided Jupyter notebook scripts.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We provide the code for reproducing the 6 figures in the appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: The paper is theoretical. We provide the code with all the details to reproduce the 6 figures in the appendix which serve an illustrative purpose.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [No]

Justification: The paper is theoretical. In the simulations in the appendix, we observe similar qualitative results for different runs, we only report the average cost for legibility. Since the main goal is to consider adversarial disturbances that are not necessarily stochastic, we report individual runs for different configurations rather than averaging possibly unrelated results corresponding to different sequences of disturbances.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Experiments were run on Google Colab and only take a few seconds to minutes. This is reported in the simulations section in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is theoretical and we do not foresee any potentially harmful consequence.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.