Online Learning for Receding Horizon Control with Provable Regret Guarantees

Deepan Muthirayan
University of California, Irvine

Jianjun Yuan
Expedia Group

Dileep Kalathil
Texas A&M University

Pramod P. Khargonekar
University of California, Irvine

Abstract

We address the problem of learning to control an unknown linear dynamical system with time varying cost functions through the framework of online Receding Horizon Control (RHC). 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 an 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 two different online RHC algorithms to address this problem, namely Certainty Equivalence RHC (CE-RHC) algorithm and Optimistic RHC (O-RHC) algorithm. We show that under the standard stability assumption for the model estimate, the CE-RHC algorithm achieves $O(T^{2/3})$ dynamic regret. We then extend this result to the setting where the stability assumption hold only for the true system model by proposing the O-RHC algorithm. We show that O-RHC algorithm achieves $O(T^{2/3})$ dynamic regret but with some additional computation.

1 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 (Skogestad and Postlethwaite, 2007) while the adaptive control literature studies the control of systems with parametric uncertainty (Sastry and Bodson, 2011). Typically, these classical approaches are concerned with stability and asymptotic performance guarantees.

Recently, there has been an increased attention on online control algorithms for linear dynamical systems with adversarial cost functions and disturbances, focusing on their finite time performance guarantees (Dean et al., 2018; Cohen et al., 2019; Mania et al., 2019; Agarwal et al., 2019a,b; Simchowitz et al., 2020). 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. However, most of the existing works only consider the setting where the online algorithm only has access to the past observations (of states, cost functions, and disturbances) to compute the current control input. On the other hand, in many practical problems such as robotics (Baca et al., 2018; Shi et al., 2019), energy systems (Vazquez et al., 2016), data-center management (Laiz et al., 2018), 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, Receding Horizon Control (RHC), also known as Model Predictive Control (MPC), addresses the class problems where a preview of future cost functions are available to compute the current control input. The RHC is a well studied methodology in the control literature (Mayne et al., 2000; Angeli et al., 2011; Camacho and Alba, 2013; Grüne and Stieler, 2014; Angeli et al., 2016; Borrelli et al., 2017; Rosolia and Borrelli, 2017).
However, the control theory literature on RHC primarily focuses on asymptotic performance guarantees. In sharp contrast to these existing works, our goal is to develop an online RHC algorithm with provable finite time performance guarantees. We will 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). The optimal sequence of control actions with full information is also called the optimal dynamic offline policy. The dynamic regret is a stronger performance metric compared to the static regret. Our objective is to show that an optimally designed online RHC algorithm can achieve sub-linear dynamic regret using the preview information.

Recently, some works have addressed the online RHC problem, focusing on minimizing the dynamic regret. Li et al. (2019) considered linear systems with strongly cost functions and a fixed-length preview. Given a baseline algorithm, they proposed an approach to improve its dynamic regret with the length of the preview. However, the performance is entirely determined by the baseline policy, which is left unspecified. In another work, Yu et al. (2020) considered the online linear quadratic regulator (LQR) problem with a preview of the disturbances. They showed that the proposed algorithm achieves an $O(1)$ dynamic regret given an increasing ($O(\log T)$) preview. In addition to the requirement of this increasing preview, they also restricted their setting to known quadratic cost function. Moreover, both works (Li et al., 2019; Yu et al., 2020) assume that the system model is known to the algorithm a priori. Significantly different from these works, we consider the more challenging setting of the online RHC problem where the model of the system is unknown, the cost functions are adversarial, and only a fixed-length preview is available. Our main contributions are the following.

**Main Contributions:** We address the problem of learning to control an unknown linear dynamical system with time varying cost functions through the framework of online RHC algorithm. 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. Under the standard assumptions used for establishing the asymptotic stability of RHC controllers (Grimm et al., 2005), we show that our proposed online RHC algorithm can achieve sublinear dynamic regret. In particular, we propose two different online RHC algorithms, namely Certainty Equivalence RHC (CE-RHC) algorithm and Optimistic RHC (O-RHC) algorithm. We show that under the standard stability assumption for the model estimate, the CE-RHC algorithm achieves $O(T^{2/3})$ dynamic regret. We then extend this result to the setting where the stability assumption hold only for the true system model by proposing the O-RHC algorithm. We show that O-RHC algorithm achieves $O(T^{2/3})$ dynamic regret but with some additional computation. To the best of our knowledge, this is the first work that gives a sublinear dynamic regret guarantee for the online RHC problem with unknown systems.

### 1.1 Related Work

**Online control:** Online regret analysis is an extensively studied topic in online learning problems (Hazan, 2016; Shalev-Shwartz et al., 2011; Yuan and Lamperski, 2020). Recently, many papers studied the online regret performance in control problems with general time-varying costs, disturbances and known system model (Abbasi-Yadkori et al., 2011; Cohen et al., 2013; Agarwal et al., 2019b). Goel and Wierman (2019). A few others also studied the problem with unknown linear systems. In (Dean et al., 2018), the authors provide an algorithm for the LQR problem with unknown dynamics that achieves a regret of $O(T^{2/3})$. In (Cohen et al., 2019), the authors improved this result by providing an algorithm that achieves a regret of $O(T^{1/2})$ for the same problem. Simchowitz et al. (2020) generalized these results to provide sub-linear regret guarantee for online control with partial observation for both known and unknown systems. All these works, however, focus on static regret minimization.

**RHC:** Many receding horizon control based methods have been proposed for managing disturbances and uncertainties in the system dynamics. For example, some works handle disturbances or uncertainties by robust or chance constraints (Langson et al., 2004; Goulart et al., 2006; Limon et al., 2010; Tempo et al., 2012; Goulart et al., 2016). Adaptive RHC techniques that adapt online when the system model is unknown have also been proposed (Fukushima et al., 2007; Adetola et al., 2009; Aswani et al., 2013; Tanaskovic et al., 2019; Bujarbaruah et al., 2019). These methods primarily focus on constraint satisfaction, stability and in some cases performance improvement using the adapted models. In contrast to these works, we consider non-asymptotic performance of an online RHC algorithm. There are considerable amount of papers that provide performance analysis of RHC under both time-invariant costs (Angeli et al., 2011; Grimm and Stieler, 2014; Grimm and Panin, 2015) and time varying costs (Ferramosca et al., 2010; Angeli et al., 2016; Ferramosca et al., 2014; Grimm and Pirkelmann, 2017). However, most of
these studies also focus on asymptotic performance.

1.2 Notation

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 $[m_1, m_2]$, and the sequence $(x_{m_1}, x_{m_1+1}, ..., x_{m_2})$ compactly by $x_{m_1:m_2}$.

2 Problem Formulation and Preliminaries

2.1 Problem Statement

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

$$x_{t+1} = A^* x_t + B^* u_t, \quad y_t = x_t + \epsilon_t,$$  \hspace{2cm} (1)

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 of the system, 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 $u_t^T$ at each time $t$ depending on the available information, resulting in a sequence of actions $u_{1:T}^T$ and the corresponding state trajectory $x_{1:T}^T$. The cumulative cost of a policy $\pi$ under the system dynamics (1) is given by

$$J_T(\pi; \theta^*) = \sum_{t=1}^{T} c_t(x_t^T, u_t^T),$$  \hspace{2cm} (2)

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

In most real-world control problems, it is not possible to find the optimal policy directly as described above because of two important practical concerns: (i) the true system parameter $\theta^*$ may be unknown to the decision maker a priori, and (ii) the current and future cost functions, $c_{1:T}$, may be unknown to the decision maker at any time step $t$. The policy that can achieve the minimum possible cumulative cost should then depend on the information available to the policy at each time step.

In this work, we consider a 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 has only access to a fixed-length preview of the next $M$ cost functions, $c_{t+1:M-1}$, for making the control decision $u_t$ 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. So, such a policy is called an online learning policy.

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

$$R_T(\pi) = J_T(\pi; \theta^*) - J_T(\pi^*; \theta^*).$$  \hspace{2cm} (3)

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. Our goal is to find an online learning policy that minimizes the dynamic regret. We note that the dynamic regret is a stronger performance metric compared to the more commonly used static regret (Agarwal et al., 2019; Simchowitz et al., 2020) where the cost of the online algorithm is compared with that of the best fixed policy from a specific class.

We give a motivating example of the LQ tracking controller with preview, which is an important application in a number of areas (Anderson and Moore, 2007).

