Online Learning for Predictive Control with Provable Regret Guarantees

Deepan Muthirayan, Jianjun Yuan, Dileep Kalathil, and Pramod P. Khargonekar

Abstract—We address the problem of online learning in predictive control of an unknown linear dynamical system with time varying cost functions. We consider the setting where the control algorithm does not know the true system model and has only access to a fixed-length (that does not grow with the control horizon) preview of the future cost functions. We characterize the performance of the algorithm using the metric of dynamic regret, which is defined as the difference between the cumulative cost incurred by the algorithm and that of the best sequence of actions in hindsight. We propose a novel online learning predictive control algorithm called Optimistic MPC (O-MPC) algorithm. We show that under the standard stability assumption for the true underlying system, the O-MPC algorithm achieves $O(T^{2/3})$ dynamic regret.

I. INTRODUCTION

The control of dynamical systems with uncertainties such as modeling errors, parametric uncertainty, and disturbances is a central challenge in control theory. There is vast literature in the field on control synthesis for systems with such uncertainties. The robust control literature studies the problem of feedback control with modeling uncertainty and disturbances [1] while the adaptive control literature studies the control of systems with parametric uncertainty [2]. Typically, these classical approaches are concerned with stability and asymptotic performance guarantees.

Recently, there has been an increasing attention on the online control algorithms for dynamical systems with uncertainty in disturbances, system parameters and cost functions. This online control literature typically focuses on the finite time performance guarantees of the algorithms [3]–[6]. The typical objective in these works is to minimize the static regret, which is defined as the difference between the cumulative cost incurred by the online algorithm and the best policy from a certain class of policies. This is a key difference and challenge compared to the conventional adaptive control literature, and it requires combining techniques from statistical learning, online optimization and control. Most of the existing works in online control consider the setting where the online algorithm has access to only the past observations (of states, cost functions, and disturbances). On the other hand, in many practical problems such as robotics [7], energy systems [8], data-center management [9] etc., a finite-length preview of the future cost functions and/or disturbances are available to the control algorithm to compute the current control input. The question then is how do we develop online control algorithms that can exploit this preview to provably achieve better performance guarantees?

In the control theory literature, Model Predictive Control (MPC) addresses the class of problems where a preview of future cost functions are available to compute the current control input. The MPC is a well studied methodology in the control literature [10], [11]. However, the MPC literature primarily focuses on asymptotic performance guarantees. In sharp contrast to these existing works, our goal is to develop an online learning MPC algorithm with provable finite time performance guarantees. We focus on minimizing the metric of dynamic regret, which is defined as the difference between the cumulative cost of the online algorithm and that of the optimal sequence of control actions in hindsight (with full information). Thus, the dynamic regret is a stronger performance metric compared to the static regret. Our objective is to show that an optimally designed online learning MPC algorithm can achieve sub-linear dynamic regret using the preview information under minimal standard assumptions.

Related work: Recently, some works have addressed characterization of online performance (dynamic regret) of MPC algorithms [12]–[14]. In [12], the authors characterize the effect of preview on the dynamic regret of any baseline policy. They present an algorithm that improves the dynamic regret of any baseline policy exponentially with the length of the preview. In [13], the authors present guarantees for dynamic regret for a fixed LQR cost function with preview of disturbances. In [14], the authors extend these results to strongly convex cost functions. However, these works: (i) require the preview to grow at least logarithmically with the time horizon, and (ii) assume the system model is known. Significantly different from these works, we consider the more challenging setting of online learning in predictive control where only a fixed-length preview is available and the system model is unknown.

Another recent work [15] has also studied the problem of online learning in predictive control with a certainty equivalent approach, where the estimated parameter is directly used to compute the control policy. They consider a setting similar to ours, an unknown dynamical system without disturbances, with noise perturbed observations of the state. They show that their approach achieves $O(T^{2/3})$ regret. However, this approach crucially depends on a major assumption that the MPC policy computed using the estimated parameter is stabilizing for all systems within a certain radius from this estimate. In contrast, we develop a novel approach using the principle of optimism for online learning predictive control,
where the stability assumption is required only for the true underlying system, not for all the systems whose parameters lies within a ball around the true system parameter. We show that under this minimal and standard assumption, our proposed approach achieves $O(T^{2/3})$ dynamic regret.

