Online Control with Adversarial Disturbance for Continuous-time Linear Systems

Anonymous Authors[1]

Abstract

We study online control for continuous-time linear systems with finite sampling rates, where the objective is to design an online procedure that learns under non-stochastic noise and performs comparably to a fixed optimal linear controller. We present a novel two-level online algorithm, by integrating a higher-level learning strategy and a lower-level feedback control strategy. This method offers a practical and robust solution for online control, which achieves sublinear regret. Our work provides one of the first nonasymptotic results for controlling continuous-time linear systems a with finite number of interactions with the system.

1. Introduction

A major challenge in robotics is to deploy simulated controllers into real-world. This process, known as sim-to-real transfer, can be difficult due to misspecified dynamics, unanticipated real-world perturbations, and non-stationary environments. Various strategies have been proposed to address these issues, including domain randomization, meta-learning, and domain adaptation (Höfer et al., 2021; Chen et al., 2022; Hu et al., 2022).

In this work, we provide an analysis of the sim-to-real transfer problem from an online control perspective. Online control focuses on iteratively updating the controller after deployment (i.e., online) based on collected trajectories. Significant progress has been made in this field by applying insights from online learning to linear control problems (Abbasi-Yadkori & Szepesvári, 2011; Abbasi-Yadkori et al., 2014; Cohen et al., 2018; Hazan et al., 2020; Chen & Hazan, 2021; Basei et al., 2022; Andrew et al., 2013; Goel & Wierman, 2019).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML). Do not distribute.

Following this line of work, we approach the sim-to-real transfer issue for continuous-time linear systems as a non-stochastic control problem, as explored in previous works (Hazan et al., 2020; Chen & Hazan, 2021; Basei et al., 2022). These studies provide regret bounds for an online controller that lacks prior knowledge of system perturbations. However, a gap remains as no previous analysis has specifically investigated continuous-time systems, but real world systems often evolve continuously in time.

Existing literature on online continuous control is limited (Vrabie et al., 2009; Jiang & Jiang, 2012; Duncan et al., 1992; Rizvi & Lin, 2018). Most continuous control research emphasizes the development of model-free algorithms, such as policy iteration, under the assumption of noise absence. Recently, (Basei et al., 2022) examined online continuous-time linear quadratic control in the presence of standard Brownian noise that may not always hold true in real-world applications. This leads us to the crucial question:

Is it possible to design an online non-stochastic control algorithm in a continuous-time setting that achieves sublinear regret?

Our work addresses this question by proposing a two-level online controller. The higher-level controller symbolizes the policy learning process and updates the policy at a low frequency to minimize regret. Conversely, the lower-level controller delivers high-frequency feedback control input to reduce discretization error.

Our proposed algorithm results in regret bounds for continuous-time linear control in the face of non-stochastic disturbances. More importantly, our analyses suggest that online learning algorithms, with potentially nontrivial adaptations, could also benefit continuous-time control problems. We believe this direction holds promising potential for further exploration.

2. Related Works

The control theory of linear dynamical systems under the disturbance has been thoroughly examined in various contexts, such as the linear quadratic stochastic control (Athans, 1971), robust control (Stengel, 1994; Khalil et al., 1996),

system identification (Goodwin et al., 1981; Kumar, 1983; Campi & Kumar, 1998; Ljung, 1998). However, most of these problems are investigated in non-robust settings, with robust control being the sole exception where adversarial perturbations in the dynamic are permitted. In this scenario, the controller solves for the optimal linear controller in the presence of worst-case noise. Nonetheless, the algorithms designed in this context can be overly conservative as they optimize over the worst-case noise, a scenario that is rare in real-world applications. We will elaborate on the difference between robust control and online non-stochastic control in Section 3.

Online control There has been a recent surge of interest in online control, as demonstrated by studies such as (Abbasi-Yadkori & Szepesvári, 2011; Abbasi-Yadkori et al., 2014; Cohen et al., 2018). In online control, the player interacts with the environment and updates the policy in each round aiming to achieve sublinear regret. In scenarios with stochastic Gaussian noise, (Cohen et al., 2018) has provided the first efficient algorithm with an $O(\sqrt{T})$ regret bound. However, in real-world applications, the assumption of Gaussian distribution is often unfulfilled.

(Agarwal et al., 2019) pioneered research on non-stochastic online control, where the noises can be adversarial. Under general convex cost, they introduced the Disturbance-Action Policy Class. Using an online convex optimization (OCO) algorithm with memory, they achieved an $O(\sqrt{T})$ regret bound. Subsequent studies extended this approach to other scenarios, such as quadratic cost (Basei et al., 2022), partial observations (Simchowitz et al., 2020; Simchowitz, 2020) or unknown dynamical systems (Hazan et al., 2020; Chen & Hazan, 2021), yielding varying theoretical guarantees like online competitive ratio (Goel et al., 2022; Shi et al., 2020).

Online Continuous Control Compared to online control, there has been relatively little research on model-based continuous-time control. Most continuous control literature has focused on developing model-free algorithms such as policy iteration (e.g. (Vrabie et al., 2009; Jiang & Jiang, 2012; Rizvi & Lin, 2018)), typically assume zero-noise. This is because analyzing the system when transition dynamics are represented by differential equations, rather than recurrence formulas, poses a significant challenge.

Recently, (Basei et al., 2022) studied online continuous-time linear quadratic control with standard Brownian noise and unknown system dynamics. They proposed an algorithm based on the least-square method, which estimates the system's coefficients and solves the corresponding Riccati equation. However, it should be noted that standard Brownian noise can be quite stringent and may fail in real-world applications.

3. Problem Setting

In this paper, we consider the online non-stochastic control for continuous-time linear systems. Therefore, we provide a brief overview below and define our notations.

3.1. Continuous-time Linear Systems

The Linear Dynamical System can be considered a specific case of a continuous Markov decision process with linear transition dynamics. The state transitions are governed by the following equation:

$$\dot{x}_t = Ax_t + Bu_t + w_t\,,$$

where x_t is the state at time t, u_t is the action taken by the controller at time t, and w_t represents the disturbance at time t. We assume access to $\dot{x}_t$ at each time step, which is dependent on the state, action, and disturbance at time t. We do not make any strong assumptions about the distribution of w_t, and we also assume that the distribution of w_t is unknown to the learner beforehand. This implies that the disturbance sequence w_t can be selected adversarially.

When the action u_t is applied to the state x_t, a cost $c_t(x_t, u_t)$ is incurred. Here, we assume that the cost function c_t is convex. However, this cost is not known in advance and is only revealed after the action u_t is implemented at time t. In the system described above, an online policy π is defined as a function that maps known states to actions, i.e., $u_t = \pi(\{x_\xi | \xi \in [0, t]\})$. Our goal, then, is to design an algorithm that determines such an online policy to minimize the cumulative cost incurred. Specifically, for any algorithm $\mathcal{A}$, the cost incurred over a time horizon T is:

$$J_T(\mathcal{A}) = \int_0^T c_t(x_t, u_t)dt\,.$$

In scenarios where the policy is linear (i.e., a linear controller) $\pi(K)$, such that $u_t = -Kx_t$, we use $J(K)$ to denote the cost of a policy from a certain class $K \in \mathcal{K}$.

3.2. Difference between Robust and Online Non-stochastic Control

While both robust and online non-stochastic control models incorporate adversarial noise, it's crucial to understand that their objectives differ significantly.

The objective function for robust control, as seen in (Stengel, 1994; Khalil et al., 1996), is defined as:

$$\min_{u_1} \max_{w_{1:T}} \min_{u_2} \ldots \min_{u_t} \max_{w_T} J_T(\mathcal{A})\,,$$

Meanwhile, the objective function for online non-stochastic control, as discussed in (Agarwal et al., 2019), is:

$$\min_{\mathcal{A}} \max_{w_{1:T}} (J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K))\,.$$

Note that the robust control approach seeks to directly minimize the cost function, while online non-stochastic control targets the minimization of regret, which is the discrepancy between the actual cost and the cost associated with a baseline policy. Additionally, in robust control, the noise at each step can depend on the preceding policy, whereas in online non-stochastic control, all the noise is predetermined (though unknown to the player).

3.3. Assumptions

Throughout this paper, we operate under the following assumptions, starting with the initial condition $x_0 = 0$. We denote $\|\cdot\|$ as the L_2 norm of the vector and matrix. Firstly, we make assumptions concerning the system dynamics and noise:

Assumption 3.1. The matrices that govern the dynamics are bounded, meaning $\|A\| \leq \kappa_A$ and $\|B\| \leq \kappa_B$, where κ_A and κ_B are constants. Moreover, the perturbation and its derivative are both continuous and bounded: $\|w_t\|, \|\dot{w}_t\| \leq W$, with W being a constant.

These assumptions ensure that we can bound the states and actions, as well as their first and second-order derivatives. Next, we make assumptions regarding the cost function:

Assumption 3.2. The costs $c_t(x, u)$ are convex. Additionally, if there exists a constant D such that $\|x\|, \|u\| \leq D$, then $|c_t(x, u)| \leq \beta D^2, \|\nabla_x c_t(x, u)\|, \|\nabla_u c_t(x, u)\| \leq GD$.

This assumption implies that if the differences between states and actions are small, then the error in their cost will also be relatively small. Finally, we describe our baseline policy class:

Definition 3.3. A linear policy K is (κ, γ)-strongly stable if, for any $h > 0$ that is sufficiently small, there exist matrices L_h, P such that $I + h(A - BK) = PL_hP^{-1}$, with the following two conditions:

1. The norm of L_h is strictly smaller than unity and dependent on h, i.e., $\|L_h\| \leq 1 - h\gamma$.
2. The controller and transforming matrices are bounded, i.e., $\|K\| \leq \kappa$ and $\|P\|, \left\|P^{-1}\right\| \leq \kappa$.

This definition ensures the system can be stabilized by a linear controller K.

3.4. Regret Formulation

To evaluate the designed algorithm, we use regret, which is defined as the cumulative difference between the cost incurred by the policy of our algorithm and the cost incurred by the best policy in hindsight. Let $\mathcal{K}$ denotes the class of strongly stable linear policy, i.e. $\mathcal{K} = \{K : K \text{ is } (\kappa, \gamma)\text{-strongly stable}\}$. Then, for an algorithm $\mathcal{A}$, the regret is defined as follows.

$$\text{Regret}(\mathcal{A}) = J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K)\,.$$

4. Algorithm Design

In this section, we outline the design of our algorithm and discuss the technical challenges encountered in deriving our main theorem.