**Example 1** (Tracking controller). The goal of a tracking controller is to track a given sequence of way points $\bar{x}_{1:T}$ at the minimum cumulative cost. In the linear quadratic tracking problem, the dynamics is assumed to be linear as in (1), and the cost is quadratic of the form $c_t(x, u) = (x - \bar{x}_t)^	op Q(x - \bar{x}_t) + u^	op R u$. Often, only a finite preview of the way points $\bar{x}_{t+1:M-1}$ is available before making the decision $u_t$, which is equivalent to having a finite preview of cost functions $c_{t+1:M-1}$. In addition to this, the system parameter $\theta^*$ may also be unknown a priori.

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 a martingale difference sequence w.r.t. a filtration $\mathcal{F}_t$. Also, the observation
noise 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 \( t \).

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 (Abbasi-Yadkori and Szepesvári, 2011; Dean et al., 2018; Cohen et al., 2019; Simchowitz et al., 2020). The assumption that the noise sequence is a martingale difference sequence is also standard in prior works with stochastic disturbances and noise (Abbasi-Yadkori and Szepesvári, 2011; Dean et al., 2018; Cohen et al., 2019). Our assumption that the noise is bounded is similar to (Simchowitz et al., 2020), which addresses the online control of unknown linear systems with general cost functions.

2.2 Receding Horizon Control: Preliminaries

Receding horizon control is one of the most popular approaches for control design when only a preview of the cost functions are available (Morari and Lee, 1999; Mayne et al., 2000; Borrelli et al., 2017). The standard RHC algorithm needs the knowledge of 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 RHC algorithm with an \( M \)-step preview. Given the current time step \( t \), current state \( x_t \), preview of the cost functions \( c_{t:M-1} \), and the system parameter \( \theta \) as the input, the RHC algorithm provides the control action \( u_t = \text{RHC}(t, x_t, c_{t:M-1}, \theta) \) as the output. Unlike the optimal policy for the standard LQR problems, the optimal RHC policy for a linear system with general cost functions need not be linear. So, characterizing the stability properties of the RHC algorithm with general cost functions is much more challenging compared to the LQR setting. There has been significant works on analyzing the stability of systems that employ RHC policies under various assumptions (Lazar et al., 2008; Grimm et al., 2005). Clearly, any online learning RHC 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 RHC policies.

Define the \( M \)-step cost-to-go function, denoted \( V_t \), 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 \( a, \sigma \) and a continuous function \( \sigma : \mathbb{R}^n \to \mathbb{R}_+ \) such that:

(i) \( c_t(x, u) \geq a \sigma(x) \), \( \forall x, \forall u, \forall t \),  
(ii) \( V_t(x; \theta) \leq \sigma(x), \forall x, \forall t \), and  
(iii) \( \lim_{\|x\| \to \infty} \sigma(x) = \infty \).

Under the above assumption, (Grimm et al., 2005) showed that the system with the parameter \( \theta \) under the RHC policy has global asymptotic stability. We will make use of the assumption in analyzing our approach. In contrast to other RHC approaches like (Lazar et al., 2008), the stability assumption we use from (Grimm et al., 2005), does not assume the existence of a Lyapunov-like function directly, which is a stronger stability assumption.

We note that the above assumptions are satisfied for a number of systems and cost functions. For example, the linear quadratic control problem given in Example 1 satisfies the above assumption; see appendix for the proof. In appendix, we also give an example of non-quadratic, non-convex cost functions.

We note that most of the works analyzing the stability of RHC policies provide only asymptotic guarantees, including (Grimm et al., 2005). Our focus is on analyzing the finite-time performance of the RHC policy using the metric of dynamic regret in a more challenging setting with unknown system parameter.

We note that the prior online predictive control works (Li et al., 2019; Yu et al., 2020) do not use this stability assumption. However, they require a preview of length at least the logarithm of the whole control horizon \( O(\log T) \) to achieve sub-linear dynamic regret. With fixed-length preview, the algorithms proposed in (Li et al., 2019; Yu et al., 2020) will yield linear regret. This clearly shows the hardness of the fixed-length preview setting. In this work, we show that sub-linear dynamic regret is achievable with the fixed-length preview under an additional stability assumption that is standard in RHC literature.

3 Algorithms and Regret Performance Guarantees

In this section, we present two different online learning RHC algorithms, namely Certainty Equivalence RHC (CE-RHC) algorithm and Optimistic RHC (O-RHC) algorithm. Both algorithms operate in two phases: (i)
exploration phase, and (ii) control phase. In the exploration phase, both algorithms follow a pure exploration strategy to estimate the unknown system parameter. In the control phase, the algorithms employ an RHC policy with the parameter estimated using the observation from the exploration phase. The algorithms, however, differ in the parameter estimation approach. The CE-RHC algorithm uses a certainty equivalence approach which treats the least squares parameter estimate as the true parameter and employs an RHC policy based on this parameter. The O-RHC algorithm selects an optimistic parameter from the confidence region around the least squares parameter estimate and employs an RHC policy based on this optimistic parameter.

Both the CE-RHC algorithm and the O-RHC algorithm provide the same regret guarantees, but under two different assumptions. The CE-RHC algorithm is computationally more tractable than the O-RHC algorithm, but requires a stronger assumption for the regret guarantee. The O-RHC algorithm eliminates the need for this assumption, but at the expense of additional computational complexity. The algorithms are presented in Algorithm 2 and Algorithm 3.

### 3.1 Exploration Phase

The goal of the exploration phase is to explore the system and collect the observations to estimate 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. Both algorithms use a periodic sequence of control actions which almost surely guarantees the persistence of excitation condition \( \text{Moore, 1983} \) necessary for estimation.

Our approach for estimating the system parameter differs from the prior online learning works with partial observation of the state \( \text{Mania et al., 2019; Oymak and Ozay, 2019; Simchowitz et al., 2020; Oymak and Ozay, 2019} \) only consider the estimation of the system up to a similarity transformation whereas \( \text{Mania et al., 2019; Oymak and Ozay, 2019; Simchowitz et al., 2020} \) consider the control of the measured output. In contrast, we consider the control of the state with the cost as a function of the true state instead of the observed output. Therefore, our approach requires the direct estimate of the system matrices with a high confidence.

Let \( e_i \in \mathbb{R}^m \) be the unit vector in the \( i \)th dimension of \( \mathbb{R}^m \). For \( t \in [1 : T_0] \), the algorithms select the control action \( u_t \) as

\[
  u_t = \begin{cases} 
  (n+1)m e_j & j \in [1, m], \text{if } j = f(t) \\ 
  0 & \text{otherwise}
  \end{cases}
\]

where \( f(t) = \max_{i \leq n} i \) s.t. \((t-1) \mod (n+1) i = 0\), and \( \mod \) denotes the modulo function.

For example, when \( m = 2, n = 2 \), the controller applies \( u_1 = [0, 1]^\top, u_2 = [0, 0]^\top, u_3 = [0, 0]^\top, u_4 = [1, 0]^\top, u_5 = [0, 0]^\top, u_6 = [0, 0]^\top \) for the first six time steps and repeats this sequence.

At the end of the exploration phase, both algorithms perform a least squares estimation using the observation \( y_{1:T_0}, u_{1:T_0} \). More precisely, the algorithms compute \( \hat{\theta}_n = [\hat{\theta}_n], B(\hat{\theta}_n) \) as

\[
  \hat{\theta}_n = \arg \min_{\theta} \sum_{t=1}^{T_0} \| y_{t+1} - [A(\theta), B(\theta)] [y_t, u_t]^\top \|^2_2.
\]

The O-RHC algorithm also computes a confidence region \( \hat{\Theta}(\delta) \) around the least squares estimate \( \hat{\theta}_n \) for a given \( \delta \in (0, 1) \). This confidence region is given by

\[
  \hat{\Theta}(\delta) = \left\{ \theta : \| \theta - \hat{\theta}_n \|_F \leq \beta(\delta) \right\},
\]

where

\[
  \beta(\delta) = \frac{n(n+1)c_{\epsilon}}{c_p T_0} \sqrt{\log \left( \frac{(n(n+2))(\sqrt{2})^{n+m}}{\delta} \right)},
\]

and \( c_p \) is a system dependent constant (exact expression is given in the proof of Proposition 1). For \( \beta(\delta) \) as defined above, the true parameter \( \theta^* \) lies in the confidence region with a probability greater than \((1 - \delta) \). We formally state this result below.

**Proposition 1.** Let Assumption 1 holds, \( u_{1:T_0} \) is given by \( \mathbb{X} \), \( T_0 > (4n^2c_\epsilon^2/c_p) \log((n(n+2))(\sqrt{2})^{n+m})/\delta \). Then with probability greater than \( 1 - \delta \), \( \theta^* \in \hat{\Theta} \).

We note that the lower bound for \( T_0 \) given in the above proposition does not depend on the horizon \( T \).

