2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

# On the performance of memory-augmented controllers*

Farnaz Adib Yaghmaie[1], Hamidreza Modares[2], and Bahare Kiumarsi[2]

*Abstract*— **Recently, online convex optimization techniques have been utilized to develop online algorithms for controlling linear systems under adversarial disturbances [1], [2], [3]. This approach involves introducing a class of memory-augmented controllers, also known as disturbance-action controllers, and learning their parameters online to optimize general convex functions. The performance of the controller is measured using the concept of regret, which compares its performance to a benchmark. However, while regret is an important metric for algorithm performance, it does not directly address the boundedness of the state variable. In this paper, we investigate the conditions under which boundedness can be inferred from regret, and vice versa, for the class of memory-augmented controllers. Our analysis is independent of the specific controller design, making it applicable to any algorithm or learning procedure, as long as the specified conditions are satisfied.**

## I. INTRODUCTION

Online Convex Optimization (OCO) techniques can be used to design algorithms making optimal decisions under uncertainty and disturbances [4], [5]. In this setting, a decision variable is selected and an *a priori* unknown cost is suffered. An algorithm is then designed to map the available history of measurements and cost functions to a decision variable. Regret quantifies the performance of the algorithms by comparing the incurred cost of by the online algorithm with a baseline. It is common to select the the baseline as the cost assuming full knowledge of the problem [6].

Recently, OCO has been used to design online optimal policies for dynamical systems subject to adversarial disturbances. The proposed algorithms within the OCO framework are online, capable of adjusting according to the properties of the disturbances and costs. When the policy is selected from a predefined policy class, the associated regret is called policy regret while dynamic regret refers to the case with no restrictions on policy class is exposed [7].

A common framework for studying optimal control problem is the class of linear systems, usually with a quadratic cost. If the linear system is subject to Gaussian disturbance (noise) on the system dynamics and no noise on the system's state measurements, linear quadratic regulator (LQR) control can be used to design an optimal controller by minimizing a quadratic cost [8], [9]. The policy regret of the LQR problem is studied in [10], [11], [12], [13]. If the disturbance is non-Guassian but has a limited-energy, one can use the $H_\infty$-control theory to guarantee an $\mathcal{L}_2$-gain performance bound [14], [15]. The $H_\infty$ approach is typically overly conservative as the resulting controller hedges against the worst-case disturbance, which rarely occurs in reality. The regret of $H_\infty$-controller is discussed in [16].

However, in many practical control systems, the distribution of the disturbance is neither Gaussian nor worst-case. To avoid the design of an overly-conservative controller while optimizing a general convex cost function, the class of memory-augmented policies (also called disturbance-action policy) is introduced in [1], [2], [3], [17]. The main property of the memory-augmented policy class is a neat parameterization of the policy from which any general convex cost function can be optimized using OCO.

The proposed algorithms in [1], [2], [3], [17] achieve a sublinear regret bound implying that the cost by the algorithm converges with at least a sublinear rate to the baseline. To study memory-augmented policy in the context of control theory, one needs to go beyond the regret and analyzes if stability can be concluded from regret. The relationship between the regret and stability for disturbance-free linear systems with linear feedback controllers and nonlinear systems is studied in [18] and [19] respectively.

In this paper, we consider linear systems subject to general adversarial disturbances. For the class of memory-augmented policies, we aim to specify the conditions to conclude boundedness of the state variable from a linear regret and vice versa. Note that we study boundedness instead of stability as the dynamical system is subject to general adversarial disturbances. The stability in the absence of disturbance is concluded as a special case in our analysis. To cover a wider class of problems, we bring a linear tracking problem, where a controller is designed to track a linear reference signal [20], [21]. The regulation problem can be considered as a special case by setting the reference signal to zero. The contribution of this paper is as follows:

- We give the conditions to guarantee linear regret bounds for the class of memory-augmented policies independent of how the memory-augmented controller is designed.
- We specify the conditions to infer boundedness of the state variable from linear regrets.

The organization of this paper is as follows. In Section II we define the optimal tracking problem and give the

*Farnaz Adib Yaghmaie is supported by the Excellence Center at Linköping–Lund in Information Technology (ELLIIT), ZENITH, and partially by Sensor informatics and Decision-making for the Digital Transformation (SEDDIT). Hamidreza Modares and Bahare Kiumarsi are supported by the Department of Navy award N00014-22-1-2159 issued by the Office of Naval Research

[1]Farnaz Adib Yaghmaie is with the Faculty of Electrical Engineering, Linköping University, Linköping, Sweden farnaz.adib.yaghmaie@liu.se.

[2]Hamidreza Modares is with the Department of Mechanical Engineering, Michigan State University, Michigan, USA modaresh@msu.edu. Bahare Kiumarsi is with the Department of Electrical and Computer Engineering (ECE), Michigan State University, Michigan, USA kiumarsi@msu.edu.

assumptions. In Section III, we define the class of memory-augmented control policy and regret. Section IV contains the main results of this paper and specifies the conditions to guarantee boundedness of the state variable from a linear regret and vice versa. Section V concludes the paper.