- First, we discretize the total time period T into smaller intervals of length h. We use the information at each point $x_h, x_{2h}, \ldots$ and $u_h, u_{2h}, \ldots$ to approximate the actual cost of each time interval, leveraging the continuity assumption. This process does introduce some discretization errors.
- Next, we employ the Disturbance-Action policy (DAC) (Agarwal et al., 2019). This policy selects the action based on the current time step and the estimations of disturbances from several past steps. This policy can approximate the optimal linear policy in hindsight when we choose suitable parameters. However, the optimal policy K^* is unknown, so we cannot directly acquire the optimal choice. To overcome this, we employ the OCO with memory framework (Anava et al., 2015) to iteratively adjust the DAC policy parameter M_t to approximate the optimal solution M^*.
- After that, we introduce the concept of the ideal state y_t and ideal action v_t that approximate the actual state x_t and action u_t. Note that both the state and policy depend on all DAC policy parameters $M_1, M_2, \ldots, M_t$. Yet, the OCO with memory framework only considers the previous H steps. Therefore, we need to consider ideal state and action. y_t and v_t represent the state the system would reached if it had followed the DAC policy $\{M_{t-H}, \ldots, M_t\}$ at all time steps from $t - H$ to t, under the assumption that the state x_{t-H} was 0.

From all the analysis above, we can decompose the regret as four parts: the discretization error, the approximation error of the DAC policy compared to the optimal policy, the regret of the OCO with memory, and the approximation error between the ideal cost and the actual cost.

4.1. New Challenges in Online Continuous Control

In transitioning online control to continuous systems, we cannot directly apply the methods from (Agarwal et al., 2019) to our work. We must overcome several challenges:

Challenge 1. Unbounded States In a discrete-time system, it is straightforward to demonstrate that the state sequence x_t is bounded using the DAC policy. This can be easily shown by applying the dynamics inequality $\|x_{t+1}\| \leq a\|x_t\|+b$ (where $a < 1$) and the induction method presented in (Agarwal et al., 2019). However, for a continuous-time system, a different approach is necessary because we only have the differential equation instead of the state recurrence formula.

One naive approach is to use the Taylor expansion of each state to derive the recurrence formula of the state. However, this argument requires the prerequisite knowledge that the states within this neighborhood are bounded by the dynamics, leading to circular reasoning.

To overcome this challenge, we employ Gronwall's inequality to bound the first and second-order derivatives in the neighborhood of the current state. We then use these bounded properties, in conjunction with an estimation of previous noise, to bound the distance to the next state. Through an iterative application of this method, we can argue that all states and actions are bounded.

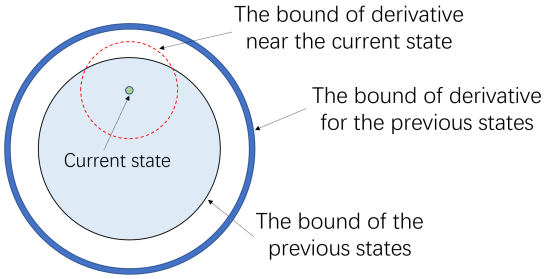

Figure 1. Bounding the states and their derivatives separately. We prove the bounded derivative of the current state based on the information of the previous state and then use the information of derivative to bound the next state. We inductively show that this is true for all states.

Challenge 2. The Ideal Cost Approximation and Discretization Error Trade-off We propose a new definition of a strongly stable policy for continuous-time systems, as the definition used for discrete systems is not immediately applicable. In Definition 3.3, we describe a strongly stable policy K that is dependent on a discretization parameter h. It is crucial to understand that the selection of h influences the convergence rate of the ideal cost approximation error, as measured by $|c_t(x_t, u_t) - c_t(y_t, v_t)|$, resulting in a rate of $O(T(1 - h\gamma)^H)$. On the one hand, choosing an overly small value for h may lead to a slow convergence rate. On the other hand, an excessively large h may cause the discretization error to become unmanageable, being of the order $O(hT)$. Hence, the design of an appropriate discretization parameter presents a significant challenge.

Challenge 3. The Curse of Dimensionality Caused by Discretization In discrete-time systems, where the number of states is predetermined, the parameters for the DAC policy and the OCO memory buffer can be selected with relative ease (Agarwal et al., 2019). However, in continuous-time systems, the number of states can be inversely proportional to the discretization parameter h, which also determines the size of the OCO memory buffer. Thus, if we set the OCO memory buffer size as $H = O(\log(T)/h)$ to attain a sublinear ideal cost approximation error $|c_t(x_t, u_t) - c_t(y_t, v_t)|$, the associated regret of OCO with memory will be $O(\sqrt{T}/h^{2.5})$. This regret could become excessively large if h is small enough to allow for minimal discretization error.

4.2. Main Algorithm

In the subsequent discussion, we use shorthand notation to denote the cost, state, control, and disturbance variables c_{ih}, x_{ih}, u_{ih}, and w_{ih} as c_i, x_i, u_i, and w_i, respectively.

We now introduce our algorithm, which is built upon a two-level controller update approach. Our algorithm employs two controllers, working in concert to enhance the performance of the policy. The higher-level controller implements the OCO with memory to sporadically update the policy, while the lower-level controller offers high-frequency control input to minimize discretization error.

Our lower-level controller utilizes the DAC policy. To formally define the DAC policy for continuous systems, we start by dividing the interval into multiple sub-intervals. This division ensures that the current state is influenced only by all disturbance that occur within each time interval, rather than exclusively at the moment of disturbance. Following this property, we introduce our definition of the DAC policy in the continuous system:

Definition 4.1. The Disturbance-Action Policy Class(DAC) is defined as:

$$u_t = -Kx_t + \sum_{i=1}^{l} M_t^i \hat{w}_{t-i}\,,$$

where K is a fixed strongly stable policy, l is a parameter that signifies the dimension of the policy class, $M_t = \{M_t^1, \ldots, M_t^l\}$ is the weighting parameter of the disturbance at step t, and $\hat{w}_t$ is the estimated disturbance:

$$\hat{w}_t = \frac{x_{t+1} - x_t - h(Ax_t + Bu_t)}{h}\,. \tag{1}$$

We note that this estimation of disturbance is readily implementable as it only requires information from the previous state. Furthermore, it counteracts the second-order residue term of the Taylor expansion of x_t, which greatly simplifies the analysis of state evolution.

Our higher-level controller adopts the OCO with memory framework. A technical challenge lies in balancing the approximation error and OCO regret. To achieve a low approximation error, we desire the policy update interval H to be inversely proportional to the sampling distance h. However, this relationship lead to large OCO regret. To mitigate this issue, we introduce a new parameter $m = \Theta(\frac{1}{h})$, representing the lookahead window. We update the parameter M_t only once every m iteration, further reducing the OCO regret without negatively impacting the approximation error:

$$M_{t+1} = \begin{cases} \Pi_{\mathcal{M}}\left(M_t - \eta \nabla g_t(M)\right) & \text{if } t \bmod m == 0\,, \\ M_t & \text{otherwise}\,. \end{cases}$$

For notational convenience and to avoid redundancy, we denote $\tilde{M}_{[t/m]} = M_t$. We can then define the ideal state and action. Due to the properties of the OCO with memory structure, we need to consider only the previous Hm states and actions, rather than all states. As a result, we introduce the definition of the ideal state and action. During the interval $t \in [im, (i+1)m - 1]$, the learning policy remains unchanged, so we could define the ideal state and action in the following:

Definition 4.2. The ideal state y_t and action v_t at time $t \in [im, (i+1)m - 1]$ are defined as

$$y_t = x_t(\tilde{M}_{i-H}, ..., \tilde{M}_i), v_t = -Ky_t + \sum_{j=1}^{l} M_i^j w_{t-i}\,.$$

where the notation indicates that we assume the state x_{t-H} is 0 and that we apply the DAC policy $\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right)$ at all time steps from $t - Hm$ to t.

We can also define the ideal cost in this interval:

Definition 4.3. The ideal cost function during the interval $t \in [im, (i+1)m - 1]$ is defined as follows:

$$f_i\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right) = \sum_{t=im}^{(i+1)m-1} c_t\left(y_t\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right), v_t\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right)\right)\,.$$

With all the concepts presented above, we are now prepared to introduce our algorithm:

Algorithm 1 Continuous two-level online control algorithm

Input: step size η, sample distance h, policy update parameters H, m, parameters κ, γ, T.
Define sample numbers $n = \lceil T/h \rceil$, OCO policy update times $p = \lceil n/m \rceil$.
Define DAC policy update class $\mathcal{M} = \left\{\tilde{M} = \left\{\tilde{M}^{[1]} \ldots \tilde{M}^{[Hm]}\right\} : \left\|\tilde{M}^{[i]}\right\| \le 2\kappa^3(1-\gamma)^{i-1}\right\}$.
Initialize $M_0 \in \mathcal{M}$ arbitrarily.
for $k = 0, \ldots, p-1$ **do**
  for $s = 0, \ldots, m-1$ **do**
    Denote the discretization time $t = km + s$.
    Use the action $u_t = -Kx_t + h\sum_{i=1}^{Hm} \tilde{M}_k^i \hat{w}_{t-i}$ during the time period $[th, (t+1)h]$.
    Observe the new state x_{t+1} at time $(t+1)h$ and record $\hat{w}_t$ according to Equation (1).
  end for
  Update OCO policy $\tilde{M}_{k+1} = \Pi_{\mathcal{M}}\left(\tilde{M}_k - \eta \nabla g_t(\tilde{M}_k)\right)$.
end for

5. Main Result

In this section, we present the primary theorem of online continuous control regret analysis:

Theorem 5.1. *Under Assumption 3.1, 3.2, a step size of $\eta = \Theta(\sqrt{\frac{m}{Th}})$, and a DAC policy update frequency $m = \Theta(\frac{1}{h})$, Algorithm 1 attains a regret bound of*

$$J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K) \le O(n(1-h\gamma)^{\frac{H}{h}}) + O(\sqrt{nh}) + O(Th)\,.$$

With the sampling distance $h = \Theta(\frac{1}{\sqrt{T}})$, and the OCO policy update parameter $H = \Theta(\log(T))$, Algorithm 1 achieves a regret bound of

$$J_T(\mathcal{A}) - \min_{K \in \mathcal{K}} J_T(K) \le O\left(\sqrt{T}\log(T)\right)\,.$$

Theorem 5.1 demonstrates a regret that matches the regret of a discrete system (Agarwal et al., 2019). Despite the analysis of a continuous system differing from that of a discrete system, we can balance discretization error, approximation error, and OCO with memory regret by selecting an appropriate update frequency for the policy. Here, $O(\cdot)$ and $\Theta(\cdot)$ are abbreviations for the polynomial factors of universal constants in the assumption.

While we defer the detailed proof to the appendix, we outline the key ideas and highlight them below.

Proof Sketch We denote x_t^*, $u_t^* = K^* x_t^*$ as the optimal state and action following the policy specified by K^* respectively, where $K^* = \arg\max_{K \in \mathcal{K}} J_T(K)$. We use

the shorthand c_{ih}, x_{ih}, u_{ih}, and w_{ih} for c_i, x_i, u_i, and w_i, respectively.

Initially, we need to prove Lemma 6.1, confirm that the state x_t and action u_t are bounded by some constant D when using either the DAC policy or the optimal policy.