Recently, many papers have studied the online regret performance in control problems with general time-varying costs, disturbances and known system model [5], [16]–[19]. A few others have also studied the problem with unknown linear systems. In [3], the authors provide an algorithm for the LQR problem with unknown dynamics that achieves a regret of $O(T^{2/3})$. In [20], the authors improve this result by providing an algorithm that achieves a regret of $O(T^{1/2})$ for the same problem. In [6], the authors generalize these results to provide sub-linear regret guarantee for online control with partial observation for both known and unknown systems. However, these works do not address online learning in predictive control which is the focus of this paper.

Main Contributions: We address the problem of online learning in predictive control of an unknown linear dynamical system with time varying cost functions. We assume that the system model is unknown to the control algorithm a priori and the control algorithm has only access to a fixed-length preview of the future cost functions. We propose a novel online learning MPC called Optimistic MPC (O-MPC) algorithm for this setting. We show that the O-MPC algorithm achieves a sublinear dynamic regret of $O(T^{2/3})$, under a standard assumption used for establishing the asymptotic stability of MPC controllers for the true underlying system.

To the best of our knowledge, this is the first work that gives a sub-linear dynamic regret guarantee for the online learning MPC problem with unknown system parameter and time varying cost function under standard assumptions.

Due to the page limits, we omit the proofs. All the proofs are given in the arXiv version [21].

A. Notations

We denote the spectral radius of a matrix $A$ by $\rho(A)$, the 2-norm of a vector by $\| \cdot \|_2$, the Frobenious norm of a matrix $X$ by $\| X \|_F$, the non-negative part of the real line by $\mathbb{R}_+$, the discrete time interval from $m_1$ to $m_2$ by $[m_1, m_2]$, the sequence $(x_1, x_2, \ldots, x_m)$ compactly by $x_{1:m}$. We denote the $j$th element of a vector $x$ by $x[j]$. We denote the $m$-ary cartesian power of a set $A$ by $A^m$. When a sequence of $m$-dimensional vectors $x_{1:m}$ are i.i.d. over the support $X^m$, we denote this by $x_t \overset{i.i.d.}{\sim} A^m$, where it is implicit that $t \in [m_1, m_2]$.

II. Problem Formulation and Preliminaries

We consider the online control of an unknown and partially observed linear dynamical system. The system evolution and the observation models are given by the equations

$$x_{t+1} = A^* x_t + B^* u_t, \quad y_t = x_t + \epsilon_t,$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$, $y_t \in \mathbb{R}^n$, $\epsilon_t \in \mathbb{R}^n$ are the state, control action, observation, and observation noise at time $t$, respectively. The system model is characterized by the parameters $A^* \in \mathbb{R}^{n \times n}$ and $B^* \in \mathbb{R}^{n \times m}$. For conciseness, we denote $\theta^* = [A^*, B^*]$, and we assume that $\theta^* \in \Theta \subset \mathbb{R}^{n \times (n + m)}$, where $\Theta$ is a known compact set.

A control policy $\pi$ selects a control action $u^*_t$ at each time $t$ depending on the available information, resulting in a sequence of actions $u^*_{1:T}$ and the corresponding state trajectory $x^*_{1:T}$. The cumulative cost of a policy $\pi$ under the system dynamics (1) is given by

$$J_T(\pi; \theta^*) = \sum_{t=1}^{T} c_t(x^*_t, u^*_t),$$

where $c_t$ is the cost function at time $t$. The typical goal is to find the optimal policy $\pi^*$ such that $\pi^* = \arg\min_{\pi} J_T(\pi; \theta^*)$. Clearly, computing $\pi^*$ requires the knowledge of the system model $\theta^*$ and the entire sequence of cost functions $c_{1:T}$.