## II. Optimal Reference Tracking Problem

*Notations and preliminaries:* Let $I$ denote an identity matrix with appropriate dimension. Let $\mathbf{1}$ and $\mathbf{0}$ denote one and zero matrices with appropriate dimensions respectively. Let $\|x_k\|$ denote the instantaneous Euclidean norm of the vector $x_k$. For matrix $A$, the spectral norm is denoted by $\|A\|$ and the Frobenius norm is denoted by $\|A\|_F$. Let $\mathbb{I}_E$ be an indicator function on set $E$. For a time-dependent variable $x_k$, the notation $x_{i:j}$, $j \geq i$ is defined as $x_{i:j} = \{x_i, x_{i+1}, .., x_j\}$. The notation $\mathcal{O}()$ is leveraged throughout the paper to express the regret upper bound as a function of $T$.

*Definition 1:* [3] Consider

$$x_{k+1} = Ax_k + Bu_k$$

and $\gamma \in [0, 1)$, $\kappa > 1$. A linear controller $K$ is $(\kappa, \gamma)$-stable if $\|K\| \leq \kappa$ and $\|\tilde{A}_K^t\|_2 \leq \kappa^2(1-\gamma)^t \ \forall\, t \geq 0$ where $\tilde{A}_K = A + BK$.

### A. The tracking problem

Consider the following linear dynamical system

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ and $u_k \in \mathbb{R}^m$ denote the state and the control input of the system, respectively. In (1), $w_k \in \mathbb{R}^n$ denotes the adversarial (arbitrary and unknown) disturbance. We assume that $x_0 = \mathbf{0}$ and absorb the initial condition of the system into $w_0$ without loss of generality.

In this paper, we consider a tracking problem to design $u_k$ such that the state of the system $x_k$ tracks an unknown linear reference signal $r_k$ generated by

$$\begin{aligned} z_{k+1} &= Sz_k, \\ r_k &= Fz_k, \end{aligned} \tag{2}$$

where $z_k \in \mathbb{R}^p$ and $r_k \in \mathbb{R}^n$ denote the state and output of the reference signal, respectively. Let $e_k$ denote the state tracking error

$$e_k = x_k - r_k. \tag{3}$$

If regulation of the dynamical system in (1) is of concern, one can neglect the reference signal and set the relevant variables equal to zero in the derivations.

We made standard assumptions regarding (1)-(2).

*Assumption 1 (dynamical system):* The pair $(A, B)$ is known and stabilizable. Moreover, the system matrices are bounded, i.e., $\|A\| \leq \kappa_a$ and $\|B\| \leq \kappa_b$.

*Assumption 2 (disturbance):* The disturbance sequence $w_k$ is bounded, i.e., $\|w_k\| \leq \kappa_w$ for some $\kappa_w > 0$. Moreover, $w_k = \mathbf{0}$ for $k < 0$.

*Assumption 3 (reference signal):* The following assumptions are made on the reference signal

- The pair $(S, F)$ is unknown, but observable.
- The state of the reference signal $z_k$ is not measurable but the output $r_k$ is measurable.
- The reference signal $r_k$ is bounded, i.e., $\|r_k\| \leq \kappa_r$.

Since the system dynamics are assumed to be known in Assumption 1, at each time $k$, $w_{1:k-1}$ are known. This is because $w_{k-1} = x_k - Ax_{k-1} - Bu_{k-1}$ and the state $x_k$ is assumed measurable.

The following theorem brings the necessary and sufficient condition to the reference tracking in the absence of disturbances.

*Theorem 1:* [22][Theorem 1.35 and Remark 1.36] Consider (1)-(2) and let $w_k \equiv \mathbf{0}$, $k > 0$. Assume that $(A, B)$ is stabilizable and $(S, F)$ is detectable. Select $K_{fb}$ such that $A + BK_{fb}$ is strongly stable. Then, the controller

$$u_k = K_{fb}x_k + (\Gamma - K_{fb}\Pi)z_k \tag{4}$$

solves the classical state tracking problem $x_k \to r_k$ if and only if there exist matrices $\Pi \in \mathbb{R}^{n\times p}$ and $\Gamma \in \mathbb{R}^{m\times p}$ such that

$$\Pi S = A\Pi + B\Gamma, \quad \Pi - F = \mathbf{0}. \tag{5}$$

It has been shown in Lemma 1 of [1] that one can extract $z_k$ from the current and past outputs of the reference.

*Lemma 1:* Assume that $(S, F)$ is observable. Let $l$ denote the observability index of (2); i.e., the smallest positive integer $l \geq 1$ such that

$$\mathcal{O}_l = \begin{bmatrix} F \\ \vdots \\ FS^{l-1} \end{bmatrix} \in \mathbb{R}^{nl\times p} \tag{6}$$

has full column rank. That is, $\text{rank}(\mathcal{O}_l) = p$. Let

$$\begin{aligned} &\mathcal{O}_l^+ = (\mathcal{O}_l^T\mathcal{O}_l)^{-1}\mathcal{O}_l^T, \\ &N = \begin{bmatrix} N^{[1]} & \dots & N^{[l]} \end{bmatrix} = S^{l-1}\mathcal{O}_l^+, \\ &N^{[s]} \in \mathbb{R}^{p\times n}, s = 1, ..., l. \end{aligned} \tag{7}$$

Then, the state of the reference signal can be expressed as a linear function of the current and $l-1$ past outputs of the reference

$$z_k = \sum_{q=0}^{l-1} N^{[l-q]} r_{k-q}. \tag{8}$$

### B. The performance index

For the dynamical system in (1) and the reference signal in (2), it is common to define a total cost function and optimize it with designing a control policy $\pi : (x_{1:k}, w_{1:k-1}, r_{1:k}) \to u_k$. The total cost associated with a control policy $\pi$ is defined as

$$J_T(\pi) = \sum_{k=1}^{T} c_k(e_k, u_k). \tag{9}$$

Note in this paper we only consider total costs and do not study discounted costs as the regret (to be defined in the next section) is usually defined for total costs [23] and considering discounted costs result in pathological cases as detailed in [18].