### 3.2 Control Phase

In the control phase, both algorithms employ an RHC policy, but with different estimates for the true model.

#### 3.2.1 CE-RHC Algorithm

The CE-RHC algorithm treats the least squares estimate \( \hat{\theta}_n \) as the true parameter and selects control actions according to the standard RHC algorithm. More precisely, at each time \( t \in [T_0+1, T] \), the CE-RHC algorithm takes the control action

\[
  u_{t+c} = RHC(t, \hat{x}_t, c_{t+c+M-1}, \hat{\theta}_n),
\]

where \( \hat{x}_t = A(\hat{\theta}_n) \hat{x}_{t-1} + B(\hat{\theta}_n) u_{t+c-1} \), with the initialization \( \hat{x}_{T_0+1} = y_{T_0+1} \). The CE-RHC algorithm is formally presented in Algorithm 2.
The O-RHC algorithm uses an optimistic approach that simultaneously selects the optimistic parameter from the confidence region \( \tilde{\theta} \) and the optimal control action with respect to this optimistic parameter. More precisely, at each time \( t \in [T_0 + 1 : T] \), the O-RHC algorithm selects the control action \( u_t^{\hat{\pi}_o} \) optimistic parameter \( \tilde{\theta}_t \) as

\[
(u_t^{\hat{\pi}_o}, \tilde{\theta}_t) = \text{O-RHC}(t, \hat{x}_t, c_{t:t+M-1}, \tilde{\theta}_t),
\]

where the O-RHC subroutine is given in Algorithm 3.

Algorithm 3 O-RHC Algorithm
1: With the initialization \( \hat{x}_t = x_t \), compute
\[
(\hat{u}_{t:t+M-1}, \tilde{\theta}_t) = \arg \min_{u_{t:t+M-1}, \theta \in \tilde{\Theta}} \sum_{k=t}^{t+M-1} c_k(\hat{x}_k, \hat{u}_k),
\]
\[
\text{s.t. } \hat{x}_{k+1} = A(\theta)\hat{x}_k + B(\theta)\hat{u}_k
\]
2: Output: \((\hat{u}_t, \tilde{\theta}_t)\)

3.3 Regret Performance Guarantees

We now formally present the regret guarantees of the CE-RHC algorithm (denoted as \( \pi_{ce} \)) and O-RHC algorithm (denoted as \( \pi_o \))

**Theorem 1** (Regret of the CE-RHC Algorithm). Suppose Assumption 7 holds and \( \tilde{\theta}_t \) satisfies the conditions given in Assumption 4. Suppose \( M > (\bar{\pi}/\alpha)^2 + 1 \), \( T_0 = T^{2/3} \) and \( T_0 \) satisfies the conditions in Proposition 7. Then, with probability greater than \((1 - \delta)\),

\[
R_T(\pi_{ce}) \leq \mathcal{O}(T^{2/3})
\]

**Theorem 2** (Regret of the O-RHC Algorithm). Suppose Assumption 7 holds, \( \theta^* \) satisfies Assumption 2, and \( T_0 \) satisfies the conditions in Proposition 7. Assume that \( \bar{\pi}/\alpha < 2 \). Fix \( T_0 = T^{2/3} \). Then there exists \( M, T \) such that, for \( T > T' \), with probability greater than \((1 - \delta)\),

\[
R_T(\pi_o) \leq \mathcal{O}(T^{2/3})
\]

**Remark 1.** We observe that the dynamic regret guarantee is valid only when the preview \( M \) is greater than the threshold given by \((\bar{\pi}/\alpha)^2 + 1 \). Such a lower bound requirement is typical in RHC algorithms; see for example [Grimm et al., 2003]. This is expected because, the control computed from a short preview might be very inaccurate and can potentially lead to instability.

**Remark 2.** The first theorem implies that CE-RHC achieves sub-linear dynamic regret provided the stability assumption is satisfied by the least-squares estimate \( \theta_o \). This is feasible provided a neighborhood of parameters around \( \theta^* \) satisfy the assumption. In contrast, the O-RHC algorithm achieves the same dynamic regret under a significantly milder condition that the stability assumption holds only for the system parameter \( \theta^* \). We note that O-RHC achieves this at the cost of the additional computation to estimate the optimistic model alongside the control input.

4. Regret Analysis

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

\[
J_{T_1:T_2}(u_{T_1:T_2} ; \theta_{T_1:T_2}) = \sum_{t=T_1}^{T_2} c_t(x_t, u_t),
\]
\[
\text{s.t. } x_{t+1} = A(\theta_t)x_t + B(\theta_t)u_t, \quad x_{T_1} = x.
\]
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 $u^*_{T}$ and $u^*_{T}$ 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:

\[
R_T(\pi) = J_{1:T_1}(u^*_{1:T_1}; \theta^*) - J_{1:T_1}(u^*_{1:T_1}; \theta^*) + J_{T_1+1:T}(u^*_{T_1+1:T}; \theta^*) - J_{T_1+1:T}(u^*_{T_1+1:T}; \theta^*) + J_{T_1+1:T}(u^*_{T_1+1:T}; \theta^*) - J_{T_1+1:T}(u^*_{T_1+1:T}; \theta^*)
\]

We will characterize the regret due to each term separately for both policies $\pi_{ce}$ and $\pi_o$. Note that for $\pi_{ce}$, $\theta_t = \theta_{bs}, \forall t \in [T_0 + 1, T]$.

### 4.1 Regret of Term I

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. Since the exploration strategy is identical for both the CE-RHC algorithm and the O-RHC algorithm, regret of term I is also identical for both algorithms.

The key challenge involved here is to show that the system state will 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. From here, it follows that the regret of Term I is $O(T_0)$. We formally state the result below.

**Proposition 2** (Regret of Term I). Suppose Assumption 1 holds. Let Term I be as defined in (12). Then, for $\pi \in \{\pi_{ce}, \pi_o\}$,

\[
\text{Term I} \leq O(T_0)
\]

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

### 4.2 Regret of Term II

The CE-RHC algorithm generates the sequence of control actions $u^*_{T_0+1:T}$ using the least squares parameter estimate $\hat{\theta}_{bs}$. However, this control actions are applied on the true system with parameter $\theta^*$. Term II characterizes the regret due to this estimation error.

To analyze this term, we first show that, contingent on the states being bounded, the states of any two systems driven by the same sequence of control actions differ by a term that is bounded by the norm of the difference of the parameters of the two systems. Recall that, in Proposition 1, we have already proved that $||\theta^* - \hat{\theta}_{bs}|| \leq O(1/\sqrt{T_0})$ with high probability.

We then separately show that states are indeed bounded under CE-RHC algorithm when Assumption 2 holds for $\hat{\theta}_{bs}$. This will also imply that the control actions are bounded. The cost functions being locally Lipschitz and the states and control actions being bounded, the cumulative cost can now be upper bounded by the length of the horizon $T - T_0$. Combining this with the observation made in the above paragraph, we get a net upper bound $O(T/\sqrt{T_0})$. We state this result formally below.

**Proposition 3** (Regret of Term II for CE-RHC). Suppose Assumption 1 holds and $\hat{\theta}_{bs}$ satisfies Assumption 2. Suppose $T_0$ satisfies the conditions in Proposition 1 and $M > (\pi/\alpha)^2 + 1$. Let Term II be as defined in (12) and let $\pi = \pi_{ce}$. Then, with probability greater than $(1 - \delta)$,

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

Here also, if we set $T_0 = T^{2/3}$, the regret of Term II is $O(T^{2/3})$.

The analysis of Term II is more challenging for the O-RHC algorithm because we need to consider the sequence of parameters $\theta_{T_0 + 1:T}$ selected by the algorithm. We overcome this issue by first characterizing an upper bound on the difference of the states of the two systems at a time $t$ by a decaying sum of the parameter difference from $t$ to $T_0 + 1$. Since $||\theta^* - \hat{\theta}_t|| \leq O(1/\sqrt{T_0})$ for all $t$ using Proposition 1, we can now use techniques similar to the proof of Proposition 3 to establish the following bound.

**Proposition 4** (Regret of Term II for O-RHC). Suppose Assumption 1 holds, $\theta^*$ satisfies Assumption 2 and $T_0$ satisfies the conditions in Proposition 1. Assume that $\pi/\alpha < 2$. Let Term II be as defined in (12) and let $\pi = \pi_o$. Then there exists $M, T$ such that, for $T > T$, with probability greater than $(1 - \delta)$

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

Here also, if we set $T_0 = T^{2/3}$, the regret of Term II is $O(T^{2/3})$. 