In this work, we consider the setting where the decision maker (control policy) does not know the system parameter $\theta^*$ a priori, and has to learn the system parameter from the online observations. Moreover, the policy only has access to a fixed-length preview of the next $M$ cost functions, $c_{t+1:M-1}$ at each time step $t$. More precisely, the policy has only the following information available at each time $t$ for selecting the action $u_t$: (i) past observations $y_{1:t-1}$, current observation $y_t$, and past control inputs $u_{1:t-1}$, (ii) past cost functions $c_{1:t-1}$, and (iii) a preview of the next $M$ cost functions $c_{t+1:M-1}$. The policy has to learn the unknown system parameter from the online observations and adapt with respect to the revealed future cost functions. Thus, such a policy is called an online learning policy and the problem is a online learning problem.

The performance of an online learning policy $\pi$ is measured in terms of the dynamic regret, defined as

$$R_T(\pi) = J_T(\pi; \theta^*) - J_T(\pi^*; \theta^*).$$

In other words, dynamic regret is the difference between the policy $\pi$ and that of the best policy $\pi^*$ which has the complete information of the model parameter and loss functions. Note that the dynamic regret is different from the more commonly used metric of static regret [5], [6], where the cost of the online algorithm is compared with that of the best fixed policy from a specific class. Thus, the dynamic regret metric is a harder online performance metric compared to the standard static regret in online control. Our goal is to find an online learning policy that minimizes the dynamic regret in the above information setting.

We make the following assumptions on the system model.

Assumption 1 (System model). (i) The set of possible system parameters $\Theta$ is a known compact set. Moreover, $\|\theta\|_F \leq S$, $\forall \theta \in \Theta$. (ii) $\rho(A^*) < 1$, where $\rho(\cdot)$ is the spectral radius. The pair $(A^*, B^*)$ is controllable. (iii) The observation noise $\epsilon_t$ is uniformly bounded, i.e., $\|\epsilon_t\|_2 \leq \epsilon_c$, $\forall t$. (iv) The cost functions $c_{1:T}$ are continuous and locally Lipschitz with a uniform Lipschitz constant for all $t$. 

6667
The assumptions on the boundedness of $\Theta$ and the spectral radius are standard in the online learning and control literature when the system model is unknown [3], [6], [20], [22]. Our assumption that the noise is bounded is similar to [6], which is the only other work that also studies online control of unknown systems with general cost functions. Similar to [15], which is the closest to our work, we also do not consider stochastic disturbances in our dynamics. The problem with disturbances is more challenging, although we believe the proof techniques we employ can be extended to the setting with the disturbances. We plan to address the problem with disturbance as a subsequent work.

A. Model Predictive Control: Preliminaries

Model Predictive control (MPC) is one of the most popular approaches for control design when only a preview of the cost functions are available [10], [11]. The standard MPC algorithm needs the knowledge the system parameter $\theta = [A(\theta), B(\theta)]$ for computing the optimal control sequence for that system. Algorithm 1 gives the formal description of the MPC algorithm with an $M$-step preview. Given the current time step $t$, current state $x_t$, preview of the cost functions $c_{t:t+M-1}$, and the system parameter $\theta$ as the input, the MPC algorithm gives the control action $u_t = \text{MPC}(t, x_t, c_{t:t+M-1}, \theta)$ as the output.

**Algorithm 1** MPC($t, x_t, c_{t:t+M-1}, \theta$)

1. With the initialization $\tilde{x}_t = x_t$, compute $\tilde{u}_{t:t+M-1} = \arg\min_{\tilde{u}_{t:t+M-1}} \sum_{k=t}^{t+M-1} c_k(\tilde{x}_k, \tilde{u}_k)$, s.t. $\tilde{x}_{k+1} = A(\theta)\tilde{x}_k + B(\theta)\tilde{u}_k$.