We make the following assumption regarding the cost function $c_k(e_k, u_k)$ in (9).

*Assumption 4 (cost function):* The cost $c_k(e_k, u_k)$ is convex in $e_k$, $u_k$. Moreover, when $\|e\|$, $\|u\| \leq D < \infty$, it holds that $|c_k(e_k, u_k)| \leq \beta D^2$ and $\|\nabla_e c_k(e, u)\|, \|\nabla_u c_k(e, u)\| \leq G_c D$ for some $0 < \beta < \infty$ and $0 < G_c < \infty$.

Assumption 4 limits the cost function to be convex, which is more general than typical quadratic cost functions.

## III. Memory-augmented control policy

For (1) and the reference signal in (2), in the presence of an adversarial or arbitrary disturbance, it is common to design a linear feedback controller using $H_\infty$-control. However, besides conservativeness imposed by $H_\infty$-design, the cost function $c_k(e_k, u_k)$ is not convex in the linear feedback controller gains which makes the online control design intractable. To circumvent this difficulty, one can define the class of Memory-augmented control policies [1], [3].

In this section, we define the class of Memory-augmented control policies. We show that a linear feedback policy can be considered as a special case of this class. At the end, we define the regret function to quantify the performance of the memory-augmented control policies.

### A. Memory-augmented control policy

A memory-augmented control policy is defined as follows.

*Definition 2:* A memory-augmented control policy $\pi(K, M, P)$ is specified by

$$\begin{aligned} u_k^\pi(K, M, P) =& Kx_k + \sum_{t=1}^{m_w} M^{[t-1]} w_{k-t} \\ &+ \sum_{s=0}^{m_r - 1} P^{[s]} r_{k-s}, \end{aligned} \tag{10}$$

where $K$ is a fixed controller gain and the parameters

$$\begin{aligned} M &= [M^{[0]}, ..., M^{[m_w - 1]}], \\ P &= [P^{[0]}, ..., P^{[m_r - 1]}], \\ Y &= [M,\ P], \end{aligned}$$

are learnable.

Since the parameters $M$, $P$ are being learned and thus are changing over time, we refer to $Y_k = [M_k,\ P_k]$ as the policy parameters at time $k$. Let $\tilde{A}_K = A + BK$ and define

$$\begin{aligned} \Psi_{k,y}^{K,h}(M_{k-h-1:k-1}) :=& \tilde{A}_K^y \mathbb{I}_{y \leq h-1} \\ &+ \sum_{j=0}^{h-1} \tilde{A}_K^j B M_{k-j-1}^{[y-j-1]} \mathbb{I}_{1 \leq y-j \leq m_w}, \end{aligned} \tag{11}$$

$$\psi_{k,z}^{K,h}(P_{k-h-1:k-1}) := \sum_{j=0}^{h-1} \tilde{A}_K^j B P_{k-j-1}^{[z-j-1]} \mathbb{I}_{1 \leq z-j \leq m_r}.$$

In (11), $h > 0$ is the memory length, usually used in the online convex optimization context.

Let $x_k^\pi$ and $e_k^\pi$ denote the trajectory of the system (1) and the tracking error upon execution of the memory-augmented policy $u_k^\pi$. The following lemma specifies $x_k^\pi$.

*Lemma 2 (Lemma 3 in [1]):* Let $x_k^\pi$ be the state attained upon execution of the policy $\pi(K, M_{0:k-1}, P_{0:k-1})$ that generates the control input in (10) at time $k$. Then,

$$x_k^\pi = \sum_{y=0}^{k-1} \Psi_{k,y}^{K,k}(M_{0:k-1}) w_{k-y-1} + \sum_{z=0}^{k-1} \psi_{k,z}^{K,k}(P_{0:k-1}) r_{k-z}. \tag{12}$$

### B. Linear feedback policy

It is also useful in our analysis to define a linear feedback policy

$$u_k^{\text{lin}}(K_{fb}, K_{ff}) = K_{fb} x_k + K_{ff} z_k. \tag{13}$$

*Remark 1:* One can see that linear feedback policies can be considered as a special case of memory-augmented control policies. Set $K \equiv K_{fb}$, $M \equiv \mathbf{0}, m_r = l$ and $P^{[s]} = N^{[l-s]}$, $s = 0, ..., m_r - 1$ in the memory-augmented policy (10) to get a linear feedback policy.

### C. Regret

The standard measure for online control algorithm is the policy regret [4], [24] which is defined as the difference between the total cost of Algorithm $\mathcal{A}$ and a baseline

$$\mathcal{R}_T(\mathcal{A}) = J_T(\mathcal{A}) - b_T. \tag{14}$$

Note that $b_T$ is user specified and it usually denotes the optimal total cost achievable from a specified policy class. A linear regret bound is defined as follows.

*Definition 3:* The algorithm $\mathcal{A}$ has a linear regret $\mathcal{O}(T)$ if there exists $C_0, C_1 \in \mathbb{R}^+$ such that [18]

$$\mathcal{R}_T(\mathcal{A}) \leq C_0 + C_1 T. \tag{15}$$

One can see that a sublinear regret bound is favourable as this implies that the total cost by the algorithm converges with at least a sublinear rate to the base cost.