We then discretize and decompose the regret as follows:

$$\begin{aligned}
&J_T(\mathcal{A}) - \min_{K\in\mathcal{K}} J_T(K) \\
&= \int_0^T c_t(x_t,u_t)dt - \int_0^T c_t(x_t^*,u_t^*)dt \\
&= \sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t,u_t)dt - \sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t^*,u_t^*)dt \\
&= h\left(\sum_{i=0}^{n-1} c_i(x_i,u_i) - \sum_{i=0}^{n-1} c_i(x_i^*,u_i^*)\right) + R_0\,,
\end{aligned}$$

where R_0 represents the discretization error:

$$\begin{aligned}
R_0 = &\sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} (c_t(x_t,u_t) - c_t(x_t^*,u_t^*))dt \\
&-h\sum_{i=0}^{n-1}\left(c_i(x_i,u_i) - c_i(x_i^*,u_i^*)\right)\,.
\end{aligned}$$

By the discussion in section 4, the first term can be further decomposed as

$$\sum_{i=0}^{n-1} c_i(x_i,u_i) - \sum_{i=0}^{n-1} c_i(x_i^*,u_i^*) = R_1 + R_2 + R_3\,,$$

where

$$\begin{aligned}
R_1 &= \sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - f_i(\tilde{M}_{i-H},\ldots,\tilde{M}_i)\right)\,, \\
R_2 &= \sum_{i=0}^{p-1} f_i\left(\tilde{M}_{i-H},\ldots,\tilde{M}_i\right) - \min_{M\in\mathcal{M}}\sum_{i=0}^{p-1} f_i\left(M,\ldots,M\right)\,, \\
R_3 &= \min_{M\in\mathcal{M}}\sum_{i=0}^{p-1} f_i\left(M,\ldots,M\right) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*)\,.
\end{aligned}$$

Thus, we have the regret decomposition as $\mathrm{Regret}(T) = h(R_1+R_2+R_3)+R_0$. We then separately upper bound each of the four terms.

The term R_0 represents the error caused by discretization, which decreases as the number of sampling points increases and the sampling distance h decreases. This is because more sampling points make our approximation of the continuous system more accurate. Using Lemma 6.2, we get the following upper bound: $R_0 \le O(hT)$.

The term R_1 represents the difference between the actual cost and the approximate cost. For a fixed h, this error decreases as the number of sample points looked ahead m increases, while it increases as the sampling distance h decreases. This is because the closer adjacent points are, the slower the convergence after approximation. By Lemma 6.3 we can bound it as $R_1 \le O(n(1-h\gamma)^{Hm})$.

The term R_2 is incurred due to the regret of the OCO with memory algorithm. Note that this term is determined by learning rate η and the policy update frequency m. Choosing suitable parameters and using Lemma 6.4, we can obtain the following upper bound: $R_2 \le O(\sqrt{n/h})$.

The term R_3 represents the difference between the ideal optimal cost and the actual optimal cost. Since the accuracy of the DAC policy approximation of the optimal policy depends on its degree of freedom l, a higher degree of freedom leads to a more accurate approximation of the optimal policy. We use Lemma 6.5 and choose $l = Hm$ to bound this error: $R_3 \le O(n(1-h\gamma)^{Hm})$.

By summing up these four terms and taking $m = \Theta(\frac{1}{h})$, we get:

$$\mathrm{Regret}(T) \le O(nh(1-h\gamma)^{\frac{H}{h}}) + O(\sqrt{nh}) + O(hT)\,.$$

Finally, we choose $h = \Theta\left(\frac{1}{\sqrt{T}}\right)$, $m = \Theta\left(\frac{1}{h}\right)$, $H = \Theta(log(T))$, the regret is bounded by

$$\mathrm{Regret}(T) \le O(\sqrt{T}\log(T))\,.$$

6. Key Lemmas

In this section, we will primarily discuss the rationale behind the proof of our key lemmas. Due to space limitations, detailed proofs of these lemmas are provided in the appendix.

Bounding the States and Actions First, we need to prove all the states and actions are bounded.

Lemma 6.1. *Under Assumption 3.1 and 3.2, choosing arbitrary h in the interval $[0,h_0]$ where h_0 is a constant only depends on the parameters in the assumption, we have for any t and policy M_i, $\|x_t\|, \|y_t\|, \|u_t\|, \|v_t\| \le D$. $\|x_t - y_t\|, \|u_t - v_t\| \le \kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}D$. In particular, taking all the $M_t = 0$ and $K = K^*$, we obtain the actual optimal solution $\|x_t^*\|, \|u_t^*\| \le D$.*

The proof of this Lemma mainly use the Gronwall inequality and the induction method.

Then we analyze the discretization error of the system.

Bounding the Discretization Error Analyzing a continuous system can be arduous; hence, we employ discretization with distance h to facilitate the analysis.

Lemma 6.2. *Under Assumption 3.2, Algorithm 1 attains the following bound of R_0:*

$$R_0 = \sum_{i=0}^{n-1} \int_{ih}^{(i+1)h} (c_t(x_t, u_t) - c_t(x_t^*, u_t^*))dt - h\sum_{i=0}^{n-1} (c_i(x_i, u_i) - c_i(x_i^*, u_i^*)) \leq 2GDhT\,.$$

This lemma indicates that the discretization error is directly proportional to the sample distance h. In other words, increasing the number of sampling points leads to more accurate estimation of system.

Based on Lemma 6.1, we know that $\|x_t\|, \|u_t\| \leq D$. By utilizing assumption 3.2, we can deduce that:

$$\begin{aligned}&|c_t(x_t, u_t) - c_{ih}(x_{ih}, u_{ih})| \\ &\leq \max_{x,u}(\|\nabla_x c_t(x,u)\| + \|\nabla_u c_t(x,u)\|)(t - ih) \\ &\leq GD(t - ih)\,.\end{aligned}$$

Summing up all these terms, we obtain the bound for the discretization error.

Bounding the Difference between Ideal Cost and Actual Cost The following lemma describes the upper bound of the error by approximating the ideal state and action:

Lemma 6.3. *Under Assumption 3.1 and 3.2, Algorithm 1 attains the following bound of R_1:*

$$\begin{aligned}R_1 &= \sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i, u_i) - f_i\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right)\right) \\ &\leq nGD^2\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}\,.\end{aligned}$$

From this lemma, it is evident that for a fixed sample distance h, the error diminishes as the number of sample points looked ahead m increases. However, as the sampling distance h decreases, the convergence rate of this term becomes slower. Therefore, it is not possible to select an arbitrarily small value for h in order to minimize the discretization error R_0.

We need to demonstrate that the discrepancy between x_t and y_t, as well as u_t and v_t, is sufficiently small, given assumption 3.1. This can be proven by analyzing the state evolution under the DAC policy.

By utilizing Assumption 3.2 and Lemma 6.1, we can deduce the following inequality:

$$\begin{aligned}&|c_t(x_t, u_t) - c_t(y_t, v_t)| \\ &\leq |c_t(x_t, u_t) - c_t(y_t, u_t)| + |c_t(y_t, u_t) - c_t(y_t, v_t)| \\ &\leq GD\|x_t - y_t\| + GD\|u_t - v_t\|\,.\end{aligned}$$

Summing over all the terms and use Lemma 6.1, we can derive an upper bound for R_1.

Next, we analyze the regret of Online Convex Optimization (OCO) with a memory term.

Bounding the Regret of OCO with Memory To analyze OCO with a memory term, we provide an overview of the framework established by (Anava et al., 2015) in online convex optimization. The framework considers a scenario where, at each time step t, an online player selects a point x_t from a set $\mathcal{K} \subset \mathbb{R}^d$. At each time step, a loss function $f_t : \mathcal{K}^{H+1} \to \mathbb{R}$ is revealed, and the player incurs a loss of $f_t(x_{t-H}, \ldots, x_t)$. The objective is to minimize the policy regret, which is defined as

$$\text{PolicyRegret} = \sum_{t=H}^{T} f_t(x_{t-H}, \ldots, x_t) - \min_{x\in\mathcal{K}} \sum_{t=H}^{T} f_t(x, \ldots, x)\,.$$

In this setup, the first term corresponds to the DAC policy we choose, while the second term is used to approximate the optimal strongly stable linear policy.

Lemma 6.4. *Under Assumption 3.1 and 3.2, choosing $m = \frac{C}{h}$ and $\eta = \Theta(\frac{m}{Th})$, Algorithm 1 attains the following bound of R_2:*

$$\begin{aligned}R_2 &= \sum_{i=0}^{p-1} f_i(\tilde{M}_{i-H}, \ldots, \tilde{M}_i) - \min_{M\in\mathcal{M}} \sum_{i=0}^{p-1} f_i(M, \ldots, M) \\ &\leq \frac{4a}{\gamma}\sqrt{\left(\frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma} + C^2\kappa^3\kappa_B W_0 H^2\right)\frac{n}{h}}\,.\end{aligned}$$

To analyze this term, we can transform the problem into an online convex optimization with memory and utilize existing results presented by (Anava et al., 2015) for it. By applying their results, we can derive the following bound:

$$\begin{aligned}&\sum_{t=H}^{T} f_t(x_{t-H}, \ldots, x_t) - \min_{x\in\mathcal{K}} \sum_{t=H}^{T} f_t(x, \ldots, x) \\ &\leq O\left(D\sqrt{G_f\left(G_f + LH^2\right)T}\right)\,.\end{aligned}$$

Taking into account the bounds on the diameter, Lipschitz constant, and the gradient, we can ultimately derive an upper bound for R_2.

Bounding the Approximation Error of DAC Policy Lastly, we aim to establish a bound on the approximation error between the optimal DAC policy and the unknown optimal linear policy.

Lemma 6.5. *Under Assumption 3.1 and 3.2, Algorithm 1*

attains the following bound of R_3:

$$R_3 = \min_{M\in\mathcal{M}} \sum_{i=0}^{p-1} f_i(M,...,M) - \sum_{i=0}^{p-1} \sum_{j=im}^{(i+1)m-1} c_i(x_i^*, u_i^*)$$
$$\leq 3n(1-h\gamma)^{Hm} GDW_0\kappa^3 a(lh\kappa_B + 1)\,.$$

The intuition behind this lemma is that the evolution of states leads to an approximation of the optimal linear policy in hindsight, where $u_t^* = -K^* x_t$ if we choose $M^* = \{M^i\}$, where $M^i = (K-K^*)(I+h(A-BK^*))^i$. Although the optimal policy K^* is unknown, such an upper bound is attainable because the left-hand side represents the minimum of $M \in \mathcal{M}$.

7. Conclusions and Future Directions

In this paper, we propose a two-level online controller to achieve sublinear regret in online continuous-time control of linear system with adversarial disturbances. The higher-level controller updates the policy using the Online Convex Optimization (OCO) with memory framework at a low frequency to reduce regret, while the lower-level controller employs the DAC policy to approximate the actual state with an idealized setting. Through our analysis, we observe that the regret primarily depends on the time T and the sampling distance h. By selecting suitable sampling distance, we are able to achieve sublinear regret of T.