2. Output: $\tilde{u}_t$

There has been significant works on analyzing the stability of systems that employ MPC policies under various assumptions [23], [24]. Clearly, any online learning MPC algorithm also has to ensure the stability of the system. So, we follow the same assumptions used in the literature that ensure stability of systems that employ MPC policies.

Define the $M$-step cost-to-go function, denoted by $V_t$, as

$$V_t(x; \theta) = \min_{u_{t:t+M-1}} \sum_{k=t}^{t+M-1} c_k(x_k, u_k),$$

s.t. $x_{k+1} = A(\theta)x_k + B(\theta)u_k, x_t = x$. (4)

**Assumption 2** (Stability assumptions). There exist positive scalars $\alpha, \sigma$ and a continuous function $\sigma: \mathbb{R}^n \to \mathbb{R}_+$ such that: (i) $c_t(x, u) \geq \alpha \sigma(x)$, $\forall x, \forall u, \forall t$, (ii) $V_t(x; \theta) \leq \sigma(x)$, $\forall x, \forall t$, and (iii) $\lim_{||x|| \to \infty} \sigma(x) = \infty$.

Under the above stability assumption, [23] showed that the system with the parameter $\theta$ under the MPC policy has global asymptotic stability. We will make use of this assumption in analyzing our approach. In contrast to other MPC approaches like [24], the stability assumption we use from [23] does not assume the existence of a Lyapunov-like function directly, which is a stronger form of stability assumption. Unlike [15], which is the only other work to discuss online learning predictive control, we do not require the stability assumption to hold for all the systems within a neighborhood around the true underlying system.

**Remark 1.** The cost functions satisfying 2 trivially include time varying quadratic, for eg., $c_t(x) = (x - b)^TQ_t(x - b) + u^TR_tu$, where $b$ is an offset, and non-convex functions of the state (see [23] for examples). Thus, the cost functions we consider are quite general.

III. ALGORITHM AND REGRET GUARANTEE

In this section, we present our online learning MPC algorithm called the Optimistic MPC (O-MPC) algorithm, and its regret performance guarantee. The algorithm operates in two phases: (i) exploration phase, and (ii) control phase. In the exploration phase, the algorithm follows a pure exploration strategy to estimate the unknown system parameter. In the control phase, the algorithm selects an optimistic parameter from a high confidence region around the parameter estimate and employs an MPC policy based on this optimistic parameter.

A. Exploration Phase

The goal of the exploration phase is to explore the system and collect the observations to estimate the unknown parameter $\theta^*$ up to a desired accuracy with high probability. This is achieved by a pure exploration strategy for the first $T_0$ steps.

We adopt the approach of [25]. The algorithm sets the control actions in the exploration phase as an i.i.d. random sequence as given by

$$u_t \sim \{\pm 1\}^m,$$

where $P(u_t[j] = 1/2$, where $\{\pm 1\}$ denotes the set with elements 1 and $-1$ and $\{\pm 1\}^m$ denotes the $m$-ary cartesian power.

At the end of the exploration phase, the algorithm computes the following quantities

$$N_j = \frac{1}{T_0 - n} \sum_{i=1}^{T_0 - n - 1} y_{t+j+1}u_t^T, \forall j \in [1, n].$$

Let

$$\tilde{C}_0 = [N_0 \ N_1 \ldots N_{n-1}], \ \tilde{C}_1 = [N_1 \ N_2 \ldots N_n].$$

Then, the algorithm computes the estimate of $\theta^*$ as

$$\tilde{\theta}_{ls} = [\tilde{A} \ \tilde{B}], \ \tilde{B} = N_0, \ \tilde{A} = \tilde{C}_1\tilde{C}_0^{-1}(\tilde{C}_0\tilde{C}_0^{-1})^{-1}. \ (6)$$