In the sequel we define two regrets for the class of memory-augmented policies in Definition 2

$$\mathcal{R}_T^{\text{lin}^*}(\pi) = J_T(\pi) - J_T(u_k^{\text{lin}^*}), \tag{16}$$

$$\mathcal{R}_T^{\pi^*}(\pi) = J_T(\pi) - J_T(\pi^*). \tag{17}$$

The first regret $\mathcal{R}_T^{\text{lin}^*}(\pi)$ in (16) quantifies the difference between the cumulative cost of the designed memory-augmented control policy $\pi = \pi(K, M, P)$ and that of the best linear control policy $u_k^{\text{lin}^*} = u_k^{\text{lin}^*}(K_{fb}^*, K_{ff}^*)$ in hindsight and is studied in [1], [3].

The second regret $\mathcal{R}_T^{\pi^*}(\pi)$ in (17) quantifies the difference between the cumulative cost of the designed memory-augmented control policy $\pi = \pi(K, M, P)$ and that of the best memory-augmented control policy denoted by $\pi^* = \pi^*(K^*, M^*, P^*)$. Since the class of memory-augmented

control policies is more general than the class of linear control policies, $\mathcal{R}_T^{\pi^*}(\pi)$ provides a stronger performance guarantee. Our results in the sequel are valid for both regrets in (16)-(17).

## IV. Boundedness guarantee from the regret and vice versa

This section contains the main results of this paper. Our aim is to give conditions to infer boundedness of the state from a linear regret and vice versa.

### A. Inferring linear regret form bounded state and control

In the next lemmas we specify the conditions such that the state, error, and input remain bounded using memory-augmented and linear feedback policies. We define the following constants to be used in the results below.

$$\begin{aligned}\kappa_z :=&\|\psi_{k,z}^{K,k}(P_{0:k-1}) - FN^{[l-z]}\|,\\ \kappa_n :=&\|K_{ff}\sum_{q=0}^{l-1} N^{[l-q]}\|,\\ D :=&\kappa^3\gamma^{-1}\kappa_w + \kappa^3\gamma^{-1}\kappa_b(1-\gamma)^{-1}(\kappa_w m_w\kappa_m + \kappa_r m_r\kappa_p)\\ &+ \kappa^3\gamma^{-1}\kappa_r\kappa_b\kappa_n + \gamma^{-1}(\kappa_w\kappa_m + \kappa_r\kappa_p) + \kappa_r\sum_{z=0}^{l-1}\kappa_z.\end{aligned} \tag{18}$$

*Lemma 3:* Consider (1)-(2) and the memory-augmented control policy in (10). Let Assumptions 1-3 hold. Assume that $K$ is selected such that $A+BK$ is $(\kappa,\gamma)$-stable and $M_k$, $P_k$ are learned such that $\|M_k^{[t]}\| \le \kappa_m(1-\gamma)^t$, $\|P_k^{[t]}\| \le \kappa_p(1-\gamma)^t$. Then,

$$\|x_k^\pi\| \le D,\ \|u_k^\pi\| \le D,\ \|e_k^\pi\| \le D. \tag{19}$$

Moreover, if $w_k \equiv 0$, $r_k \equiv 0$, $k > 0$, then the system is asymptotically stable.

*Proof:* The proof is similar to the proof of Lemma 5 in [1] and is given in Appendix V-A for completeness of the results. ∎

The conditions in Lemma 3 are necessary for the boundedness of the state and error. Indeed if the conditions do not hold, for example if $\|M_k^{[t]}\| \le \kappa_m(1-\gamma)^t$, $\|P_k^{[t]}\| \le \kappa_p(1-\gamma)^t$ is not satisfied, the boundedness cannot be guaranteed anymore.

*Lemma 4:* Consider (1)-(2) and the linear feedback policy in (13). Let Assumptions 1-3 hold. Assume that $K_{fb}$ is selected such that $A+BK_{fb}$ is $(\kappa,\gamma)$-stable. Then,

$$\|x_k^{\text{lin}}\| \le D,\ \|u_k^{\text{lin}}\| \le D,\ \|e_k^{\text{lin}}\| \le D. \tag{20}$$

*Proof:* The proof is similar to the proof of Lemma 5 in [1] and is given in Appendix V-B for completeness of the results. ∎

The results of Lemmas 3-4 will be used to show that using a memory-augmented control policy results in linear regret bounds.

*Theorem 2:* Consider (1)-(2) and the memory-augmented control policy in (10). Let Assumptions 1-4 hold. Assume that $K$ is selected such that $A+BK$ is $(\kappa,\gamma)$-stable and $M_k$, $P_k$ are learned such that $\|M_k^{[t]}\| \le \kappa_m(1-\gamma)^t$, $\|P_k^{[t]}\| \le \kappa_p m(1-\gamma)^t$. Then,

$$\begin{aligned}\mathcal{R}_T^{\text{lin}^*}(\pi) &= \mathcal{O}(T),\\ \mathcal{R}_T^{\pi^*}(\pi) &= \mathcal{O}(T).\end{aligned} \tag{21}$$

*Proof:* We use $|c_k(e_k,u_k)| \le \beta D^2$ from Assumption 4 to derive an upper bound for the total cost in (9)

$$J_T(\pi) = \sum_{k=1}^{T} c_k(e_k, u_k^\pi(K,M_k,P_k)) \le \sum_{k=1}^{T}\beta D^2 = T\beta D^2,$$

where $D$ is given in Lemma 3. This shows that the total cost of a memory-augmented policy is linear. Similarly, using Lemma 4, one can conclude that the total associated with a linear feedback policy $u_k^{\text{lin}}$ is linear

$$J_T(u_k^{\text{lin}}) = \sum_{k=1}^{T} c_k(e_k, u_k^{\text{lin}}) \le \sum_{k=1}^{T}\beta D^2 \le T\beta D^2.$$