4.3 Regret of Term III

For the CE-RHC algorithm, we bound Term III by bounding its first term, which is the cumulative cost for the standard RHC controller for a system with parameter $\theta_0$. To bound this term, we use the stability assumption for the estimated parameter $\hat{\theta}_0$. We note that the stability assumption does not directly imply the existence of a Lyapunov-like function. The key part of the proof is in establishing that under Assumption 2, for sufficiently large $M$, the function $V_t(:;\theta_0)$ becomes a Lyapunov-like function. This guarantees exponential convergence for the system resulting in $O(1)$ regret for Term III CE-RHC. We formally state this result below.

**Proposition 5** (Regret of Term III for CE-RHC).
Suppose Assumption 7 holds and $\hat{\theta}_0$ satisfies Assumption 2. Let $M > (\bar{\pi} / \alpha)^2 + 1$. Let Term III be as defined in (12) and let $\pi = \pi_{\text{ce}}$. Then,

$$\text{Term III} \leq O(1)$$

The proof for the O-RHC algorithm is significantly more challenging because we assume that the stability assumption is true only for the true system $\theta^*$, whereas the quantity to be analyzed is the cumulative cost of the RHC controller for a time-varying system. 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 $\theta_t$. The novelty of the proof technique is how the optimistic estimate is used to establish the Lyapunov-like condition. This additional term at every time step leads to the $O(T / \sqrt{T_0})$ overall regret instead of $O(1)$ as in CE-RHC.

**Proposition 6** (Regret of Term III for O-RHC). Suppose Assumption 7 holds, $\theta^*$ satisfies Assumption 2 and $T_0$ satisfies the conditions in Proposition 1. Assume that $\bar{\pi} / \alpha < 2$. Let Term III be as defined in (12) and let $\pi = \pi_{\text{ce}}$. Then, there exists $M, T$ such that, for $T > T$, with probability greater than $(1 - \delta)$

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

Note that, by setting $T_0 = T_2/3$, the regret of Term III becomes $O(T^{2/3})$.

4.4 Proofs of the Main Results

Proof of the main results now immediately follow by using the upperbounds obtained for Term I, II, and III. We state this formally below.

**Proof of Theorem 2.** The proof follows by setting $T_0 = T^{2/3}$ and combining Proposition 2, Proposition 8 and Proposition 5.

**Proof of Theorem 5.** The proof follows by setting $T_0 = T^{2/3}$ and combining Proposition 2, Proposition 4 and Proposition 6.

5 Conclusion

In this work, we presented online learning and control algorithms for receding horizon control of linear dynamical systems. Our work sheds light on methods, conditions and analysis for predictive control of unknown systems with limited preview. We show that by using a stability assumption that is standard in the asymptotic analysis of RHC algorithms, we can guarantee $O(T^{2/3})$ dynamic regret for this setting. In future, we plan to extend this algorithm and analysis to systems with adversarial disturbances.

References

Abbasi-Yadkori, Y., Bartlett, P., and Kanade, V. (2014). Tracking adversarial targets. In *International Conference on Machine Learning*, pages 369–377.

Abbasi-Yadkori, Y. and Szepesvári, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26.

Adetola, V., DeHaan, D., and Guay, M. (2009). Adaptive model predictive control for constrained nonlinear systems. *Systems & Control Letters*, 58(5):320–326.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S., and Singh, K. (2019a). Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119.

Agarwal, N., Hazan, E., and Singh, K. (2019b). Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184.

Anderson, B. D. and Moore, J. B. (2007). *Optimal control: linear quadratic methods*. Courier Corporation.

Angeli, D., Amrit, R., and Rawlings, J. B. (2011). On average performance and stability of economic model predictive control. *IEEE transactions on automatic control*, 57(7):1615–1626.

Angeli, D., Casavola, A., and Tedesco, F. (2016). Theoretical advances on economic model predictive control with time-varying costs. *Annual Reviews in Control*, 41:218–224.
Aswani, A., Gonzalez, H., Sastry, S. S., and Tomlin, C. (2013). Provably safe and robust learning-based model predictive control. *Automatica*, 49(5):1216–1226.

Baca, T., Hert, D., Loianno, G., Saska, M., and Kumar, V. (2018). Model predictive trajectory tracking and collision avoidance for reliable outdoor deployment of unmanned aerial vehicles. pages 6753–6760.

Borrelli, F., Bemporad, A., and Morari, M. (2017). *Predictive control for linear and hybrid systems*. Cambridge University Press.

Bujarbaruah, M., Zhang, X., Tanaskovic, M., and Borrelli, F. (2019). Adaptive mpc under time varying uncertainty: Robust and stochastic. *arXiv preprint arXiv:1909.13473*.

Camacho, E. F. and Alba, C. B. (2013). *Model predictive control*. Springer science & business media.

Cohen, A., Hassidim, A., Koren, T., Lazic, N., Mansour, Y., and Talwar, K. (2018). Online linear quadratic control. *arXiv preprint arXiv:1806.07104*.

Cohen, A., Koren, T., and Mansour, Y. (2019). Learning linear-quadratic regulators efficiently with only √T regret. In *International Conference on Machine Learning*, pages 1300–1309.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. (2018). Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197.

Ferramosca, A., Limon, D., and Camacho, E. F. (2014). Economic mpc for a changing economic criterion for linear systems. *IEEE Transactions on Automatic Control*, 59(10):2657–2667.

Ferramosca, A., Rawlings, J. B., Limón, D., and Camacho, E. F. (2010). Economic mpc for a changing economic criterion. In *49th IEEE Conference on Decision and Control (CDC)*, pages 6131–6136. IEEE.

Fukushima, H., Kim, T.-H., and Sugie, T. (2007). Adaptive model predictive control for a class of constrained linear systems based on the comparison model. *Automatica*, 43(2):301–308.

Goel, G. and Wierman, A. (2019). An online algorithm for smoothed regression and lqr control. *Proceedings of Machine Learning Research*, 89:2504–2513.

Goulart, P., Zhang, X., Kamgarpour, M., Georghiou, A., and Lygeros, J. (2016). Robust optimal control with adjustable uncertainty sets. *Automatica*, 75.

Goulart, P. J., Kerrigan, E. C., and Maciejowski, J. M. (2006). Optimization over state feedback policies for robust control with constraints. *Automatica*, 42(4):523–533.

Green, M. and Moore, J. B. (1986). Persistence of excitation in linear systems. *Systems & Control Letters*, 7(5):351–360.

Grimm, G., Messina, M. J., Tuna, S. E., and Teel, A. R. (2005). Model predictive control: for want of a local control lyapunov function, all is not lost. *IEEE Transactions on Automatic Control*, 50(5):546–558.

Grüne, L. and Panin, A. (2015). On non-averaged performance of economic mpc with terminal conditions. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 4332–4337. IEEE.

Grüne, L. and Pirkelmann, S. (2017). Closed-loop performance analysis for economic model predictive control of time-varying systems. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5563–5569. IEEE.

Grüne, L. and Pirkelmann, S. (2020). Economic model predictive control for time-varying system: Performance and stability results. *Optimal Control Applications and Methods*, 41(1):42–64.

Grüne, L. and Stieler, M. (2014). Asymptotic stability and transient optimality of economic mpc without terminal conditions. *Journal of Process Control*, 24(8):1187–1196.

Hazen, E. (2019). Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*.

Langson, W., Chryssochoos, I., Raković, S., and Mayne, D. Q. (2004). Robust model predictive control using tubes. *Automatica*, 40(1):125–133.

Lazar, M., De La Peña, D. M., Heemels, W., and Alamo, T. (2008). On input-to-state stability of min–max nonlinear model predictive control. *Systems & Control Letters*, 57(1):39–48.

Lazic, N., Lu, T., Boutilier, C., Ryu, M., Wong, E. J., Roy, B., and Imwalle, G. (2018). Data center cooling using model-predictive control.

Li, Y., Chen, X., and Li, N. (2019). Online optimal control with linear dynamics and predictions: Algorithms and regret analysis. In *Advances in Neural Information Processing Systems*, pages 14887–14899.

Limon, D., Alvarado, I., Alamo, T., and Camacho, E. (2010). Robust tube-based mpc for tracking of constrained linear systems with additive disturbances. *Journal of Process Control*, 20(3):248–260.

Mania, H., Tu, S., and Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. *arXiv preprint arXiv:1902.07826*.

Mayne, D. Q., Rawlings, J. B., Rao, C. V., and Scokaert, P. O. (2000). Constrained model predictive control: Stability and optimality. *Automatica*, 36(6):789–814.
Moore, J. (1983). Persistence of excitation in extended least squares. *IEEE Transactions on Automatic Control*, 28(1):60–68.