There are several potential directions for future research in online non-stochastic control of continuous-time systems. Firstly, this paper focuses on solving the problem when the dynamics are known. It would be valuable to extend this work to address the case of unknown dynamics, where a trade-off between system identification and regret minimization exists. Secondly, while we assume convexity of the cost function in this paper, it would be interesting to explore whether assuming strong convexity can lead to even smaller regret. Finally, it would be intriguing to shift the focus from regret to the competitive ratio in this setup, as it presents a different perspective on performance evaluation.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

Abbasi-Yadkori, Y., Bartlett, P., and Kanade, V. Tracking adversarial targets. In *International Conference on Machine Learning*, 2014.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. Online control with adversarial disturbances. In *International Conference on Machine Learning*, 2019.

Anava, O., Hazan, E., and Mannor, S. Online learning for adversaries with memory: price of past mistakes. *Advances in Neural Information Processing Systems*, 2015.

Andrew, L., Barman, S., Ligett, K., Lin, M., Meyerson, A., Roytman, A., and Wierman, A. A tale of two metrics: Simultaneous bounds on competitiveness and regret. In *Conference on Learning Theory*, pp. 741–763. PMLR, 2013.

Athans, M. The role and use of the stochastic linear-quadratic-gaussian problem in control system design. *IEEE transactions on automatic control*, 16(6):529–552, 1971.

Basei, M., Guo, X., Hu, A., and Zhang, Y. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 2022.

Campi, M. C. and Kumar, P. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Chen, X. and Hazan, E. Black-box control for linear dynamical systems. In *Conference on Learning Theory*, 2021.

Chen, X., Hu, J., Jin, C., Li, L., and Wang, L. Understanding domain randomization for sim-to-real transfer. In *International Conference on Learning Representations*, 2022.

Cohen, A., Hasidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. Online linear quadratic control. In *International Conference on Machine Learning*, 2018.

Duncan, T. E., Mandl, P., and Pasik-Duncan, B. On least squares estimation in continuous time linear stochastic systems. *Kybernetika*, 28(3):169–180, 1992.

Goel, G. and Wierman, A. An online algorithm for smoothed regression and lqr control. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2504–2513. PMLR, 2019.

Goel, G., Agarwal, N., Singh, K., and Hazan, E. Best of both worlds in online control: Competitive ratio and policy regret. *arXiv preprint arXiv:2211.11219*, 2022.

Goodwin, G. C., Ramadge, P. J., and Caines, P. E. Discrete time stochastic adaptive control. *SIAM Journal on Control and Optimization*, 19(6):829–853, 1981.

Hazan, E. Introduction to online convex optimization. *CoRR*, abs/1909.05207, 2019.

Hazan, E., Kakade, S., and Singh, K. The nonstochastic control problem. In *Algorithmic Learning Theory*, 2020.

Höfer, S., Bekris, K., Handa, A., Gamboa, J. C., Mozifian, M., Golemo, F., Atkeson, C., Fox, D., Goldberg, K., Leonard, J., et al. Sim2real in robotics and automation: Applications and challenges. *IEEE transactions on automation science and engineering*, 2021.

Hu, J., Zhong, H., Jin, C., and Wang, L. Provable sim-to-real transfer in continuous domain with partial observations. *arXiv preprint arXiv:2210.15598*, 2022.

Jiang, Y. and Jiang, Z.-P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 2012.

Khalil, I., Doyle, J., and Glover, K. *Robust and optimal control*. Prentice hall, 1996.

Kumar, P. Optimal adaptive control of linear-quadratic-gaussian systems. *SIAM Journal on Control and Optimization*, 21(2):163–178, 1983.

Ljung, L. *System identification*. Springer, 1998.

Rizvi, S. A. A. and Lin, Z. Output feedback reinforcement learning control for the continuous-time linear quadratic regulator problem. In *2018 Annual American Control Conference (ACC)*, 2018.

Shi, G., Lin, Y., Chung, S.-J., Yue, Y., and Wierman, A. Online optimization with memory and competitive control. *Advances in Neural Information Processing Systems*, 33: 20636–20647, 2020.

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

Simchowitz, M., Singh, K., and Hazan, E. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pp. 3320–3436. PMLR, 2020.

Stengel, R. F. *Optimal control and estimation*. Courier Corporation, 1994.

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., and Lewis, F. L. Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, 2009.

In the appendix we define n as the smallest integer greater than or equal to $\frac{T}{h}$, and we use the shorthand c_{ih}, x_{ih}, u_{ih}, and w_{ih} as c_i, x_i, u_i, and w_i, respectively.

A. Proof of Theorem 5.1

Theorem 5.1. *Under Assumption 3.1, 3.2, a step size of $\eta = \Theta(\sqrt{\frac{m}{Th}})$, and a DAC policy update frequency $m = \Theta(\frac{1}{h})$, Algorithm 1 attains a regret bound of*

$$J_T(\mathcal{A}) - \min_{K\in\mathcal{K}} J_T(K) \leq O(n(1-h\gamma)^{\frac{H}{h}}) + O(\sqrt{nh}) + O(Th)\,.$$

With the sampling distance $h = \Theta(\frac{1}{\sqrt{T}})$, and the OCO policy update parameter $H = \Theta(\log(T))$, Algorithm 1 achieves a regret bound of

$$J_T(\mathcal{A}) - \min_{K\in\mathcal{K}} J_T(K) \leq O\left(\sqrt{T}\log(T)\right)\,.$$

Proof. We denote $u_t^* = K^* x_t^*$ as the optimal state and action that follows the policy specified by K^*, where $K^* = \arg\max_{K\in\mathcal{K}} J_T(K)$.

We then discretize and decompose the regret as follows:

$$\begin{aligned}
J_T(\mathcal{A}) - \min_{K\in\mathcal{K}} J_T(K) &= \int_0^T c_t(x_t,u_t)dt - \int_0^T c_t(x_t^*,u_t^*)dt \\
&= \sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t,u_t)dt - \sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t^*,u_t^*)dt \\
&= h\left(\sum_{i=0}^{n-1} c_i(x_i,u_i) - \sum_{i=0}^{n-1} c_i(x_i^*,u_i^*)\right) + R_0\,,
\end{aligned}$$

where R_0 represents the discretization error.

We define p as the smallest integer greater than or equal to $\frac{n}{m}$, then the first term can be further decomposed as

$$\begin{aligned}
&\sum_{i=0}^{n-1} c_i(x_i,u_i) - \sum_{i=0}^{n-1} c_i(x_i^*,u_i^*) \\
=&\sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*) \\
=&\sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - \sum_{j=im}^{(i+1)m-1} c_i(y_i,v_i)\right) + \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(y_i,v_i) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*) \\
=&\sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - f_i(\tilde{M}_{i-H},\ldots,\tilde{M}_i)\right) + \sum_{i=0}^{p-1} f_i(\tilde{M}_{i-H},\ldots,\tilde{M}_i) \\
&- \min_{M\in\mathcal{M}}\sum_{i=0}^{p-1} f_i(M,\ldots,M) + \min_{M\in\mathcal{M}}\sum_{i=0}^{p-1} f_i(M,\ldots,M) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*)\,,
\end{aligned}$$

where the last equality is by the definition of the idealized cost function (Definition 4.3).

Let us denote

$$R_1 = \sum_{i=0}^{p-1} \left(\sum_{j=im}^{(i+1)m-1} c_i(x_i, u_i) - f_i(\tilde{M}_{i-H}, \ldots, \tilde{M}_i) \right) ,$$
$$R_2 = \sum_{i=0}^{p-1} f_i(\tilde{M}_{i-H}, \ldots, \tilde{M}_i) - \min_{M \in \mathcal{M}} \sum_{i=0}^{p-1} f_i(M, \ldots, M) ,$$
$$R_3 = \min_{M \in \mathcal{M}} \sum_{i=0}^{p-1} f_i(M, \ldots, M) - \sum_{i=0}^{p-1} \sum_{j=im}^{(i+1)m-1} c_i(x_i^*, u_i^*) .$$

Then we have the regret decomposition as

$$\text{Regret}(T) = h(R_1 + R_2 + R_3) + O(hT) .$$

We then separately upper bound each of the four terms.

The term R_0 represents the error caused by discretization, which decreases as the number of sampling points increases and the sampling distance h decreases. This is because more sampling points make our approximation of the continuous system more accurate. Using Lemma 6.2, we get the following upper bound: $R_0 \leq O(hT)$.

The term R_1 represents the difference between the actual cost and the approximate cost. For a fixed h, this error decreases as the number of sample points looked ahead m increases, while it increases as the sampling distance h decreases. This is because the closer adjacent points are, the slower the convergence after approximation. By Lemma 6.3 we can bound it as $R_1 \leq O(n(1-h\gamma)^{Hm})$.

The term R_2 is incurred due to the regret of the OCO with memory algorithm. Note that this term is determined by learning rate η and the policy update frequency m. Choosing suitable parameters and using Lemma 6.4, we can obtain the following upper bound: $R_2 \leq O(\sqrt{n/h})$.

The term R_3 represents the difference between the ideal optimal cost and the actual optimal cost. Since the accuracy of the DAC policy approximation of the optimal policy depends on its degree of freedom l, a higher degree of freedom leads to a more accurate approximation of the optimal policy. We use Lemma 6.5 and choose $l = Hm$ to bound this error: $R_3 \leq O(n(1-h\gamma)^{Hm})$.

By summing up these four terms and taking $m = \Theta(\frac{1}{h})$, we get:

$$\text{Regret}(T) \leq O(nh(1-h\gamma)^{\frac{H}{h}}) + O(\sqrt{nh}) + O(hT) .$$

Finally, we choose $h = \Theta\left(\frac{1}{\sqrt{T}}\right)$, $m = \Theta\left(\frac{1}{h}\right)$, $H = \Theta(log(T))$, the regret is bounded by

$$\text{Regret}(T) \leq O(\sqrt{T}\log(T)) .$$

□

B. The evolution of the state

In this section we will prove that using the DAC policy, the states and actions are uniformly bounded. The difference between ideal and actual states and the difference between ideal and actual action is very small.