Thus, one can conclude that the regret is linear

$$\mathcal{R}_T^{\text{lin}^*}(\pi) = J_T(\pi) - J_T(u_k^{\text{lin}*}) = 2T\beta D^2. \tag{22}$$

Linearity of $\mathcal{R}_T^{\pi^*}(\pi)$ is stablished similarly. ∎

If the conditions in Theorem 2 are satisfied, a memory-augmented control policy results, in the worst case, a linear regret bound $\mathcal{O}(T)$. This is independent of how the memory-augmented control policy is designed or learned. If the conditions in Theorem 2 are not satisfied, for example if $\|M_k^{[t]}\| \le \kappa_m(1-\gamma)^t$, $\|P_k^{[t]}\| \le \kappa_p(1-\gamma)^t$ do not hold, a linear regret cannot be guaranteed anymore. Note that it is possible to achieve sublinear regret bounds $\mathcal{O}(T^\alpha)$, $0<\alpha<1$ by designing algorithms, see [1], [3].

One can get stronger results regarding the performance of the memory-augmented policy $\pi = \pi(K,M,P)$ if the cost function $c_k$ are all equal $c_k = c$, $\forall k$. Let

$$\begin{aligned}\bar{M} &:= \frac{1}{T}\sum_{k=1}^{T} M_k,\\ \bar{P} &:= \frac{1}{T}\sum_{k=1}^{T} P_k\end{aligned} \tag{23}$$

denote the average of the learnable parameters $M_k$, $P_k$ over $T$ steps.

*Corollary 1:* Consider (1)-(2) and the memory-augmented control policy in (10). Let Assumptions 1-4 hold. Let $c_k = c$, $\forall k$. Assume that $K$ is selected such that $A+BK$ is $(\kappa,\gamma)$-stable and $M_k$, $P_k$ are learned such that $\|M_k^{[t]}\| \le \kappa_m(1-\gamma)^t$, $\|P_k^{[t]}\| \le \kappa_p m(1-\gamma)^t$. Then

$$\begin{aligned}c(e_k,u_k^\pi(K,\bar{M},\bar{P})) - c(e_k,u_k^{\text{lin}^*}) &\le \mathcal{O}(T^{\alpha-1}),\\ c(e_k,u_k^\pi(K,\bar{M},\bar{P})) - c(e_k,u_k^{\pi^*}) &\le \mathcal{O}(T^{\alpha-1}),\end{aligned} \tag{24}$$

where $0<\alpha\le 1$.

*Proof:* By Jensen's inequality [25],

$$\begin{aligned} c(e_k,&u_k^{\pi}(K,\bar{M},\bar{P})) - c(e_k,u_k^{\text{lin}^*}) \\ &\leq \frac{1}{T}\sum_{k=1}^{T}(c(e_k,u_k^{\pi}(K,M_k,P_k)) - c(e_k,u_k^{\text{lin}^*})) \\ &\leq \frac{1}{T}\mathcal{R}_T^{\text{lin}^*}(\pi). \end{aligned}$$

Based on Theorem 2, $\mathcal{R}_T^{\text{lin}^*}(\pi) \leq \mathcal{O}(T)$, so in general $c(e_k,u_k^{\pi}(K,\bar{M},\bar{P})) - c(e_k,u_k^{\text{lin}^*}) \leq \mathcal{O}(T^{\alpha-1})$, $0<\alpha\leq 1$, where $\alpha = 1$ for a linear regret $\mathcal{R}_T^{\text{lin}^*}(\pi) \leq \mathcal{O}(T)$ and $0<\alpha<1$ in a sublinar regret regime $\mathcal{R}_T^{\text{lin}^*}(\pi) \leq \mathcal{O}(T^{\alpha})$. The second inequality in (24) is concluded similarly. ■

### *B. Inferring boundedness of the state variable from linear regret*

To conclude boundedness of $\|x_k^{\pi}\|$, $\|e_k^{\pi}\|$ from a linear regret, we need to restrict the cost function further.

*Assumption 5 (cost function):* There exists $0<\underline{\kappa}_c<\infty$ such that

$$\underline{\kappa}_c(\|e_k\|^2+\|u_k\|^2) \leq c_k(e_k,u_k).$$

Assumption 5 is more stringent than Assumption 4, as it requires a positive lower bound as (5). A similar assumption is also used in proving stability from linear regret for disturbance-free nonlinear systems in [19] using a general controller and disturbance-free linear systems in [18] using a linear feedback controller.

*Theorem 3:* Consider (1)-(2) and the memory-augmented control policy in (10). Let Assumptions 1-5 hold. If the regret bound $\mathcal{R}_T^{\text{lin}^*}$ or $\mathcal{R}_T^{\pi^*}$ associated with the memory-augmented control policy in (10) is linear, then $\|x_k^{\pi}\|$, $\|e_k^{\pi}\|$ are bounded.