Given the parameter estimate $\tilde{\theta}_{ls}$, the algorithm computes a high confidence set as given by

$$\tilde{\Theta} = \left\{ \theta : ||\theta - \tilde{\theta}_{ls}||_F \leq \beta(\delta) \right\}, \ (7)$$

where

$$\beta(\delta) = \sqrt{2 \times 10^3 n^2 k^3 (\sqrt{m} + c_F \gamma^{-1}) m \log(mn^2/\delta) / T_0^{1/2}}. \ (8)$$

6668
Algorithm 2 O-MPC($t, \hat{x}_t, c_{t:t+M-1}, \Theta$)

**Input:**
With the initialization $\hat{x}_t = \hat{x}_t$, compute
$$(\hat{u}_{t:t+M-1}, \hat{\theta}_t) = \arg\min_{\hat{u}_{t:t+M-1},\theta \in \Theta} \sum_{k=t}^{t+M-1} c_k(\hat{x}_k, \hat{u}_k),$$
s.t. $\hat{x}_{k+1} = A(\theta)\hat{x}_k + B(\theta)\hat{u}_k$

**Output:** $(\hat{u}_t, \hat{\theta}_t)$

Algorithm 3 O-MPC Algorithm

**Input:** $T_0, \delta$

**Estimation Phase:**
for $t = 1 : T_0$
  select control action $u_t$ according to (5)
end

Estimate the parameter $\hat{\theta}_t$ according to (6)

Estimate the confidence region $\Theta$ according to (7) -(8)

Set $\hat{x}_{T_0+1} = y_{T_0+1}$

**Control Phase:**
for $t = T_0 + 1 : T$
  Select the control action $u^\pi_t$ and the optimistic parameter $\hat{\theta}_t$ as $(u^\pi_t, \hat{\theta}_t) = O\text{-MPC}(t, \hat{x}_t, c_{t:t+M-1}, \Theta)$
  $\hat{x}_{t+1} = A(\hat{\theta}_t)\hat{x}_t + B(\hat{\theta}_t)u^\pi_t$
end

Here, the constants $c_\rho$ and $\gamma_\rho$ are constants such that $\|A^k\| \leq \gamma_\rho (1 - \rho)^k$, and $\kappa$ is a sufficiently large constant such that $\|(C_0C_0^T)^{-1}\| \leq \kappa$, where $C_0 = [B, \ldots, A^{n-1}B]$.

The implicit assumption here is that these constants are known and can be used to construct the high confidence set. We note that this assumption is essential to construct the high confidence set.

**Proposition 1.** Suppose Assumption 1 holds, and $u_1:T_0$ is given by (5). Then, for sufficiently large $T_0$, with probability greater than $1 - \delta$, $\theta^* \in \Theta$.

We note that the lower bound for $T_0$ given in the above proposition does not depend on the horizon $T$. The proof is exactly the same as the proof of [25, Theorem 19]. We refer the reader to [25, Theorem 19] for the details of the proof.

B. O-MPC Algorithm

The O-MPC algorithm uses an optimistic approach that simultaneously selects the optimistic parameter from the confidence region $\Theta$ and the optimal control action with respect to this optimistic parameter. More precisely, at each time $t \in [T_0 + 1 : T]$, the O-MPC algorithm selects the control action $u^\pi_t$ and the optimistic parameter $\hat{\theta}_t$ as
$$(u^\pi_t, \hat{\theta}_t) = O\text{-MPC}(t, \hat{x}_t, c_{t:t+M-1}, \Theta),$$

where the O-MPC subroutine is given in Algorithm 2. The complete O-MPC algorithm is formally presented in Algorithm 3.

C. Regret Performance Guarantee

We now formally present the regret guarantee of the O-MPC algorithm (denoted as $\pi_k$).

**Theorem 1** (Regret of the O-MPC Algorithm). Suppose Assumption 1 holds, $\theta^*$ satisfies Assumption 2, and $T_0$ satisfies the conditions in Proposition 1. Assume that $\bar{\sigma}/\sigma < 2$. Fix $T_0 = T^{2/3}$. Then there exists $M, T$ such that, for $T > T$, with probability greater than $(1 - \delta)$

$$R_T(O\text{-MPC}) \leq O(T^{2/3})$$

**Remark 2.** The condition on the ratio of upper and lower bound to the cost functions, i.e., $\bar{\sigma}/\sigma < 2$ can be relaxed by making additional but less restrictive assumptions. The ratio condition is only required to establish the boundedness of the state, as we shall show in the proof later. For lack of space we do not discuss this relaxation.