We begin with expressions of the state evolution using DAC policy:

Lemma B.1. *We have the evolution of the state and action:*

$$x_{t+1} = Q_h^{l+1} x_{t-l} + h \sum_{i=0}^{2l} \Psi_{t,i} \hat{w}_{t-i} \,,$$
$$y_{t+1} = h \sum_{i=0}^{2Hm} \Psi_{t,i} \hat{w}_{t-i} \,,$$
$$v_t = -Ky_t + h \sum_{j=1}^{Hm} M_t^j \hat{w}_{t-j} \,.$$

where $\Psi_{t,i}$ represent the coefficients of $\hat{w}_{t-i}$:

$$\Psi_{t,i} = Q_h^i \mathbf{1}_{i \le l} + h \sum_{j=0}^{l} Q_h^j B M_{t-j}^{i-j} \mathbf{1}_{i-j \in [1,l]} \,.$$

Proof. Define $Q_h = I + h(A - BK)$. Using the Taylor expansion of x_t and denoting r_t as the second-order residue term, we have

$$x_{t+1} = x_t + h\dot{x}_t + h^2 r_t = x_t + h(Ax_t + Bu_t + w_t) + h^2 r_t \,.$$

Then we calculate the difference between w_i and $\hat{w}_i$:

$$\hat{w}_t - w_t = \frac{x_{t+1} - x_t - h(Ax_t + Bu_t + w_t)}{h} = h r_t \,.$$

Using the definition of DAC policy and the difference between disturbance, we have

$$\begin{aligned}
x_{t+1} &= x_t + h\left(Ax_t + B\left(-Kx_t + h\sum_{i=1}^{l} M_t^i \hat{w}_{t-i}\right) + \hat{w}_t - h r_t\right) + h^2 r_t \\
&= (I + h(A - BK))x_t + h\left(Bh\sum_{i=1}^{l} M_t^i \hat{w}_{t-i} + \hat{w}_t\right) \\
&= Q_h x_t + h\left(Bh\sum_{i=1}^{l} M_t^i \hat{w}_{t-i} + \hat{w}_t\right) \\
&= Q_h^2 x_{t-1} + h\left(Q_h\left(Bh\sum_{i=1}^{l} M_{t-1}^i \hat{w}_{t-1-i} + \hat{w}_{t-1}\right)\right) + h\left(Bh\sum_{i=1}^{l} M_t^i \hat{w}_{t-i} + \hat{w}_t\right) \\
&= Q_h^{l+1} x_{t-l} + h\sum_{i=0}^{2l} \Psi_{t,i} \hat{w}_{t-i} \,,
\end{aligned}$$

where the last equality is by recursion and $\Psi_{t,i}$ represent the coefficients of $\hat{w}_{t-i}$.

Then we calculate the coefficients of w_{t-i} and get the following result:

$$\Psi_{t,i} = Q_h^i \mathbf{1}_{i \le l} + h \sum_{j=0}^{l} Q_h^j B M_{t-j}^{i-j} \mathbf{1}_{i-j \in [1,l]} \,.$$

By the ideal definition of y_{t+1} and v_t(only consider the effect of the past Hm steps while planning, assume $x_{t-Hm} = 0$),

taking $l = Hm$ we have

$$y_{t+1} = h \sum_{i=0}^{2Hm} \Psi_{t,i} \hat{w}_{t-i},$$
$$v_t = -Ky_t + h \sum_{j=1}^{Hm} M_t^j \hat{w}_{t-j}\,. \qquad \square$$

Then we prove the norm of the transition matrix is bounded.

Lemma B.2. *We have the following bound of the transition matrix:*

$$\|\Psi_{t,i}\| \le a(lh\kappa_B + 1)\kappa^2(1-h\gamma)^{i-1}\,.$$

Proof. By the definition of strongly stable policy, we know

$$\|Q_h^i\| = \|(PL_hP^{-1})^i\| = \|P(L_h)^iP^{-1}\| \le \|P\|\|L_h\|^i\|P^{-1}\| \le a\kappa^2(1-h\gamma)^i\,. \tag{2}$$

By the definition of $\Psi_{t,i}$, we have

$$\begin{aligned}\|\Psi_{t,i}\| &= \left\| Q_h^i \mathbf{1}_{i\le l} + h\sum_{j=0}^{l} Q_h^j B M_{t-j}^{i-j} \mathbf{1}_{i-j\in[1,l]} \right\| \\ &\le \kappa^2(1-h\gamma)^i + ah\sum_{j=1}^{l} \kappa_B\kappa^2(1-h\gamma)^j(1-h\gamma)^{i-j-1} \\ &\le \kappa^2(1-h\gamma)^i + alh\kappa_B\kappa^2(1-h\gamma)^{i-1} \le a(lh\kappa_B+1)\kappa^2(1-h\gamma)^{i-1}\,,\end{aligned}$$

where the first inequality is due to equation 2, assumption 3.1 and the condition of $\left\|M_t^i\right\| \le a(1-h\gamma)^{i-1}$. $\square$

After that, we can uniformly bound the state x_t and its first and second-order derivative.

Lemma B.3. *For any $t \in [0,T]$, choosing arbitrary h in the interval $[0,h_0]$ where h_0 is a constant only depends on the parameters in the assumption, we have $\|x_t\| \le D_1$, $\|\dot{x}_t\| \le D_2$, $\|\ddot{x}_t\| \le D_3$ and the estimatation of disturbance is bounded by $\|\hat{w}_t\| \le W_0$. Moreover, D_1, D_2, D_3 are only depend on the parameters in the assumption.*

Proof. We prove this lemma by induction. When $t = 0$, it is clear that x_0 satisfies this condition. Suppose $x_t \le D_1$, $\dot{x}_t \le D_2$, $\ddot{x}_t \le D_3$, $\hat{w}_t \le W_0$ for any $t \le t_0$, where $t_0 = kh$ is the k-th discretization point. Then for $t \in [t_0, t_0+h]$, we first prove that $\dot{x}_t \le D_2$, $\ddot{x}_t \le D_3$.

By Assumption 3.1 and our definition of u_t, we know that for any $t \in [t_0, t_0+h]$. Thus, we have

$$\begin{aligned}\|\dot{x}_t\| &= \|Ax_t + Bu_t + w_t\| \\ &= \|Ax_t + B(-Kx_{t_0} + h\sum_{i=1}^{l} M_k^i \hat{w}_{k-i}) + w_t\| \\ &\le \kappa_A\|x_t\| + \kappa_B\kappa\|x_{t_0}\| + h\sum_{i=1}^{l}(1-h\gamma)^{i-1}W_0 + W \\ &\le \kappa_A\|x_t\| + \kappa_B\kappa D_1 + \frac{W_0}{\gamma} + W\,,\end{aligned}$$

where the first inequality is by the induction hypothesis $\hat{w}_t \le W_0$ for any $t \le t_0$ and $M_k^i \le (1-h\gamma)^{i-1}$, the second inequality is by the induction hypothesis $x_t \le D_1$ for any $t \le t_0$.

For any $t \in [t_0, t_0 + h]$, because we choose the fixed policy $u_t \equiv u_{t_0}$, so we have $\dot{u}_t = 0$ and

$$\|\ddot{x}_t\| = \|A\dot{x}_t + B\dot{u}_t + w_t\| = \|A\dot{x}_t + w_t\| \leq \kappa_A \|\dot{x}_t\| + W \,.$$

By the Newton-Leibniz formula, we have for any $\zeta \in [0, h]$,

$$\dot{x}_{t_0+\zeta} - \dot{x}_{t_0} = \int_0^\zeta \ddot{x}_{t_0+\xi} d_\xi \,.$$

Then we have

$$\begin{aligned}
\|\dot{x}_{t_0+\zeta}\| &\leq \|\dot{x}_{t_0}\| + \int_0^\zeta \|\ddot{x}_{t_0+\xi}\| d_\xi \\
&\leq \|\dot{x}_{t_0}\| + \int_0^\zeta (\kappa_A \|\dot{x}_{t_0+\xi}\| + W) d_\xi \\
&= \|\dot{x}_{t_0}\| + W\zeta + \kappa_A \int_0^\zeta \|\dot{x}_{t_0+\xi}\| d_\xi \,.
\end{aligned}$$

By Gronwall inequality, we have

$$\|\dot{x}_{t_0+\zeta}\| \leq \|\dot{x}_{t_0}\| + W\zeta + \int_0^\zeta (\|\dot{x}_{t_0}\| + W\xi) \exp(\kappa_A(\zeta - \xi)) d_\xi \,.$$

Then we have

$$\begin{aligned}
\|\dot{x}_{t_0+\zeta}\| &\leq \|\dot{x}_{t_0}\| + W\zeta + \int_0^\zeta (\|\dot{x}_{t_0}\| + W\zeta) \exp(\kappa_A \zeta)) d_\xi \\
&= (\|\dot{x}_{t_0}\| + W\zeta)(1 + \zeta \exp(\kappa_A \zeta)) \\
&\leq \left(\kappa_A \|x_{t_0}\| + \kappa_B \kappa D_1 + \frac{W_0}{\gamma} + W + Wh \right) (1 + h \exp(\kappa_A h)) \\
&\leq \left((\kappa_A + \kappa_B \kappa) D_1 + \frac{W_0}{\gamma} + W + Wh \right) (1 + h \exp(\kappa_A h)) \\
&\leq \left((\kappa_A + \kappa_B \kappa) D_1 + \frac{W_0}{\gamma} + 2W \right) (1 + \exp(\kappa_A)) \,,
\end{aligned}$$

where the first inequality is by the relation $\xi \leq \zeta$, the second inequality is by the relation $\zeta \leq h$ and the bounding property of first-order derivative, the third inequality is by the induction hypothesis and the last inequality is due to $h \leq 1$.

By the relation $\|\ddot{x}_t\| \leq \kappa_A \|\dot{x}_t\| + W$, we have

$$\|\ddot{x}_{t_0+\zeta}\| \leq \kappa_A D_2 + W \,.$$

So we choose $D_3 = \kappa_A D_2 + W$. By the equation 1, we have

$$\begin{aligned}
\|\hat{w}_t - w_t\| &= \left\| \frac{x_{t+1} - x_t - h(Ax_t + Bu_t + w_t)}{h} \right\| \\
&= \left\| \frac{x_{t+1} - x_t - h\dot{x}_t}{h} \right\| = \left\| \frac{\int_0^h (\dot{x}_{t+\xi} - \dot{x}_t) d\xi}{h} \right\| = \left\| \frac{\int_0^h \int_0^\xi \ddot{x}_{t+\zeta} d\zeta d\xi}{h} \right\| \\
&\leq \frac{\int_0^h \int_0^\xi \|\ddot{x}_{t+\zeta}\| d\zeta d\xi}{h} \\
&\leq h D_3 \,,
\end{aligned}$$

where in the second line we use the Newton-Leibniz formula, the inequality is by the conclusion $\|\ddot{x}_t\| \leq D_3$ which we have proved before. By Assumption 3.1, we have

$$\|\hat{w}_t\| \leq W + hD_3 \,.$$

Choosing $D_3 = \kappa_A D_2 + W$, $W_0 = W + hD_3 = W + h(\kappa_A D_2 + W)$, we get

$$\begin{aligned}
\|\dot{x}_{t_0+\zeta}\| &\leq ((\kappa_A + \kappa_B\kappa)D_1 + \frac{W_0}{\gamma} + 2W)(1 + \exp(\kappa_A)) \\
&\leq ((\kappa_A + \kappa_B\kappa)D_1 + \frac{W + h(\kappa_A D_2 + W)}{\gamma} + 2W)(1 + \exp(\kappa_A)) \\
&\leq D_2 \left(\frac{h\kappa_A}{\gamma}(1 + \exp(\kappa_A))\right) + \left((\kappa_A + \kappa_B\kappa)D_1 + \frac{(1 + h + 2\gamma)W}{\gamma}\right)(1 + \exp(\kappa_A))) \,.
\end{aligned}$$

Using the notation

$$\begin{aligned}
\beta_1 &= \frac{h\kappa_A}{\gamma}(1 + \exp(\kappa_A)) \,, \\
\beta_2 &= \left((\kappa_A + \kappa_B\kappa)D_1 + \frac{2(1+\gamma)W}{\gamma}\right)(1 + \exp(\kappa_A)) \,.
\end{aligned}$$

When $h < \frac{\gamma}{2\kappa_A(1+\exp(\kappa_A))}$, we have $\beta_1 < \frac{1}{2}$. Taking $D_2 = 2\beta_2$ we get

$$\|\dot{x}_{t_0+\zeta}\| \leq \beta_1 D_2 + \beta_2 \leq D_2 \,.$$

So we have proved that for any $t \in [t_0, t_0 + h]$, $\|\dot{x}_t\| \leq D_2$, $\|\ddot{x}_t\| \leq D_3$, $\|\hat{w}_t\| \leq W_0$.