*Proof:* Based on Lemma 4, since Assumptions 1-3 hold, there exists a $K_{fb}$ such that the bounds in (20) hold. Then, as it has been shown in the proof of Theorem 2, the total cost of the optimal linear feedback policy $u_k^{\text{lin}^*}$ is linear

$$J_T(u_k^{\text{lin}^*}) \leq T\beta D^2. \tag{25}$$

Since the class of linear feedback policies is included in the class of memory-augmented policies, one can also conclude that the total cost of $u_k^{\pi^*}$ is also linear

$$J_T(u_k^{\pi^*}) \leq T\beta D^2. \tag{26}$$

We prove the theorem by contradiction. Assume that the regret $\mathcal{R}_T^{\text{lin}^*}$ or $\mathcal{R}_T^{\pi^*}$ associated with the memory-augmented policy $u_k^{\pi}$ is linear but $\|e_k^{\pi}\|$ is unbounded. Since the reference signal is bounded (Assumption 3), one can conclude that $\|x_k^{\pi}\| = \|e_k^{\pi}+r_k\|$ is unbounded. Similarly, $\|u_k^{\pi}\|$ is unbounded.

Since $u_k^{\pi}$ results in linear regret bounds

$$\begin{aligned} \mathcal{R}_T^{\text{lin}^*}(\pi) &= J_T(\pi) - J_T(u_k^{\text{lin}^*}) \leq C_0 + C_1 T, \\ \mathcal{R}_T^{\pi^*}(\pi) &= J_T(\pi) - J_T(\pi^*) \leq C_0 + C_1 T, \end{aligned}$$

and based on (25)-(26), the total cost associated with $u_k^{\pi}$ is also linear

$$J_T(\pi) \leq C_0 + (C_1 + \beta D^2)T. \tag{27}$$

By using the lower bound in Assumption (5), the total cost reads

$$J_T(\pi) = \sum_{k=1}^{T} c_k(e_k^{\pi},u_k^{\pi}) \geq \sum_{k=1}^{T} \underline{\kappa}_c(\|e_k^{\pi}\|^2+\|u_k^{\pi}\|^2).$$

However, since both $e_k^{\pi}$ and $u_k^{\pi}$ are unbounded, the total cost $J_T(\pi)$ is unbounded which leads to a contradiction with (27). This completes the proof. ■

## V. Conclusion and future works

In this paper, we have studied the relationship between policy regret and boundedness of the state variable for the class of memory-augmented control policies. We have considered two regret functions and shown that by properly bounding the parameters of the memory-augmented policy, it is possible to obtain linear regret bounds while sublinear regrets can be obtained by properly designing the learning procedure. We have also shown that the cost should be positive and properly bounded to conclude boundedness of the state variable from a linear regret. For our future works, we will consider studying dynamic regrets.

## APPENDIX

### *A. Proof of Lemma 3*

First, we give the bounds on $\Psi_{k,y}^{K,h}$, $\psi_{k,z}^{K,h}$ in (11)

$$\begin{aligned} \|\Psi_{k,y}^{K,h}\| &\leq \|\tilde{A}_K^y \mathbb{I}_{y\leq h-1}\| \\ &\quad + \|\sum_{j=0}^{h-1} \tilde{A}_K^j B M_{k-j-1}^{[y-j-1]} \mathbb{I}_{1\leq y-j\leq m_w}\| \\ &\leq \kappa^2(1-\gamma)^y \mathbb{I}_{y\leq h-1} \\ &\quad + \sum_{j=0}^{h-1} \kappa^2(1-\gamma)^j \kappa_b\kappa_m(1-\gamma)^{y-j-1}\mathbb{I}_{1\leq y-j\leq m_w} \\ &\leq \kappa^2(1-\gamma)^y \mathbb{I}_{y\leq h-1} + m_w\kappa^2\kappa_b\kappa_m(1-\gamma)^{y-1}, \end{aligned} \tag{28}$$

where the second inequality follows from $(\kappa,\gamma)$-stability of the controller gain $K$ and the condition that $\|M_k^{[t]}\| \leq \kappa_m(1-\gamma)^t$. Similarly

$$\|\psi_{k,z}^{K,h}\| \leq m_r\kappa^2\kappa_b\kappa_p(1-\gamma)^{z-1}. \tag{29}$$

We use $\|\Psi_{k,y}^{K,h}\|$, $\|\psi_{k,z}^{K,h}\|$ to derive the bounds.

**The bound of $\|x_k^\pi\|$:** Based on (12)

$$\begin{aligned}\|x_k^\pi\| \le& \kappa_w \sum_{y=0}^{k-1} \|\Psi_{k,y}^{K,k}(M_{0:k-1})\| + \kappa_r \sum_{z=0}^{k-1} \|\psi_{k,z}^{K,k}(P_{0:k-1})\| \\ \le& \kappa_w \sum_{y=0}^{k-1} (\kappa^2(1-\gamma)^y \mathbb{I}_{y\le k-1} + m_w\kappa^2\kappa_b\kappa_m(1-\gamma)^{y-1}) \\ &+ \kappa_r \sum_{z=0}^{k-1} m_r\kappa^2\kappa_b\kappa_p(1-\gamma)^{z-1} \qquad (30)\\ \le& \kappa_w\kappa^2\gamma^{-1}(1+m_w\kappa_b\kappa_m(1-\gamma)^{-1}) \\ &+ \kappa_r\kappa^2\gamma^{-1}m_r\kappa_b\kappa_p(1-\gamma)^{-1} \\ =& \kappa^2\gamma^{-1}(\kappa_w + \kappa_b(1-\gamma)^{-1}(\kappa_w m_w\kappa_m + \kappa_r m_r\kappa_p)) \le D\end{aligned}$$