Morari, M. and Lee, J. H. (1999). Model predictive control: past, present and future. *Computers & Chemical Engineering*, 23(4-5):667–682.

Oymak, S. and Ozay, N. (2019). Non-asymptotic identification of lti systems from a single trajectory. pages 5655–5661.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.

Rosolia, U. and Borrelli, F. (2017). Learning model predictive control for iterative tasks. a data-driven control framework. *IEEE Transactions on Automatic Control*, 63(7):1883–1896.

Sastry, S. and Bodson, M. (2011). *Adaptive control: stability, convergence and robustness*. Courier Corporation.

Shalev-Shwartz, S. et al. (2011). Online learning and online convex optimiztion. *Foundations and trends in Machine Learning*, 4(2):107–194.

Shi, G., Shi, X., O’Connell, M., Yu, R., Azizzadenesheli, K., Anandkumar, A., Yue, Y., and Chung, S.-J. (2019). Neural lander: Stable drone landing control using learned dynamics. pages 9784–9790.

Simchowitz, M., Singh, K., and Hazan, E. (2020). Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*.

Skogestad, S. and Postlethwaite, I. (2007). *Multivariable feedback control: analysis and design*, volume 2. Citeseer.

Tanaskovic, M., Fagiano, L., and Gligorovski, V. (2019). Adaptive model predictive control for linear time varying mimo systems. *Automatica*, 103:237–245.

Tempo, R., Calafiore, G., and Dabbene, F. (2012). *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer Science & Business Media.

Vazquez, S., Rodriguez, J., Rivera, M., Franquelo, L. G., and Norambuena, M. (2016). Model predictive control for power converters and drives: Advances and trends. *IEEE Transactions on Industrial Electronics*, 64(2):935–947.

Yu, C., Shi, G., Chung, S.-J., Yue, Y., and Wierman, A. (2020). The power of predictions in online control. *arXiv preprint arXiv:2006.07569*.

Yuan, J. and Lamperski, A. G. (2020). Trading-off static and dynamic regret in online least-squares and beyond. In *AAAI*, pages 6712–6719.
A On Assumption [2]

A.1 Proof that LQR Problem Satisfies Assumption [2]

The LQR cost is given by $c_t(x, u) = x^T Q_t x + u^T R_t u$, where $Q_t > 0$, and $R_t > 0$. We see that $c_t(x, u) \geq x^T Q_t x \geq \lambda_{\min}(Q_t) \|x\|^2$, Hence, it follows that Assumption [2](i) is satisfied with $\alpha = \min_{t \in [1, T]} \lambda_{\min}(Q_t)$, and $\sigma(x) = \|x\|^2$. With $\sigma(x) = \|x\|^2$, Assumption [2](iii) is trivially satisfied.

Suppose that $Q_t$ and $R_t$ are bounded. Suppose $\rho(A^*) < 1$, in accordance with Assumption [3]. Then, it is easy to see that the state of the system will exponentially decay with $u_t = 0$ for all $t$. Hence, Assumption [2](ii) is also trivially satisfied for the LQR case with $\sigma(x) = \|x\|^2$.

A.2 Non-Quadratic, Non-Convex Cost Function Example

Let the system $(A, B)$ be such that $\rho(A) < 1$, which is in accordance with Assumption [4]. Let the controller cost be
\[
c(x, u) = \|x\|^a + \|u\|^a,
\]
where $a$ can be any positive number. Given that $\rho(A) < 1$, there exists a $c_p > 0$, and $\gamma < 1$ when $u_t = 0$, such that $\|x_t\| \leq c_p \gamma^{t-1} \|x_1\|$. For this cost function, let $\sigma(x) = \|x\|^a$. Then, when $u_t = 0$,
\[
\sum c(x_t, u_t) \leq \left( \frac{c_p^a}{1 - \gamma^a} \right) \sigma(x_1).
\]

Then, Assumption [2] is satisfied with $\alpha = 1$, $\alpha = \left( \frac{c_p^a}{1 - \gamma^a} \right)$, $\sigma(x) = \|x\|^a$.

B Proof of Proposition [1]

Let $W_t^T := [(u_t)^\top, \ldots, (u_{t+n})^\top]$, where $(u_t, \ldots, u_{t+n}, \ldots)$ is a sequence of control inputs. We first prove the following lemmas.

**Lemma 1.** Let $u_t$ be as defined in [5], $N = (n + 1)sm$, $s \in \mathbb{N}$. Then $\sum_{k=1}^N W_k W_k^T \geq (n + 1)m = N$.

The proof follows trivially from the observation that (i) each $W_k$ is parallel to the unit vector along a dimension and is of magnitude $(n + 1)m$ and (ii) the matrix $[W_{(n+1)m+1}, W_{(n+1)m+2}, \ldots, W_{(n+1)(s+1)m}]$ is a full rank square matrix for all $j \in [0, s - 1]$.

In the next lemma, we show that the system response in the estimation phase satisfies the persistence of excitation condition (Moore, 1983).

**Lemma 2.** Suppose Assumption [7] holds and $T_0 = (n + 1)(sm) + n$, $s \in \mathbb{N}$. Then for the estimation phase
\[
\sum_{k=t}^{T_0} \begin{bmatrix} x_k^T \\ u_k^T \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \geq c_p T_0,
\]
where $c_p$ is a system dependent constant.

**Proof.** To prove our result we use an argument similar to the one used in (Moore, 1983, Lemma 3.1.). Consider an imaginary output $y_k$ of the linear time invariant system given by the dynamics $x_{k+1} = A^* x_k + B^* u_k$:
\[
y_k = \zeta^T [x_k^T \\ u_k^T]^T,
\]
where $\zeta$ is arbitrary. Let $\zeta^T = [\zeta_1^T \\ \zeta_0^T]$. The corresponding transfer function of the system (with McMillan degree $\leq n$) for the output in (14) is of the form
\[
H(z) = \frac{C_0 + C_1 z^{-1} + \ldots + C_n z^{-n}}{d_0 + d_1 z^{-1} + \ldots + d_n z^{-n}}, d_0 = 1,
\]

Deepan Muthirayan, Jianjun Yuan, Dileep Kalathil, Pramod P. Khargonekar
where the coefficients \([d_0, d_1, \ldots, d_n]\) correspond to the characteristic polynomial. The terms \(C_0, C_1, \ldots, C_n\) have a specific form given by (Green and Moore 1986)

\[
C_j = \sum_{l=0}^j d_{j-l} P_l, P_l = \zeta^T (A^*)^{j-l} B^*, P_0 = \zeta^T.
\]

Let

\[
\beta := [C_n, C_{n-1}, \ldots, C_0]^T, \varepsilon := [d_n, \ldots, d_0]^T.
\]

Then, from (15), it follows that

\[
Q = \begin{bmatrix}
\varepsilon, d_n, \ldots, d_0
\end{bmatrix}^T = \zeta^T \quad \forall \theta.
\]

We can rewrite \(\beta^T\) as

\[
\beta^T = \zeta^T G_\theta, G_\theta = \begin{bmatrix}
\theta & 0 \\
H_\theta & I
\end{bmatrix}
\]

where \(Q_\theta = [q_n, q_{n-1}, \ldots, q_1], q_j = \frac{1}{\theta} \sum_{i=1}^j d_{j-i} (A^*)^{j-i} B^*\), \(H_\theta = [d_n I, d_{n-1} I, \ldots, d_1 I], I = I_m\), the identity matrix of size \(m\). Since the pair \((A^*, B^*)\) is controllable, \(G_\theta\) is a full row rank matrix. Consider \(Q_\theta\). The term with the highest power of \(A^*\) in \(q_j\) is \((A^*)^{j-1} B^*\) and its coefficient is \(d_0 = 1\) for all \(j\). Thus, \(\text{span}\{[B^*, A^* B^*, (A^*)^2 B^*, \ldots, (A^*)^{n-1} B^*]\} = \text{span}\{[q_n, q_{n-1}, \ldots, q_1]\}. \) Thus, \(G_\theta\) is a full row rank matrix. Hence, for a non-zero \(\zeta\) there exists a non-zero element in \(\beta\) and \(\varepsilon\) is non-zero because \(d_0 = 1\).

Let \(q = s(n+1)m\). Then, \(p = T_0 = q + n\). We know from Lemma 1 that

\[
\sum_{k=1}^q W_k W_k^T \geq q > 0,
\]

where \(X \geq c_p \iff v^T X v \geq c_p \forall v \text{ s.t. } \|v\|_2 = 1\).