Then we choose suitable D_1 and prove that for any $t \in [t_0, t_0 + h]$, $\|x_t\| \leq D_1$.

Using Lemma B.1, we have

$$x_{t+1} = h\sum_{i=0}^{t} \Psi_{t,i}\hat{w}_{t-i} \,.$$

By the induction hypothesis of bounded state and estimation noise in $[0, t_0]$ together with Lemma B.2, we have

$$\begin{aligned}
\|x_{t+1}\| &\leq h\sum_{i=0}^{t}(lh\kappa_B + 1)\kappa^2(1 - h\gamma)^i(W + hD_3) \\
&\leq \frac{(lh\kappa_B + 1)\kappa^2(W + hD_3)}{\gamma} \,.
\end{aligned}$$

Then, by the Taylor expansion and the inequality $\dot{x}_t \leq D_2$, we have for any $\zeta \in [0, h]$,

$$\|x_{t+1} - x_{t+\zeta}\| = \|\int_{\zeta}^{h} \dot{x}_{t+\xi} d\xi\| \leq (h - \zeta)D_2 \leq hD_2 \,.$$

Therefore we have

$$\begin{aligned}
\|x_{t+\zeta}\| &\leq \|x_{t+1}\| + hD_2 \leq \frac{(lh\kappa_B + 1)\kappa^2(W + hD_3)}{\gamma} + hD_2 \\
&= \frac{(lh\kappa_B + 1)\kappa^2 W(1 + h)}{\gamma} + hD_2\left(\frac{(lh\kappa_B + 1)\kappa^2\kappa_A}{\gamma} + 1\right) \\
&\leq \frac{(l\kappa_B + 1)2\kappa^2 W}{\gamma} + hD_2\left(\frac{(l\kappa_B + 1)\kappa^2\kappa_A}{\gamma} + 1\right) \,.
\end{aligned}$$

In the last inequality we use $h \leq 1$.

By the relation $D_2 = \beta_2/(1-\beta_1)$ and $\beta_1 \leq \frac{1}{2}$, we know that

$$D_2 \leq 2\left((\kappa_A + \kappa_B\kappa)D_1 + \frac{2(1+\gamma)W}{\gamma}\right)(1+\exp(\kappa_A)).$$

Using the notation

$$\gamma_1 = 2h(\kappa_A + \kappa_B\kappa)(1+\exp(\kappa_A))\,,$$
$$\gamma_2 = \frac{(l\kappa_B+1)2\kappa^2 W}{\gamma} + 4\frac{(1+\gamma)W}{\gamma}(1+\exp(\kappa_A))\left(\frac{(l\kappa_B+1)\kappa^2\kappa_A}{\gamma}+1\right).$$

We have $\|x_{t+\zeta}\| \leq \gamma_1 D_1 + \gamma_2$.

From the equation of γ_1 we know that when $h \leq \frac{1}{4(\kappa_A+\kappa_B\kappa)(1+\exp(\kappa_A))}$ we have $\gamma_1 \leq \frac{1}{2}$. Then we choose $D_1 = 2\gamma_2$, we finally get

$$\|x_{t+\zeta}\| \leq \gamma_1 D_1 + \gamma_2 \leq D_1\,.$$

Finally, set

$$h_0 = \min\left\{1, \frac{\gamma}{\kappa_A(1+\exp(\kappa_A))}, \frac{1}{4(\kappa_A+\kappa_B\kappa)(1+\exp(\kappa_A))}\right\},$$

By the relationship $D_1 = 2\gamma_2$, $D_2 = 2\beta_2$, $D_3 = \kappa_A D_2 + W$, $W_0 = W + hD_3$,

we can verify the induction hypothesis. Moreover, we know that D_1, D_2, D_3 are not depend on h. Therefore we have proved the claim.

□

The last step is then to bound the action and the approximation errors of states and actions.

Lemma 6.1. *Under Assumption 3.1 and 3.2, choosing arbitrary h in the interval $[0, h_0]$ where h_0 is a constant only depends on the parameters in the assumption, we have for any t and policy M_i, $\|x_t\|, \|y_t\|, \|u_t\|, \|v_t\| \leq D$. $\|x_t - y_t\|, \|u_t - v_t\| \leq \kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}D$. In particular, taking all the $M_t = 0$ and $K = K^*$, we obtain the actual optimal solution $\|x_t^*\|, \|u_t^*\| \leq D$.*

Proof. By Lemma B.2, we have

$$\|\Psi_{t,i}\| \leq a(lh\kappa_B + 1)\kappa^2(1-h\gamma)^{i-1}\,.$$

By Lemma B.3 we know that for any h in $[0, h_0]$, where

$$h_0 = \min\left\{1, \frac{\gamma}{\kappa_A(1+\exp(\kappa_A))}, \frac{1}{4(\kappa_A+\kappa_B\kappa)(1+\exp(\kappa_A))}\right\},$$

we have $\|x_t\| \leq D_1$.

By Lemma B.1, Lemma B.2 and Lemma B.3, we have

$$\begin{aligned}\|y_{t+1}\| &= \|h\sum_{i=0}^{2Hm}\Psi_{t,i}\hat{w}_{t-i}\| \\ &\leq hW_0\sum_{i=0}^{2Hm} a(lh\kappa_B+1)\kappa^2(1-h\gamma)^{i-1} \\ &\leq \frac{aW_0(lh\kappa_B+1)\kappa^2}{\gamma} = \tilde{D}_1\,.\end{aligned}$$

Via the definition of x_t, y_t, we have

$$\|x_t - y_t\| \le \kappa^2(1-h\gamma)^{Hm+1}\|x_{t-Hm}\| \le \kappa^2(1-h\gamma)^{Hm+1}D_1\,.$$

For the actions

$$u_t = -Kx_t + h\sum_{i=1}^{Hm} M_t^i \hat{w}_{t-i}\,,$$
$$v_t = -Ky_t + h\sum_{i=1}^{Hm} M_t^i \hat{w}_{t-i}\,,$$

we can derive the bound

$$\|u_t\| \le \|Kx_t\| + h\sum_{i=1}^{Hm}\left\|M_t^i\hat{w}_{t-i}\right\| \le \kappa\|x_t\| + W_0 h\sum_{i=1}^{Hm} a(1-h\gamma)^{i-1} \le \kappa D_1 + \frac{aW_0}{\gamma}\,,$$
$$\|v_t\| \le \|Ky_t\| + h\sum_{i=1}^{Hm}\left\|M_t^i\hat{w}_{t-i}\right\| \le \kappa\|y_t\| + W_0 h\sum_{i=1}^{Hm} a(1-h\gamma)^{i-1} \le \kappa\tilde{D}_1 + \frac{aW_0}{\gamma}\,,$$
$$\|u_t - v_t\| \le \|K\|\|x_t - y_t\| \le \kappa^3(1-h\gamma)^{Hm+1}D_1\,.$$

Taking $D = \max\{D_1, \tilde{D}_1, \kappa D_1 + \frac{W_0}{\gamma}, \kappa\tilde{D}_1 + \frac{W_0}{\gamma}\}$, we get $\|x_t\|, \|y_t\|, \|u_t\|, \|v_t\| \le D$.

We also have

$$\|x_t - y_t\| + \|u_t - v_t\| \le \kappa^2(1-h\gamma)^{Hm+1}D_1 + \kappa^3(1-h\gamma)^{Hm+1}D_1 \le \kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}D\,.$$

In particular, the optimal policy can be recognized as taking the DAC policy with all the M_t equal to 0 and the fixed strongly stable policy $K = K^*$. So we also have $\|x_t^*\|, \|u_t^*\| \le D$.

□

Now we have finished the analysis of evolution of the states. It will be helpful to prove the key lemmas in this paper.

C. Proof of Lemma 6.2

In this section we will prove the following lemma:

Lemma 6.2. *Under Assumption 3.2, Algorithm 1 attains the following bound of R_0:*

$$R_0 = \sum_{i=0}^{n-1}\int_{ih}^{(i+1)h}(c_t(x_t,u_t) - c_t(x_t^*,u_t^*))dt$$
$$- h\sum_{i=0}^{n-1}(c_i(x_i,u_i) - c_i(x_i^*,u_i^*)) \le 2GDhT\,.$$

Proof. By Assumption 3.2 and Lemma B.3, we have

$$|c_t(x_t,u_t) - c_{ih}(x_{ih},u_{ih})| \le \max_{x,u}(\|\nabla_x c_t(x,u)\| + \|\nabla_u c_t(x,u)\|)(t-ih) \le GD(t-ih)\,.$$

Therefore we have

$$
\begin{aligned}
&|\sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t,u_t)dt - h\sum_{i=0}^{n-1} c_i(x_i,u_i)| \\
=&|\sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} (c_t(x_t,u_t) - c_{ih}(x_{ih},u_{ih}))dt| \\
\le& GD\sum_{i=0}^{n-1}\int_{ih}^{(i+1)h}(t-ih)dt \le GDnh^2 = GDhT\,.
\end{aligned}
$$

A similar bound can easily be established by lemma B.3 about the optimal state and policy:

$$
|\sum_{i=0}^{n-1}\int_{ih}^{(i+1)h} c_t(x_t^*,u_t^*)dt - \sum_{i=0}^{n-1} c_i(x_i^*,u_i^*)| \le GDhT\,.
$$

Taking sum of the two terms we get $R_0 \le 2GDhT$.

□

D. Proof of Lemma 6.3

In this section we will prove the following lemma:

Lemma 6.3. *Under Assumption 3.1 and 3.2, Algorithm 1 attains the following bound of R_1:*

$$
\begin{aligned}
R_1 &= \sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - f_i\left(\tilde{M}_{i-H},\dots,\tilde{M}_i\right)\right) \\
&\le nGD^2\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}\,.
\end{aligned}
$$

Proof. Using Lemma 6.1 and Assumption 3.2, have the approximation error between ideal cost and actual cost bounded as,

$$
\begin{aligned}
|c_t(x_t,u_t) - c_t(y_t,v_t)| &\le |c_t(x_t,u_t) - c_t(y_t,u_t)| + |c_t(y_t,u_t) - c_t(y_t,v_t)| \\
&\le GD\|x_t - y_t\| + GD\|u_t - v_t\| \\
&\le GD^2\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}\,,
\end{aligned}
$$

where the first inequality is by triangle inequality, the second inequality is by Assumption 3.2, Lemma 6.1, and the third inequality is by Lemma 6.1.