where we have used the fact that $\sum_{n=0}^{N}(1-\gamma)^n \le \frac{1}{\gamma}$ to get the third inequality. To show asymptotic stability in the case of disturbance- and reference-free system, set $w_k \equiv 0$, $r_k \equiv 0$ for $k > 0$ in (12) (note that $w_0$ is basically the initial condition of the system, (see our explanations after (1))

$$x_k^\pi = \Psi_{k,k-1}^{K,k}(M_{0:k-1})w_0$$

resulting in

$$\|x_k^\pi\| \le \|\Psi_{k,k-1}^{K,k}(M_{0:k-1})\|\|w_0\|.$$

In (11), set $h = k$, $y = k-1$. For $k \to \infty$

$$\|\Psi_{k,k-1}^{K,k}(M_{0:k-1})\| \le \|\tilde{A}_K^{k-1}\| \le \kappa^2(1-\gamma)^{k-1}.$$

As a result, $\lim_{k\to 0}\|x_k^\pi\| = \mathbf{0}$ and the system is asymptotically stable.

**The bound of $\|u_k^\pi\|$:** Based on (10)

$$\begin{aligned}\|u_k^\pi\| =& \|Kx_k^\pi + \sum_{t=1}^{m_w} M^{[t-1]}w_{k-t} + \sum_{s=0}^{m_r-1} P^{[s]}r_{k-s}\| \\ \le& \kappa\|x_k^\pi\| + \kappa_w \sum_{t=1}^{m_w} \kappa_m(1-\gamma)^{(t-1)} \\ &+ \kappa_r \sum_{s=0}^{m_r-1} \kappa_p(1-\gamma)^s \\ \le& \kappa^3\gamma^{-1}(\kappa_w + \kappa_b(1-\gamma)^{-1}(\kappa_w m_w\kappa_m + \kappa_r m_r\kappa_p)) \\ &+ \gamma^{-1}(\kappa_w\kappa_m + \kappa_r\kappa_p) \le D.\end{aligned}$$

**The bound of $\|e_k^\pi\|$:** The tracking error is defined as

$$e_k^\pi = x_k^\pi(Y_{0:k-1}) - Fz_k$$

where $x_k^\pi(Y_{0:k-1})$ is defined in (12). Using Lemma 1 to replace $z_k$ with a linear combination of the outputs of the reference

$$\begin{aligned}e_k^\pi =& \sum_{y=0}^{k-1} \Psi_{k,y}^{K,k}(M_{0:k-1})w_{k-y-1} + \sum_{z=0}^{k-1} \psi_{k,z}^{K,k}(P_{0:k-1})r_{k-z} \\ &- F\sum_{q=0}^{l-1} N^{[l-q]}r_{k-q} \\ =& \sum_{y=0}^{k-1} \Psi_{k,y}^{K,k}(M_{0:k-1})w_{k-y-1} \\ &+ \sum_{z=0}^{l-1} (\psi_{k,z}^{K,k}(P_{0:k-1}) - FN^{[l-z]})r_{k-z} \\ &+ \sum_{z=l}^{k-1} \psi_{k,z}^{K,k}(P_{0:k-1})r_{k-z}.\end{aligned}$$

Using the bounds in (28)-(29)

$$\begin{aligned}\|e_k^\pi\| \le& \sum_{y=0}^{k-1} (\kappa^2(1-\gamma)^y\mathbb{I}_{y\le k-1} + m_w\kappa^2\kappa_b\kappa_m(1-\gamma)^{y-1})\kappa_w \\ &+ \sum_{z=0}^{l-1} \|\psi_{k,z}^{K,k}(P_{0:k-1}) - FN^{[l-z]}\|\kappa_r \\ &+ \sum_{z=l}^{k-1} m_r\kappa^2\kappa_b\kappa_p(1-\gamma)^{z-1}\kappa_r \\ \le& \kappa^2\gamma^{-1}\kappa_w(1+m_w\kappa_b\kappa_m(1-\gamma)^{-1}) \\ &+ \kappa_r\sum_{z=0}^{l-1}\kappa_z + \kappa_r\kappa^2\gamma^{-1}m_r\kappa_b\kappa_p(1-\gamma)^{l-1} \\ \le& \kappa^2\gamma^{-1}\kappa_w(1+\kappa_b(1-\gamma)^{-1}m_w\kappa_m) \\ &+ \kappa_r\sum_{z=0}^{l-1}\kappa_z + \kappa^2\gamma^{-1}\kappa_r\kappa_b(1-\gamma)^{-1}m_r\kappa_p \le D\end{aligned}$$

where we have used the fact that $\sum_{n=0}^{N}(1-\gamma)^n \le \frac{1}{\gamma}$ to get the second inequality and $(1-\gamma)^{(l-1)} \le (1-\gamma)^{-1}$ to get the third inequality.