IV. REGRET ANALYSIS

In this section, we give the key proof ideas used in the analysis of O-MPC algorithm and obtaining the $O(T^{2/3})$ dynamic regret.

Given a sequence of system parameters $\theta_{T_1:T_2}$, a sequence of control actions $u_{1:T}$, and an initial state $x$, we define $J_{T_1:T_2}(u_{T_1:T_2}; \theta_{T_1:T_2})$ as

$$J_{T_1:T_2}(u_{T_1:T_2}; \theta_{T_1:T_2}) = \sum_{t=T_1}^{T_2} c_t(x_t, u_t),$$
s.t. $x_{t+1} = A(\theta_t)x_t + B(\theta_t)u_t, \quad x_{T_1} = x,$

(10)

for any $T_1, T_2 \in [1, T]$. We make the dependence on the initial state $x$ implicit as it will be clear from the context. If $\theta_t = \theta, \forall t \in [T_1, T_2]$, we will simplify the above notation as $J_{T_1:T_2}(u_{T_1:T_2}; \theta)$. Let $u^\pi_{1:T}$ and $u^*_1:T$ be the sequence of control actions generated by the policies $\pi$ and $\pi^*$, respectively. For analyzing the regret, we decompose it into three terms as follows:

$$R_T(\pi) = J_{1:T_0}(u^\pi_{1:T_0}; \theta^*) - J_{1:T_0}(u^*_1:T_0; \theta^*)$$

$$+ J_{T_0+1:T}(u^\pi_{T_0+1:T}; \theta^*) - J_{T_0+1:T}(u^*_1:T; \theta^*)$$

$$+ J_{T_0+1:T}(u^\pi_{T_0+1:T}; \theta_{T_0+1:T}) - J_{T_0+1:T}(u^*_1:T; \theta_{T_0+1:T})$$

(11)

Term I characterizes the regret due to the exploration phase. We show that, under the exploration strategy we use, the regret due to exploration is bounded by the length of the exploration phase. The key challenge involved here is to show that the system state does not grow unbounded during the exploration phase. For this, we make use of the fact that the spectral radius of $A^*$ is strictly less than one and the control sequences are bounded. We then use the fact that the cost functions are locally Lipschitz to show that the realized cost at each time step of the estimation phase is bounded.
Combining these, one can show that the regret of Term I is $O(T_0)$. We formally state the bound on Term I below.

**Proposition 2** (Regret of Term I). Suppose Assumption 1 holds. Let Term I be as defined in (11). Then, under the O-MPC algorithm,

$$\text{Term I} \leq O(T_0)$$

Note that, if we set $T_0 = T^{2/3}$ as specified in Theorem 1, then the regret due to Term I is $O(T^{2/3})$. Due to page limit, the proof is deferred to [21].

The O-MPC algorithm generates the sequence of control actions $u_{T_0+1:T}^*$ using the parameter sequence $\theta_{T_0+1:T}$. However, these control actions are applied on the true system with parameter $\theta^*$. Term II characterizes the regret due to this difference. The analysis of Term II is challenging because we need to consider the sequence of parameters $\theta_{T_0+1:T}$ selected by the algorithm. Contingent on the states being bounded, we upperbound the difference of the states of any two systems driven by the same sequence of control actions at a time $t$ by a decaying sum of the parameter difference from $t$ to $T_0 + 1$. We then separately show that the states are bounded. This will also imply that the control actions are bounded. The cost functions being locally Lipschitz and the states and control actions being bounded, Term II can now be upperbounded by the cumulative error of the parameter sequence $\theta_{T_0+1:T}$. Since $\|\theta^* - \theta_t\| \leq O(1/\sqrt{T_0})$ for all $t$ (using Proposition 1), we get a net upperbound of $O(T/\sqrt{T_0})$ with a high probability. We state this result formally below.