Now, it follows from (17) that

\[
\|\beta^T W_k\|_2^2 \leq \|\zeta^T \begin{bmatrix}
x_k \\
u_k
\end{bmatrix}, \ldots, \begin{bmatrix}
x_{k+n} \\
u_{k+n}
\end{bmatrix}\|_2^2 \|\varepsilon\|_2^2.
\]

Summing from \(k = 1\) to \(k = q\) on both sides we get that

\[
\beta^T \left( \sum_{k=1}^q W_k W_k^T \right) \beta \leq \|\varepsilon\|_2^2 \zeta^T \sum_{k=1}^q \begin{bmatrix}
x_k \\
u_k
\end{bmatrix}, \ldots, \begin{bmatrix}
x_{k+n} \\
u_{k+n}
\end{bmatrix} \begin{bmatrix}
x_k \\
u_k
\end{bmatrix}^T \zeta \leq (n+1) \|\varepsilon\|_2^2 \zeta^T \sum_{k=1}^{q+n} \begin{bmatrix}
x_k \\
u_k
\end{bmatrix} \begin{bmatrix}
x_k^T \\
u_k^T
\end{bmatrix} \zeta.
\]

Then, using (19) in the previous equation we get that

\[
\zeta^T \sum_{k=1}^{T_0} \begin{bmatrix}
x_k \\
u_k
\end{bmatrix} \begin{bmatrix}
x_k^T \\
u_k^T
\end{bmatrix} \zeta \geq \frac{q \beta^T \beta}{(n+1) \|\varepsilon\|_2^2}.
\]

Since the elements of \(\varepsilon\) are the coefficients of the minimal polynomial of \(A^*\), each element of \(\varepsilon\) is a polynomial function of the eigenvalues of the system matrix \(A^*\). Denote the roots (or the eigenvalues) of \(A^*\) by \(r_1, \ldots, r_n\). Let \(r = \max\{|r_1|, |r_2|, |r_3|, \ldots, |r_n|\}\). It is clear that \(r\) is finite because \(\theta \in \Theta\), and \(\Theta\) is compact. Then, using Vieta’s formula we get that

\[
(-1)^k d_{n-k} = \sum_{1 \leq i_1 \leq i_2 \leq \ldots \leq i_{n-k} \leq n} \prod_{j=1}^{n-k} r_{i_j}, \forall k < n.
\]
Applying binomial theorem, we get that \(|d_k| \leq (1 + r)^n\) and \(\|\epsilon\|^2 \leq n(1 + r)^{2n} + 1\). Using this relation in (20) we get that

\[
\zeta \sum_{k=1}^{T_0} \left[ x_k \right] \left[ x_k^T, u_k \right] \zeta \geq \frac{q(\beta^\top \beta)}{((n + 1)(n(1 + r)^{2n} + 1))}.
\]

Let \(\lambda_{\text{min}}(G_0G_0^\top) = c_g\). Since \((A^*, B^*)\) is a controllable pair, \(c_g > 0\). Applying this fact to the previous equation we get that

\[
\sum_{k=1}^{T_0} \left[ x_k \right] \left[ x_k^T, u_k \right] \geq \frac{qc_g}{((n + 1)(n(1 + r)^{2n} + 1))} \geq c_pT_0,
\]

where \(c_p = c_g/(2(n + 1)(n(1 + r)^{2n} + 1))\).

We will use the following result from Abbasi-Yadkori and Szepesvári (2011).

**Lemma 3** (Abbasi-Yadkori and Szepesvári 2011, Theorem 16). Let \(\{\mathcal{F}_i\}\) be a filtration. Let \(\{m_i\}\) be an \(\mathbb{R}^d\) valued stochastic process adapted to the filtration \(\{\mathcal{F}_i\}\), \(\{\eta_i\}\) be a real-valued martingale difference sequence adapted to \(\{\mathcal{F}_i\}\) and is conditionally sub-Gaussian with constant \(R\). Consider the martingale stochastic process

\[
S_t = \sum_{k=1}^{t} \eta_k m_{k-1}.
\]

Consider the matrix valued process

\[
\nabla_t = \tilde{V} + \tilde{V}_t, \quad \tilde{V}_t = \sum_{k=1}^{t} m_{k-1} m_{k-1}^\top.
\]

Then, with probability \(1 - \delta\), \(\delta > 0\), we get that

\[
\forall t > 0, \|S_t\|^2_{V_t^{-1}} \leq 2R^2 \log \left( \frac{\text{det}(V_t)^{1/2} \text{det}(\tilde{V})^{-1/2}}{\delta} \right).
\]

Next, we use all the above lemmas to get a bound on the least squares estimation error, which proves Proposition 1.

**Proof of Proposition 1** From the state equation (1), we have \(x_{t+1} = A^* x_t + B^* u_t\). Let \(X = [z_1, z_2, z_3, \ldots, z_{T_0}]^\top\), where \((z_i)^\top = [x_i^\top, u_i^\top]\), \(Y = [x_2, x_3, \ldots, x_{T_0+1}]^\top\). Then, it follows that \(Y = X(\theta^*)^\top\), where \(\theta^* = [A^*, B^*]\).

Let \(X^y = [z_1^y, z_2^y, z_3^y, \ldots, z_{T_0}^y]^\top\), where \((z_i^y)^\top = [y_i^T, u_i^T]\) and \(Y^y = [y_2, y_3, \ldots, y_{T_0+1}]^\top\). Let

\[
\mathcal{E}^y = [\epsilon_2, \epsilon_3, \ldots, \epsilon_{T_0+1}]^\top, \mathcal{E}^T = [\epsilon_1^T 0^T_{m \times 1}, \epsilon_2^T 0^T_{m \times 1}, \ldots, \epsilon_{T_0}^T 0^T_{m \times 1}]^\top.
\]

Then, by definition

\[
Y^y = Y + \mathcal{E}^y, \quad X^y = X + \mathcal{E}.
\]

First, we show that \((X^y)^\top X^y\) is invertible with a high probability.

Let \(V = \sum_{k=1}^{T_0} z_k z_k^\top = X^\top X\). In our case, \(z_{k+1}\) is trivially a vector-valued process adapted to \(\mathcal{F}_k\), and any component of the sequence \(\epsilon_{1:T_0}\) by definition is a martingale difference sequence adapted to \(\mathcal{F}_k\). Hence, Lemma 3 can be applied to each component of \(\epsilon_k\) by setting \(V = V/2\) and recognizing that \(m_{k-1} = z_k/\sqrt{2}\) and \(\eta_k\) can be any of the components of \(\sqrt{2}\epsilon_k\). Then, using union bound after applying Lemma 3 to each component of \(\epsilon_k\), we get with probability at least \(1 - n\delta\)

\[
\|X^\top \mathcal{E}\|_{V^{-1}} \leq \frac{n\epsilon}{2} \sqrt{4 \log \left( \frac{\text{det}(V)^{1/2} \text{det}(\tilde{V})}{\delta} \right)} \leq \frac{n\epsilon}{2} \sqrt{4 \log \left( \frac{(\sqrt{2})^{n+m}}{\delta} \right)}
\]
Online Learning for Receding Horizon Control with Provable Regret Guarantees

\[ \|X^\top E + E^\top X\|_{V^{-1}} \leq n\epsilon_c \sqrt{4\log \left( \frac{(\sqrt{2})^{n+m}}{\delta} \right)}, \]  

(22)

where the additional factor \( n \) follows from the fact that \( \eta_k \) has at most \( n \) non-zero components, and we have used the fact that \( R = \epsilon_c/2 \) and \( V = V/2 \).

Now, for any given \( T_0 \), there exists \( s \) such \( (n+1)s m + n \leq T_0 \leq (n+1)(s+1)m + n \). Then, from Lemma 2 for any unit vector \( v \),

\[ v^\top X^\top X v \geq c_p T_0. \]

For a matrix \( M \), denote \( \|M\|_{V^{-1}} = \sum_j \|V^{-1/2} M_j\|_2 \), where \( M_j \) are columns of \( M \). Then \( \|V^{-1/2}(\cdot)\|_2 \leq \|V^{-1/2}(\cdot)\|_F \leq \|V^{-1}\|_F \). Now, for any unit vector \( v \), we get with probability at least \( 1 - n\delta \)

\[ v^\top (X^\top X)^{y} v \geq \sqrt{c_p T_0} \left( \frac{\epsilon_c}{\sqrt{c_p T_0}} \sqrt{4\log \left( \frac{(\sqrt{2})^{n+m}}{\delta} \right)} \right) \]

(23)

Here, we get (a) by \( (X^\top X)^{y} = (X + E)^\top (X + E) = X^\top X + X^\top E + E^\top X + E^\top E \) and ignoring the term due to \( E^\top E \), (b) by first applying Cauchy-Schwarz inequality and then applying the inequality \( \|V^{-1/2}(\cdot)\|_2 \leq \|V^{-1}\|_F \), (c) by Eq. (22), (d) by using \( v^\top X^\top X v \geq c_p T_0 \), (e) by using the fact that \( T_0 > \frac{4n^2\epsilon^2_c}{c_p \log(\sqrt{2}^{n+m}/\delta)} \), which makes the term inside the braces positive, and then using \( v^\top X^\top X v \geq c_p T_0 \) again. This proves that \( (X^\top X)^{y} \) is invertible with probability at least \( 1 - n\delta \).