### *B. Proof of Lemma 4*

**The bound of $\|x_k^{\text{lin}}\|$:** Using the linear feedback policy in (13), the closed-loop system of(1) reads

$$\begin{aligned}x_{k+1}^{\text{lin}} =& (A+BK_{fb})x_k^{\text{lin}} + BK_{ff}z_k + w_k \\ =& (A+BK_{fb})x_k^{\text{lin}} + BK_{ff}\sum_{q=0}^{l-1} N^{[l-q]}r_{k-q} + w_k,\end{aligned}$$

where we have used (8) in Lemma 1 to replace $z_k$ in the second line. $x_k^{\text{lin}}$ reads

$$\begin{aligned}x_k^{\text{lin}} =& \sum_{i=0}^{k-1}(A+BK_{fb})^i w_{k-i-1} \\ &+ \sum_{i=0}^{k-1}(A+BK_{fb})^i BK_{ff}\sum_{q=0}^{l-1} N^{[l-q]}r_{k-i-1-q}.\end{aligned}$$

As result,

$$\begin{aligned}\|x_k^{lin}\| \leq & \kappa_w \sum_{i=0}^{k-1} \|\tilde{A}_{K_{fb}}^i\| + \kappa_r \sum_{i=0}^{k-1} \|\tilde{A}_{K_{fb}}^i B K_{ff} \sum_{q=0}^{l-1} N^{[l-q]}\| \\ \leq & \kappa_w \sum_{i=0}^{k-1} \kappa^2 (1-\gamma)^i + \kappa_r \kappa_b \kappa_n \sum_{i=0}^{k-1} \kappa^2 (1-\gamma)^i \\ \leq & \gamma^{-1} \kappa^2 (\kappa_w + \kappa_r \kappa_b \kappa_n) \leq D \end{aligned} \tag{31}$$

where we have used the fact that $\sum_{n=0}^{N} (1-\gamma)^n \leq \frac{1}{\gamma}$ in the last inequality.

**The bound of $\|u_k^{\text{lin}}\|$:** Similarly, for the linear feedback controller in (13)

$$\begin{aligned}\|u_k^{\text{lin}}\| \leq & \|K_{fb}\| x_k^{\text{lin}} + \|K_{ff} \sum_{q=0}^{l-1} N^{[l-q]}\| \|r_{k-q}\| \\ \leq & \gamma^{-1} \kappa^3 (\kappa_w + \kappa_b \kappa_n \kappa_r) + \kappa_n \kappa_r \leq D. \end{aligned}$$

**The bound of $\|e_k^{\text{lin}}\|$:** Since $\|x_k^{\pi}\|$ and $\|x_k^{\text{lin}}\|$ have the same upper bound $D$, see (19) and (31), and based on (19) $\|e_k^{\pi}\| = \|x_k^{\pi} - F z_k\| \leq D$, once can conlude that $\|e_k^{\text{lin}}\| = \|x_k^{\text{lin}} - F z_k\| \leq D$.

## REFERENCES


[1] F. A. Yaghmaie and H. Modares, "Online optimal tracking of linear systems with adversarial disturbances," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=5nVJlKgmxp

[2] N. Niknejad, F. A. Yaghmaie, and H. Modares, "Online reference tracking for linear systems with unknown dynamics and unknown disturbances," *Transactions on Machine Learning Research*, 2023.

[3] N. Agarwal, B. Bullins, E. Hazan, S. M. Kakade, and K. Singh, "Online control with adversarial disturbances," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 154–165, 2019.

[4] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[5] O. Anava, E. Hazan, and S. Mannor, "Online learning for adversaries with memory: Price of past mistakes," *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 784–792, 2015.

[6] E. Hazan, S. Kakade, and K. Singh, "The nonstochastic control problem," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 408–421.

[7] G. Goel and B. Hassibi, "The power of linear controllers in lqr control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6652–6657.

[8] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012.

[9] F. A. Yaghmaie, F. Gustafsson, and L. Ljung, "Linear quadratic control using model-free reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 737–752, 2022.

[10] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvári, "Model-free linear quadratic control via reduction to expert prediction," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 3108–3117.

[11] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator," *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[12] A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar, "Online linear quadratic control," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1029–1038.

[13] S. Tu and B. Recht, "The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint," in *Conference on Learning Theory*. PMLR, 2019, pp. 3036–3083.

[14] J. Doyle, "Robust and optimal control," *Proceedings of 35th IEEE Conference on Decision and Control*, vol. 2, pp. 1595–1598 vol.2, 1995.

[15] H. K. Khalil, *Nonlinear Systems*, 2nd ed. Prentice Hall, 2002.

[16] A. Karapetyan, A. Iannelli, and J. Lygeros, "On the regret of $H_\infty$ control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6181–6186.

[17] N. Agarwal, E. Hazan, A. Majumdar, and K. Singh, "A regret minimization approach to iterative learning control," in *International Conference on Machine Learning*. PMLR, 2021, pp. 100–109.

[18] A. Karapetyan, A. Tsiamis, E. C. Balta, A. Iannelli, and J. Lygeros, "Implications of Regret on Stability of Linear Dynamical Systems," in *22nd IFAC World Congress 2023*, 2023.

[19] M. Nonhoff and M. A. Müller, "On the relation between dynamic regret and closed-loop stability," *Systems & Control Letters*, vol. 177, p. 105532, 2023.

[20] F. Adib Yaghmaie, S. Gunnarsson, and F. L. Lewis, "Output regulation of unknown linear systems using average cost reinforcement learning," *Automatica*, vol. 110, p. 108549, 2019. [Online]. Available: https://doi.org/10.1016/j.automatica.2019.108549

[21] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, 2014.

[22] J. Huang, *Nonlinear output regulation: theory and applications*. SIAM, 2004.

[23] N. Matni, A. Proutiere, A. Rantzer, and S. Tu, "From self-tuning regulators to reinforcement learning and back again," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3724–3740.

[24] P. Gradu, E. Hazan, and E. Minasyan, "Adaptive regret for control of time-varying dynamics," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 560–572.

[25] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta mathematica*, vol. 30, no. 1, pp. 175–193, 1906.