With this, we have

$$
\begin{aligned}
R_1 &= \sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - f_i(\tilde{M}_{i-H},...,\tilde{M}_i)\right) \\
&= \sum_{i=0}^{p-1}\left(\sum_{j=im}^{(i+1)m-1} c_i(x_i,u_i) - \sum_{j=im}^{(i+1)m-1} c_i(y_i,v_i)\right) \\
&\le \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} GD^2\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1} \le nGD^2\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}\,.
\end{aligned}
$$

□

E. Proof of Lemma 6.4

Before we start the proof of Lemma 6.4, we first present an overview of the online convex optimization (OCO) with memory framework. Consider the setting where, for every t, an online player chooses some point $x_t \in \mathcal{K} \subset \mathbb{R}^d$, a loss function $f_t : \mathcal{K}^{H+1} \mapsto \mathbb{R}$ is revealed, and the learner suffers a loss of $f_t(x_{t-H}, \ldots, x_t)$. We assume a certain coordinate-wise Lipschitz regularity on f_t of the form such that, for any $j \in \{1, \ldots, H\}$, for any $x_1, \ldots, x_H, \tilde{x}_j \in \mathcal{K}$

$$|f_t(x_1, \ldots, x_j, \ldots, x_H) - f_t(x_1, \ldots, \tilde{x}_j, \ldots, x_H)| \leq L \|x_j - \tilde{x}_j\| .$$

In addition, we define $\tilde{f}_t(x) = f_t(x, \ldots, x)$, and we let

$$G_f = \sup_{t \in \{1,\ldots,T\}, x \in \mathcal{K}} \left\| \nabla \tilde{f}_t(x) \right\|, \quad D_f = \sup_{x,y \in \mathcal{K}} \|x - y\| .$$

The resulting goal is to minimize the policy regret, which is defined as

$$\text{Regret} = \sum_{t=H}^{T} f_t(x_{t-H}, \ldots, x_t) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} f_t(x, \ldots, x) .$$

Algorithm 2 Online Gradient Descent with Memory (OGD-M)

Input: Step size η, functions $\{f_t\}_{t=m}^{T}$.
Initialize $x_0, \ldots, x_{H-1} \in \mathcal{K}$ arbitrarily.
for $t = H, \ldots, T$ **do**
  Play x_t, suffer loss $f_t(x_{t-H}, \ldots, x_t)$.
  Set $x_{t+1} = \Pi_{\mathcal{K}} \left(x_t - \eta \nabla \tilde{f}_t(x) \right)$.
end for

To minimize this regret, a commonly used algorithm is the Online Gradient descent. By running the Algorithm 2, we may bound the policy regret by the following lemma:

Lemma E.1. *Let $\{f_t\}_{t=1}^{T}$ be Lipschitz continuous loss functions with memory such that $\tilde{f}_t$ are convex. Then by runnning algorithm 2 itgenerates a sequence $\{x_t\}_{t=1}^{T}$ such that*

$$\sum_{t=H}^{T} f_t(x_{t-H}, \ldots, x_t) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} f_t(x, \ldots, x) \leq \frac{D_f^2}{\eta} + T G_f^2 \eta + L H^2 \eta G_f T .$$

Furthermore, setting $\eta = \frac{D_f}{\sqrt{G_f (G_f + LH^2) T}}$ implies that

$$\text{PolicyRegret} \leq 2 D_f \sqrt{G_f (G_f + LH^2) T} .$$

Proof. By the standard OGD analysis (Hazan, 2019), we know that

$$\sum_{t=H}^{T} \tilde{f}_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=H}^{T} \tilde{f}_t(x) \leq \frac{D_f^2}{\eta} + T G^2 \eta.$$

In addition, we know by the Lipschitz property, for any $t \geq H$, we have

$$\begin{aligned} |f_t(x_{t-H}, \ldots, x_t) - f_t(x_t, \ldots, x_t)| &\leq L \sum_{j=1}^{H} \|x_t - x_{t-j}\| \leq L \sum_{j=1}^{H} \sum_{l=1}^{j} \|x_{t-l+1} - x_{t-l}\| \\ &\leq L \sum_{j=1}^{H} \sum_{l=1}^{j} \eta \left\| \nabla \tilde{f}_{t-l}(x_{t-l}) \right\| \leq L H^2 \eta G, \end{aligned}$$

and so we have that

$$\left|\sum_{t=H}^{T} f_t\left(x_{t-H}, \ldots, x_t\right) - \sum_{t=H}^{T} f_t\left(x_t, \ldots, x_t\right)\right| \leq TLH^2\eta G.$$

It follows that

$$\sum_{t=H}^{T} f_t\left(x_{t-H}, \ldots, x_t\right) - \min_{x\in\mathcal{K}} \sum_{t=H}^{T} f_t(x, \ldots, x) \leq \frac{D_f^2}{\eta} + TG_f^2\eta + LH^2\eta G_f T\,.$$

□

In this setup, the first term corresponds to the DAC policy we make, and the second term is used to approximate the optimal strongly stable linear policy. It is worth noting that the cost of OCO with memory depends on the update frequency H. Therefore, we propose a two-level online controller. The higher-level controller updates the policy with accumulated feedback at a low frequency to reduce the regret, whereas a lower-level controller provides high-frequency updates of the DAC policy to reduce the discretization error. In the following part, we define the update distance of the DAC policy as $l = Hm$, where m is the ratio of frequency between the DAC policy update and OCO memory policy update. Formally, we update the value of M_t once every m transitions, where g_t represents a loss function.

$$M_{t+1} = \begin{cases} \Pi_{\mathcal{M}}\left(M_t - \eta\nabla g_t(M)\right) & \text{if } t\%m == 0 \\ M_t & \text{otherwise}\,. \end{cases}$$

From now on, we denote $\tilde{M}_t = M_{tm}$ for the convenience to remove the duplicate elements. By the definition of ideal cost, we know that it is a well-defined definition.

By Lemma B.1 we know that

$$y_{t+1} = h\sum_{i=0}^{2Hm} \Psi_{t,i}\hat{w}_{t-i},$$

$$v_t = -Ky_t + h\sum_{j=1}^{Hm} M_t^j \hat{w}_{t-j}\,,$$

where

$$\Psi_{t,i} = Q_h^i \mathbf{1}_{i\leq l} + h\sum_{j=0}^{l} Q_h^j B M_{t-j}^{i-j}\mathbf{1}_{i-j\in[1,l]}\,.$$

So we know that y_t and y_t are linear combination of M_t, therefore

$$f_i\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right) = \sum_{t=im}^{(i+1)m-1} c_t\left(y_t\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right), v_t\left(\tilde{M}_{i-H}, \ldots, \tilde{M}_i\right)\right).$$

is convex in M_t. So we can use the OCO with memory structure to solve this problem.

By Lemma B.3 we know that y_t and v_t are bounded by D. Then we need to calculate the diameter, Lipchitz constant, and gradient bound of this function f_i. In the following, we choose the DAC policy parameter $l = Hm$.

Lemma E.2. *(Bounding the diameter) We have*

$$D_f = \sup_{M_i, M_j\in\mathcal{M}} \|M_i - M_j\| \leq \frac{2a}{h\gamma}$$

.

Proof. By the definition of $\mathcal{M}$, taking $l = Hm$ we know that

$$\begin{aligned}\sup_{M_i, M_j \in \mathcal{M}} \|M_i - M_j\| &\leq \sum_{k=1}^{Hm} \|M_i^k - M_j^k\| \\ &\leq \sum_{k=1}^{Hm} 2a(1-h\gamma)^{k-1} \\ &\leq \frac{2a}{h\gamma}.\end{aligned}$$

□

Lemma E.3. *(Bounding the Lipschitz Constant) Consider two policy sequences* $\left\{\tilde{M}_{i-H} \dots \tilde{M}_{i-k} \dots \tilde{M}_i\right\}$ *and* $\left\{\tilde{M}_{i-H} \dots \hat{M}_{i-k} \dots \tilde{M}_i\right\}$ *which differ in exactly one policy played at a time step* $t-k$ *for* $k \in \{0, \dots, H\}$*. Then we have that*

$$\left| f_i\left(\tilde{M}_{i-H} \dots \tilde{M}_{i-k} \dots \tilde{M}_i\right) - f_i\left(\tilde{M}_{i-H} \dots \hat{M}_{i-k} \dots \tilde{M}_i\right)\right| \leq C^2\kappa^3\kappa_B W_0 \sum_{j=0}^{Hm} \|\tilde{M}_{i-k}^j - \hat{M}_{i-k}^j\|,$$

where C *is a constant.*

Proof. By the definition we have

$$\begin{aligned}\|y_t - \tilde{y}_t\| &= \|h \sum_{i=0}^{2Hm} h \sum_{j=0}^{Hm} Q_h^j B(M_{t-j}^{i-j} - \tilde{M}_{t-j}^{i-j}) \mathbf{1}_{i-j\in[1,Hm]} \hat{w}_{t-i}\| \\ &\leq h^2\kappa^2\kappa_B W_0 \sum_{i=0}^{2Hm}\sum_{j=0}^{Hm} \|M_{t-j}^{i-j} - \tilde{M}_{t-j}^{i-j}\| \mathbf{1}_{i-j\in[1,Hm]} \\ &\leq h^2\kappa^2\kappa_B W_0 m \sum_{j=0}^{Hm} \|\tilde{M}_{i-k}^j - \hat{M}_{i-k}^j\| \\ &= hC\kappa^2\kappa_B W_0 \sum_{j=0}^{Hm} \|\tilde{M}_{i-k}^j - \hat{M}_{i-k}^j\|.\end{aligned}$$

Where the first inequality is by $\|Q_h^j\| \leq \kappa^2(1-h\gamma)^{j-1} \leq \kappa^2$ and lemma B.3 of bounded estimation disturbance, the second inequality is by the fact that M_{i-k} have taken m times, the last equality is by $m = \frac{C}{h}$. Furthermore, we have that

$$\|v_t - \tilde{v}_t\| = \| - K\left(y_t - \tilde{y}_t\right)\| \leq hC\kappa^3\kappa_B W_0 \sum_{j=0}^{Hm} \left\|\tilde{M}_{i-k}^j - \hat{M}_{i-k}^j\right\|.$$

Therefore using Assumption 3.2, Lemma B.3 and Lemma 6.1 we immediately get that

$$\left| f_i\left(\tilde{M}_{i-H} \dots \tilde{M}_{i-k} \dots \tilde{M}_i\right) - f_i\left(\tilde{M}_{i-H} \dots \hat{M}_{i-k} \dots \tilde{M}_i\right)\right| \leq C^2\kappa^3\kappa_B W_0 \sum_{j=0}^{Hm} \|\tilde{M}_{i-k}^j - \hat{M}_{i-k}^j\|.$$

□

Lemma E.4. *(Bounding the Gradient) We have the following bound for the gradient:*

$$\|\nabla_M f_t(M \dots M)\|_F \leq \frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma}$$