Next, we show that \( \hat{\theta}_{ls} \) is bounded by a constant that does not grow with \( T_0 \) with a high probability. Because \( (X^\top X)^{y} \) is invertible with probability at least \( 1 - n\delta \), under the same event the least squares \( \hat{\theta}_{ls} \) solution satisfies

\[ Y^y = X^y \hat{\theta}_{ls}^\top, \text{ i.e., } \hat{\theta}_{ls} = \left( (X^\top X)^{y} \right)^{-1} (X^\top X)^{y} Y^y \]

i.e. \( \left\| \hat{\theta}_{ls} \right\|_F \leq \left\| (X^\top X)^{y} \right\|_F \left\| \left( (X^\top X)^{y} \right)^{-1} \right\|_F \leq \sqrt{n + m} \left\| (X^\top X)^{y} \right\|_F \leq O \left( \frac{(n+m)^2}{\gamma^2 T_0} \right) = O(1). \)

Here, we get (f) by Cauchy-Schwarz inequality, (g) by the standard norm inequality for square matrices and the fact that \( (X^\top X)^{y} \) is of dimension \( n + m \), (h) by the fact that the eigenvalues of the inverse of a real symmetric matrix is the inverse of the eigenvalues and the singular values of positive definite matrices are the eigenvalues, (i) by Eq. (23), and (j) by the fact that each element in \( (X^\top X)^{y} \) is a summation of \( T_0 \) terms each of which is bounded by a constant that does not grow with \( T_0 \) and that there are \( (n + m)n \) terms in \( (X^\top X)^{y} \), where the boundedness follows from the fact that \( z_i \) in the estimation phase is bounded by a constant that does not grow.
with $T_0$ (since $u_t$ given $[3]$ is such that $\|u_t\| \leq (n + 1)m$ and the spectral radius $\rho(A^*) < 1$), and that the noise is also bounded by a constant.

Next, we bound the error in the least squares estimate. Let $X(j, l)$ denote the $l$th component of $j$th row of a matrix $X$. Then $(\mathcal{E}(\theta_{ls})^\top)(j, l) = [\epsilon_j^\top_0 m \times 1] \delta_{ls}^\top(:, l)$. From the definitions of $Y^y, Y, X$, we get with probability at least $1 - n\delta$

$$Y^y = Y + \mathcal{E}^y = X(\theta^*)^\top + \mathcal{E}^y,$$

i.e., $X(\hat{\theta}_{ls}^\top - (\theta^*)^\top) = \mathcal{E}^y - \mathcal{E}\hat{\theta}_{ls}^\top,$

i.e., $(\hat{\theta}_{ls}^\top - (\theta^*)^\top)^{(k)} = (X^\top X)^{-1} X^\top \left(\mathcal{E}^y - \mathcal{E}\hat{\theta}_{ls}^\top\right),$

i.e., $\|\hat{\theta}_{ls}^\top - (\theta^*)^\top\|_F \leq \left\|X^\top X\right\|_{1/2} \left\|X^\top \left(\mathcal{E}^y - \mathcal{E}\hat{\theta}_{ls}^\top\right)\right\|_{V^{-1}} = \frac{1}{\sqrt{\lambda_{\text{min}}(V)}} \left\|X^\top \left(\mathcal{E}^y - \mathcal{E}\hat{\theta}_{ls}^\top\right)\right\|_{V^{-1}}$

$$\leq \frac{1}{\sqrt{\lambda_{\text{min}}(V)}} \left(\|X^\top \mathcal{E}^y\|_{V^{-1}} + \left\|X^\top \left(\mathcal{E}\hat{\theta}_{ls}^\top\right)\right\|_{V^{-1}}\right)$$

$$\leq \frac{1}{\sqrt{\lambda_{\text{min}}(V)}} \left(\|X^\top \mathcal{E}^y\|_{V^{-1}} + \sum_{l=1}^n \left\|X^\top \left(\mathcal{E}\hat{\theta}_{ls}^\top\right)\right\|_{V^{-1}}\right)$$

$$\leq \frac{1}{\sqrt{\lambda_{\text{min}}(V)}} \left(\|X^\top \mathcal{E}^y\|_{V^{-1}} + \sum_{r=1}^n \sum_{l=1}^n \left\|X^\top \left(\mathcal{E}(r, \cdot)\hat{\theta}_{ls}^\top\right)\right\|_{V^{-1}}\right)$$

$$\leq \frac{1}{\sqrt{\lambda_{\text{min}}(V)}} \left(\sum_{r=1}^n \|X^\top \mathcal{E}^y(r, \cdot)\|_{V^{-1}} + \sum_{r=1}^n \sum_{l=1}^n \left\|X^\top \mathcal{E}(r, \cdot)\hat{\theta}_{ls}^\top\right\|_{V^{-1}}\right)$$

Here, we get (k) by the fact that $X^\top X > 0$, (l) by $\|V^{-1/2}(\cdot)\|_F \leq \|\cdot\|_{V^{-1}}$, (m) by applying triangle inequality, (n) by applying triangle inequality and the fact that $\hat{\theta}_{ls}^\top$ has $n$ columns, (o) by applying triangle inequality again and the fact that only the first $n$ columns of $\mathcal{E}$ are non-zero, and (p) by applying triangle inequality again.

Then, applying Lemma $[3]$ on each individual term within the braces (there are $n(n + 1)$ of them), and using the fact that $\hat{\theta}_{ls}$ is $O(1)$ with probability greater than $1 - n\delta$, and using union bound, we get with probability at least $1 - n(n + 2)\delta$,

$$\|\hat{\theta}_{ls}^\top - (\theta^*)^\top\|_F \leq \frac{n(n + 1)\epsilon_c}{2\sqrt{\lambda_{\text{min}}(V)}} 4\log \left(\frac{\sqrt{2}^{n+m}}{\delta}\right) \leq \frac{n(n + 1)\epsilon_c}{2\sqrt{c_P T_0}} 4\log \left(\frac{\sqrt{2}^{n+m}}{\delta}\right).$$

Thus, it follows that $\theta^* \in \hat{\Theta}$ with probability greater than $1 - \delta$.

\[\square\]

C Proof of the Results in Section [4]

C.1 Proof of Proposition [2]

Proof. In the estimation phase, the control input given by $[3]$ is clearly bounded. Also by Assumption $[1] (ii)$, the system is stable. This then implies that $x_t^\tau$ is bounded for all $t \in [1 : T_0]$. Let $b = \max_{t \in [1 : T_0]} \|x_t^\tau\|$. We note that $b$ is a constant that does not increase with $T_0$. This follows from the fact that $\|u_t^\tau\| \leq (n + 1)m$ throughout the estimation phase and $\rho(A^*) < 1$. Let $L_0$ be the uniform Lipschitz constant for all $c_t$ over the closed and bounded set $\{x, u : \|x\| \leq b, \|u\| \leq (n + 1)m\}$. It then follows that $c_t(x_t^\tau, u_t^\tau) \leq L_c(\|x_t^\tau\| + \|u_t^\tau\|) + c_t(0, 0) \leq L_c(b + (n + 1)m) + c_t(0, 0)$. Hence, $c_t \leq O(1)$ for all $t \in [1 : T_0]$. Hence, summing over all $t \in [1 : T_0]$, we get $\sum_{t=1}^{T_0} c_t(x_t^\tau, u_t^\tau) \leq O(T_0)$. Thus,

$$J_{1:T_0}(u_{1:T_0}^\tau, \theta^*) - J_{1:T_0}(u_{1:T_0}^\tau, \theta^*) \leq O(T_0)$$

\[\square\]
C.2 Proof of Proposition 4

Proposition 3 and Proposition 4 characterizes the bound on Term II. Here, we only give the proof for Proposition 4 (for the O-RHC algorithm). The proof for Proposition 3 (for the CE-RHC algorithm) is implied by this.