**Proposition 3** (Regret of Term II). Suppose Assumption 1 holds, $\theta^*$ satisfies Assumption 2, and $T_0$ satisfies the conditions in Proposition 1. Assume that $\bar{\sigma}/\sigma < 2$. Let Term II be as defined in (11). Then there exists $M,T$ such that, for $T > T$, under O-MPC algorithm, with probability greater than $(1 - \delta)$

$$\text{Term II} \leq O(T/\sqrt{T_0})$$

Here, if we set $T_0 = T^{2/3}$, Term II is $O(T^{2/3})$. Due to page limit, the proof is deferred to [21].

Bounding Term III is significantly more challenging because the stability assumption is true only for the true system $\theta^*$, whereas the quantity to be analyzed is the cumulative cost of the MPC controller for an arbitrary system within the confidence region. The proof uses the fact that estimate is optimistic to leverage Assumption 2 satisfied by $\theta^*$. This leads to a Lyapunov-like condition with an additional term that is proportional to the difference between $\theta^*$ and $\hat{\theta}_t$. The novelty of the proof technique is how the optimistic estimate is used to establish the Lyapunov-like condition. The additional term at every time step leads to the $O(T/\sqrt{T_0})$ overall regret.

**Proposition 4** (Regret of Term III). Suppose Assumption 1 holds, $\theta^*$ satisfies Assumption 2, and $T_0$ satisfies the conditions in Proposition 1. Assume that $\bar{\sigma}/\sigma < 2$. Let Term III be as defined in (11). Then there exists $M,T$ such that, for $T > T$, under O-MPC algorithm, with probability greater than $(1 - \delta)$

$$\text{Term III} \leq O(T/\sqrt{T_0})$$

Note that, by setting $T_0 = T^{2/3}$, the regret of Term III becomes $O(T^{2/3})$. Due to page limit, the proof is deferred to [21].

The proof of the main result, Theorem 1, follows by setting $T_0 = T^{2/3}$ and combining Proposition 2, Proposition 3 and Proposition 4.

### V. Numerical Experiments

In this section, we present three numerical examples to illustrate the performance of the O-MPC algorithm. In all examples, we consider a linear dynamical system as given in (1) with $n = 2$ and $m = 1$. In each example, the system matrices $A^*$ and $B^*$ are chosen randomly, with the elements of $A^*$ in the range $[0, 0.5]$ and the elements of $B^*$ in the range $[0, 1]$. The preview $M$ is set to be 5 in all the examples. This $M$ is computed as $(\bar{\sigma}/\alpha)^2 + 1$, as given in Theorem 1, for the fixed quadratic cost given by $Q = I$ and $R = I$. We select different sets of cost functions for each example, as:

**Example 1**: we select quadratic cost functions, $c_t(x,u) = (x - b)\top Q(x - b) + u\top R u$, where $b = [0.01, 0.01]$, and $Q_t$ and $R_t$ are randomly chosen diagonal matrices with each diagonal element lying in the range $[0.375, 0.625]$. This example hence illustrates a specific case of online predictive linear quadratic control with time varying cost functions and unknown system model.

**Example 2**: we select the sequence of cost functions given by $c_t(x,u) = \sigma(x) + u\top R u$, where $\sigma(x) := \inf_{y \in \mathcal{X}} \|y - x\|^2$, and $\mathcal{X}$ is the ball of radius 0.25 centered at 0.5. This example focus on the convergence to a specific region in the state space centered by $\mathcal{X}$, which is often an important objective in predictive control.

**Example 3**: we select non-quadratic cost functions given by $c_t(x,u) = \|x[1] - b\|^2 + \|x[2] - b\|^2 + u\top u$, where $b = 0.1$. This example illustrates a specific case of online predictive control with non-convex cost functions. While most of the online predictive control techniques are reliant on convexity assumption, this assumption need not hold in all control problems (see [23] for details).