Proof. Since M is a matrix, the ℓ_2 norm of the gradient $\nabla_M f_t$ corresponds to the Frobenius norm of the $\nabla_M f_t$ matrix. So it will be sufficient to derive an absolute value bound on $\nabla_{M^{[r]}_{p,q}} f_t(M,\dots,M)$ for all r,p,q. To this end, we consider the following calculation. Using lemma B.3 we get that $y_t(M\dots M), v_t(M\dots M) \le D$. Therefore, using Assumption 3.2 we have that

$$\left|\nabla_{M^{[r]}_{p,q}} c_t(M\dots M)\right| \le GD\left(\left\|\frac{\partial y_t(M)}{\partial M^{[r]}_{p,q}} + \frac{\partial v_t(M\dots M)}{\partial M^{[r]}_{p,q}}\right\|\right).$$

We now bound the quantities on the right-hand side:

$$\begin{aligned}\left\|\frac{\delta y_t(M\dots M)}{\delta M^{[r]}_{p,q}}\right\| &= \left\|h\sum_{i=0}^{2Hm} h\sum_{j=1}^{Hm}\left[\frac{\partial Q_h^j B M^{[i-j]}}{\partial M^{[r]}_{p,q}}\right]\hat{w}_{t-i}\mathbf{1}_{i-j\in[1,H]}\right\| \\ &\le h^2\sum_{i=r}^{r+Hm}\left\|\left[\frac{\partial Q_h^{i-r} B M^{[r]}}{\partial M^{[r]}_{p,q}}\right] w_{t-i}\right\| \\ &\le h^2\kappa^2 W_0\kappa_B\frac{1}{h\gamma} = \frac{h\kappa^2 W_0\kappa_B}{\gamma}.\end{aligned}$$

Similarly,

$$\left\|\frac{\partial v_t(M\dots M)}{\partial M^{[r]}_{p,q}}\right\| \le \kappa\left\|\frac{\delta y_t(M\dots M)}{\delta M^{[r]}_{p,q}}\right\| \le \kappa\frac{h\kappa^2 W_0\kappa_B}{\gamma} \le \frac{h\kappa^3 W_0\kappa_B}{\gamma}.$$

Combining the above inequalities with

$$f_i\left(\tilde{M}_{i-H},\dots,\tilde{M}_i\right) = \sum_{t=im}^{(i+1)m-1} c_t\left(y_t\left(\tilde{M}_{i-H},\dots,\tilde{M}_i\right), v_t\left(\tilde{M}_{i-H},\dots,\tilde{M}_i\right)\right).$$

gives the bound that

$$\|\nabla_M f_t(M\dots M)\|_F \le \frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma}.$$

□

Finally we prove Lemma 6.4:

Lemma 6.4. *Under Assumption 3.1 and 3.2, choosing $m = \frac{C}{h}$ and $\eta = \Theta(\frac{m}{Th})$, Algorithm 1 attains the following bound of R_2:*

$$\begin{aligned}R_2 &= \sum_{i=0}^{p-1} f_i(\tilde{M}_{i-H},\dots,\tilde{M}_i) - \min_{M\in\mathcal{M}}\sum_{i=0}^{p-1} f_i(M,\dots,M) \\ &\le \frac{4a}{\gamma}\sqrt{(\frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma} + C^2\kappa^3\kappa_B W_0 H^2)\frac{n}{h}}.\end{aligned}$$

Proof. By Lemma E.1 we have

$$R_2 \le 2D_f\sqrt{G_f\left(G_f + LH^2\right)p}$$

By Lemma E.2, Lemma E.3, and Lemma E.4 we have

$$\begin{aligned}R_2 &\le 2D_f\sqrt{G_f\left(G_f + LH^2\right)p} \\ &\le 2\frac{2a}{h\gamma}\sqrt{\frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma}(\frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma} + C^2\kappa^3\kappa_B W_0 H^2)\frac{n}{m}} \\ &\le \frac{4a}{\gamma}\sqrt{\frac{GDC^2\kappa^2(\kappa+1)W_0\kappa_B}{\gamma}(\frac{GDC\kappa^2(\kappa+1)W_0\kappa_B}{\gamma} + C^2\kappa^3\kappa_B W_0 H^2)\frac{n}{h}}.\end{aligned}$$

□

F. Proof of Lemma 6.5

In this section, we will prove the approximation value of DAC policy and optimal policy is sufficiently small. First, we introduce the following:

Lemma F.1. *For any two (κ, γ)-strongly stable matrices K^*, K, there exists $M = \left(M^1, \ldots, M^{Hm}\right)$ where*

$$M^i = (K - K^*)\left(I + h(A - BK^*)\right)^{i-1} ,$$

such that

$$c_t(x_t(M), u_t(M)) - c_t(x_t^*, u_t^*) \leq GDW_0\kappa^3 a(lh\kappa_B + 1)(1 - h\gamma)^{Hm} .$$

Proof. Denote $Q_h(K) = I + h(A - BK)$, $Q_h(K^*) = I + h(A - BK^*)$. By Lemma B.1 we have

$$x_{t+1}^* = h\sum_{i=0}^{t} Q_h^i(K^*)\hat{w}_{t-i} .$$

Consider the following calculation for $i \leq Hm$ and $M^i = (K - K^*)\left(I + h(A - BK^*)\right)^{i-1}$:

$$\begin{aligned}
\Psi_{t,i}\left(M, \ldots, M\right) &= Q_h^i(K) + h\sum_{j=1}^{i} Q_h^{i-j}(K)BM^j \\
&= Q_h^i(K) + h\sum_{j=1}^{i} Q_h^{i-j}(K)B\left(K - K^*\right)Q_h^{j-1}(K^*) \\
&= Q_h^i(K) + \sum_{j=1}^{i} Q_h^{i-j}(K)(Q_h(K^*) - Q_h(K))Q_h^{j-1}(K^*) \\
&= Q_h^i(K^*) ,
\end{aligned}$$

where the final equality follows as the sum telescopes. Therefore, we have that

$$x_{t+1}(M) = h\sum_{i=0}^{Hm} Q_h^i(K^*)\hat{w}_{t-i} + h\sum_{i=Hm+1}^{t} \Psi_{t,i}\hat{w}_{t-i} .$$

Then we obtain that

$$\left\|x_{t+1}(M) - x_{t+1}^*\right\| \leq hW_0 \sum_{i=Hm+1}^{t} \left(\left\|\Psi_{t,i}\left(M_*\right)\right\| + \left\|Q_h^i(K^*)\right\|\right) .$$

Using Definition 3.3 and Lemma B.1 we finally get

$$\begin{aligned}
\left\|x_{t+1}(M) - x_{t+1}^*\right\| &\leq hW_0\Big(\sum_{i=Hm+1}^{t} \left((lh\kappa_B + 1)a\kappa^2(1 - h\gamma)^{i-1}\right) + \kappa^2(1 - h\gamma)^i\Big) \\
&\leq W_0(lh\kappa_B + 2)a\kappa^2(1 - h\gamma)^{Hm} .
\end{aligned}$$

We also have

$$\begin{aligned}
\|u_t^* - u_t(M)\| &= \left\| -K^* x_t^* + K x_t(M) - h \sum_{i=0}^{Hm} M^i \hat{w}_{t-i} \right\| \\
&= \|(K-K^*)x_t^* + K(x_t(M) - x_t^*) - h \sum_{i=0}^{Hm} M^i \hat{w}_{t-i}\| \\
&= \|(K-K^*)h \sum_{i=0}^{t-1} Q_h^i(K^*)\hat{w}_{t-i} + K(x_t(M) - x_t^*) - h \sum_{i=0}^{Hm} M^i \hat{w}_{t-i}\| \\
&= \|K(x_t(M) - x_t^*) - h \sum_{i=Hm+1}^{t-1} (K-K^*)Q_h^{i-1}(K^*)\hat{w}_{t-i}\| \\
&= \|Kh \sum_{i=Hm+1}^{t-1} (\Psi_{t,i} - Q_h^{i-1}(K^*))\hat{w}_{t-i} - h \sum_{i=Hm+1}^{t-1} (K-K^*)Q_h^{i-1}(K^*)\hat{w}_{t-i}\| \\
&= \left\| h \sum_{i=Hm+1}^{t-1} K^* \left(Q_h^{i-1}(K^*) + \Psi_{t,i}\right) \hat{w}_{t-i} \right\| \\
&\le W_0 \kappa((1-h\gamma)^{Hm} + a(lh\kappa_B + 1)\kappa^2(1-h\gamma)^{Hm}) \\
&= W_0 \kappa(a(lh\kappa_B + 1)\kappa^2 + 1)(1-h\gamma)^{Hm}) \,,
\end{aligned}$$

where the inequality is by Definition 3.3 and Lemma B.2.

Finally, we have

$$\begin{aligned}
&|c_t(x_t(M), u_t(M)) - c_t(x_t^*, u_t^*)| \\
\le& |c_t(x_t(M), u_t(M)) - c_t(x_t^*, u_t(M))| + |c_t(x_t^*, u_t(M)) - c_t(x_t^*, u_t^*)| \\
\le& GD|x_t(M) - x_t^*| + GD|u_t(M) - u_t^*| \\
\le& GDW_0\kappa^3 a(lh\kappa_B + 1)(1-h\gamma)^{Hm} \,,
\end{aligned}$$

where the second inequality is by Assumption 3.2. □

Then we can prove our main lemma:

Lemma 6.5. *Under Assumption 3.1 and 3.2, Algorithm 1 attains the following bound of R_3:*

$$\begin{aligned}
R_3 &= \min_{M \in \mathcal{M}} \sum_{i=0}^{p-1} f_i(M, ..., M) - \sum_{i=0}^{p-1} \sum_{j=im}^{(i+1)m-1} c_i(x_i^*, u_i^*) \\
&\le 3n(1-h\gamma)^{Hm} GDW_0\kappa^3 a(lh\kappa_B + 1) \,.
\end{aligned}$$

Proof. By choosing

$$M^i = (K-K^*)\left(I + h(A - BK^*)\right)^{i-1} \,.$$

We know that

$$\|M^i\| = \|\left(K-K^*\right)\left(I + h(A-BK^*)\right)^{i-1}\| \le 2\kappa^3(1-\gamma)^{i-1} \,.$$

Therefore choose $a = 2\kappa^3$ we have $M = \{M^i\}$ in the DAC policy update class $\mathcal{M}$.

Then we have the analysis of the regret:

$$
\begin{aligned}
R_3 &= \min_{M\in\mathcal{M}} \sum_{i=0}^{p-1} f_i(M,...,M) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*) \\
&\le \min_{M\in\mathcal{M}} \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i(M),u_i(M)) - \sum_{i=0}^{p-1}\sum_{j=im}^{(i+1)m-1} c_i(x_i^*,u_i^*) + n\kappa^2(1+\kappa)(1-h\gamma)^{Hm+1}D \\
&\le 3n(1-h\gamma)^{Hm}GDW_0\kappa^3 a(lh\kappa_B+1)\,,
\end{aligned}
$$

where the first inequality is by Lemma 6.1 and the second inequality is by Lemma F.1.

□