We note that, in these examples, because the costs are offset from zero, $u_t = 0$ will not achieve a sub-linear regret.

The variation of the regret for all these examples is shown in Fig. 1. We find that the scaling of the regret in all these examples matches with our theoretical guarantee.

### VI. Conclusion

In this work, we present a novel online learning and control algorithm for predictive control of linear dynamical systems. Our work sheds light on methods, conditions and analysis for predictive control of unknown systems with limited preview. Our proposed O-MPC algorithm guarantees $O(T^{2/3})$ dynamic regret without making no additional assumption other than the standard stability assumptions required for the asymptotic analysis of MPC algorithms.
In future, we plan to extend this algorithm and analysis to systems with disturbances.

REFERENCES

[1] K. Zhou, J. Doyle, and K. Glover, Robust and optimal control. Prentice hall, 1996.

[2] S. Sastry and M. Bodson, Adaptive control: stability, convergence and robustness. Dover Publications, 2011.

[3] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, “Regret bounds for robust adaptive control of the linear quadratic regulator,” Neural Information Processing Systems (NeurIPS), pp. 4188–4197, 2018.

[4] H. Mania, S. Tu, and B. Recht, “Certainty equivalence is efficient for linear quadratic control,” Neural Information Processing Systems (NeurIPS), pp. 10 154–10 164, 2019.

[5] N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh, “Online control with adversarial disturbances,” International Conference on Machine Learning (ICML), pp. 111–119, 2019.

[6] M. Simchowitz, K. Singh, and E. Hazan, “Improper learning for non-stochastic control,” Conference on Learning Theory (COLT), pp. 3320–3436, 2020.

[7] G. Shi, X. Shi, M. O’Connell, R. Yu, K. Azizzadenesheli, A. Anandkumar, Y. Yue, and S.-J. Chung, “Neural lander: Stable drone landing control using learned dynamics,” International Conference on Robotics and Automation (ICRA), pp. 9784–9790, 2019.

[8] S. Vazquez, J. Rodriguez, M. Rivera, L. G. Franquelo, and M. Norambuena, “Model predictive control for power converters and drives: Advances and trends,” IEEE Transactions on Industrial Electronics, vol. 64, no. 2, pp. 935–947, 2016.

[9] N. Lazic, C. Boutilier, T. Lu, E. Wong, B. Roy, M. K. Ryu, and G. Imwalle, “Data center cooling using model-predictive control,” Neural Information Processing Systems (NeurIPS), pp. 3818–3827, 2018.

[10] J. B. Rawlings, D. Q. Mayne, and M. Diehl, Model predictive control: theory, computation, and design. Nob Hill Publishing, 2017.

[11] F. Borrelli, A. Bemporad, and M. Morari, Predictive control for linear and hybrid systems. Cambridge University Press, 2017.

[12] Y. Li, X. Chen, and N. Li, “Online optimal control with linear dynamics and predictions: Algorithms and regret analysis,” Neural Information Processing Systems (NeurIPS), pp. 14 887–14 899, 2019.

[13] C. Yu, G. Shi, S. Chung, Y. Yue, and A. Wierman, “The power of predictions in online control,” in Neural Information Processing Systems (NeurIPS), 2020.

[14] Y. Lin, Y. Hu, G. Shi, H. Sun, G. Qu, and A. Wierman, “Perturbation-based regret analysis of predictive control in linear time varying systems,” Neural Information Processing Systems (NeurIPS), vol. 34, 2021.

[15] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, “Model learning predictive control in nonlinear dynamical systems,” IEEE Conference on Decision and Control (CDC), pp. 757–762, 2021.

[16] Y. Abbasi-Yadkori, P. Bartlett, and V. Kanade, “Tracking adversarial targets,” International Conference on Machine Learning, pp. 369–377, 2014.