Proof. Consider the system evolution \( \hat{x}_{t+1} = A^*\hat{x}_t + B^*u^\pi_t \), \( \hat{x}_{T+1} = \hat{x}_{T+1} \). Let

\[
\hat{x}^\delta_t = \sum_{j=0+2}^t (A^*)^{t-j}(\delta \theta_{j-1})\hat{z}_{j-1}, \text{ where } \delta \theta_j = (\theta^* - \hat{\theta}_j), \hat{z}_j = [\hat{x}^\top_j, (u^\pi_t)^\top], \hat{x}^\delta_{T+1} = 0. \tag{24}
\]

We will show that \( \hat{x}_t = \hat{x}_t + \hat{x}^\delta_t \) for all \( t \geq T+1 \) by mathematical induction. This trivially holds for \( k = T+1 \). Now, let \( \hat{x}_k = \hat{x}_k + \hat{x}^\delta_k \) be true for some \( k \in [T+1 : T] \). Then

\[
\hat{x}_{k+1} = A^*\hat{x}_k + B^*u^\pi_k = A^*(\hat{x}_k + \hat{x}^\delta_k) + B^*u^\pi_k = \theta^*\hat{z}_k + A^*\hat{x}^\delta_k = \hat{\theta}_k\hat{z}_k + \delta \theta_k\hat{z}_k + A^*\hat{x}^\delta_k = \hat{x}_{k+1} + \hat{x}^\delta_{k+1}.
\tag{25}
\]

This completes the induction argument.

Similarly, we will show that \( x^\pi_{t+1} = \hat{x}_t - (A^*)^{t-T+1}\epsilon_{T+1} \) for all for all \( t \geq T+1 \) by mathematical induction. This holds at \( t = T+1 \), since \( y_{T+1} = x_{T+1} + \epsilon_{T+1} \). Let \( x^\pi_{T+1} = \hat{x}_T - (A^*)^{t-T+1}\epsilon_{T+1} \) hold for some \( k \geq T+1 \). Then

\[
x^\pi_{k+1} = A^*x^\pi_k + B^*u^\pi_{k+1} = A^*(\hat{x}_k - (A^*)^{k-T+1}\epsilon_{T+1}) + B^*u^\pi_k = A^*\hat{x}_k + B^*u^\pi_k - (A^*)^{k+1-T+1}\epsilon_{T+1}. \tag{26}
\]

This completes the induction argument.

Using Equations (24) - (26) we get

\[
\sum_{t=T+1}^T c_t(x^\pi_t, u^\pi_t) = a \sum_{t=T+1}^T c_t(\hat{x}_t - (A^*)^{t-T+1}\epsilon_{T+1}, u^\pi_t) = b \sum_{t=T+1}^T c_t(\hat{x}_t + \hat{x}^\delta_t - (A^*)^{t-T+1}\epsilon_{T+1}, u^\pi_t). \tag{27}
\]

Here, we get (a) by writing \( x^\pi_t = \hat{x}_t - (A^*)^{t-T+1}\epsilon_{T+1} \) and (b) by writing \( \hat{x}_t = \hat{x}_t + \hat{x}^\delta_t \).

According to Proposition 4, \( \theta^* \in \hat{\Theta} \), with high probability. We now claim that when \( \theta^* \in \hat{\Theta} \), under the O-RHC algorithm, \( \max_{t \in [T+1 : T]} \|x_t\| \) is bounded and the bound does not depend on \( T \). The proof for this claim is given as apart of the proof of Proposition 3.

The boundedness of \( \hat{x}_t \) implies the boundedness of \( u^\pi_t \) since \( u^\pi_t \) is the solution of O-RHC Algorithm, whose solutions are continuous in \( \hat{x}_t \) (for a proof, see Rockafellar and Wets, 2009, Theorem 1.17). Then, it follows that \( x^\pi_t \) is bounded since \( \rho(A^*) < 1 \) and \( u^\pi_t \) is bounded. The term \( (A^*)^{t-T+1}\epsilon_{T+1} \) is also bounded, since \( \rho(A^*) < 1 \) and \( \epsilon_{T+1} \) is bounded.

For convenience, denote the bound on \( \|\hat{x}_t\|, \|x^\pi_t\| \) and \( \|(A^*)^{t-T+1}\epsilon_{T+1}\| \) as \( b \) and the bound on \( u^\pi_t \) as \( d \). Note that \( b \) and \( d \) are constants that do not change with the horizon length \( T \). Let \( \alpha_0 \) be the local Lipschitz constant for all \( c_t \)s. Then, under the event \( \theta^* \in \hat{\Theta} \), using Eq. (27) we get

\[
\sum_{t=T+1}^T c_t(x^\pi_t, u^\pi_t) \leq \sum_{t=T+1}^T c_t(\hat{x}_t, u^\pi_t) + \alpha_0 \sum_{t=T+1}^T (\|\hat{x}_t\| + \|(A^*)^{t-T+1}\epsilon_{T+1}\|). \tag{28}
\]

Since \( \rho(A^*) < 1 \), there exist \( c_p > 0 \) and \( \lambda_\rho < 1 \) such that \( \|A^k\| \leq c_p\lambda^k_\rho \). This implies that

\[
\sum_{t=T+1}^T \|\hat{x}_t\| \leq c_p \sum_{t=T+1}^T \|\hat{x}_t\| \leq \sum_{t=T+1}^T \sum_{j=T+1}^t (A^*)^{t-j}(\delta \theta_{j-1})\hat{z}_{j-1} \leq \sum_{t=T+1}^T \sum_{j=T+1}^t (A^*)^{t-j}(\delta \theta_{j-1})\hat{z}_{j-1} \leq (b + d) \sum_{t=T+1}^T \sum_{j=T+1}^t (A^*)^{t-j} \|\delta \theta_{j-1}\| \|\hat{z}_{j-1}\| \leq (b + d) \sum_{t=T+1}^T \sum_{j=T+1}^t (A^*)^{t-j} \|\delta \theta_{j-1}\| \|\hat{z}_{j-1}\|.
\]
We will first show that $J$ to the CE-RHC algorithm, does not affect $x$. Proof.

C.3 Proofs of Proposition 5

Here, we get $(c)$ by definition of $\tilde{x}_t^\delta$, $(d)$ by the bound on $\dot{x}_t$ and $u_t^\pi_o$, $(e)$ by using the fact that $\|\delta \theta_{j-1}\| \leq 2\beta(\delta)$ by Proposition 1 $(f)$ by using the fact that $\|(A^*)^k\| \leq c_p A^\rho_k$, and $(b)$ by using the fact that $\beta(\delta) = O\left(1/\sqrt{T_0}\right)$.

Now,

$$\sum_{t=T_0+1}^{T} \| (A^*)^{t-T_0-1} \epsilon_T \| \leq \epsilon_c \sum_{t=T_0+1}^{T} c_p \lambda_{t-T_0-1} \leq \frac{\epsilon_c c_p}{1 - \lambda_p} = O(1).$$

Using (29) and (30) in (28), we get, with probability greater than $1 - \delta$,

$$\sum_{t=T_0+1}^{T} c_t(x_t^\pi_o, u_t^\pi_o) - \sum_{t=T_0+1}^{T} c_t(\hat{x}_t, u_t^\pi_o) \leq O\left(\frac{T}{\sqrt{T_0}}\right).$$

This implies that, with probability greater than $1 - \delta$,

$$J_{T_0+1:T}(u_{T_0+1:T}^\pi_o; \theta^*) - J_{T_0+1:T}(u_{T_0+1:T}^\pi_o; \hat{\theta}_{T_0+1:T}) \leq O\left(\frac{T}{\sqrt{T_0}}\right).$$

C.3 Proofs of Proposition 5

Proof. Given the system with parameter $\theta = [A(\theta), B(\theta)]$, initial time $t$, initial state $x_t$, and control sequence $u_{t:T}$, consider the system evolution $x_{t+k+i} = A(\theta)x_{t+k+i-1} + B(\theta)u_{t+k+i}$, for $\tau \in [t, T]$ with $x_t = x$. Let $\phi_t(k, x, u_{t:T}, \theta)$ denotes $x_{t+k}$ (the state at time $t+k$) for any $k \geq 0$. We will also denote this as $\phi_t(k, x, u_{t:t+k}, \theta)$ since $u_{t:t+k}$ does not affect $x_{t+k}$.

We will first show that $J_{T_0+1:T}(u_{T_0+1:T}^\pi_o; \hat{\theta}_{bs})$ is $O(1)$ under the assumption stated.

For any $t \in [T_0 + 1, T]$, consider the system evolution $\hat{x}_{t+1} = A(\hat{\theta}_{bs})\hat{x}_t + B(\hat{\theta}_{bs})u_{t}^\pi_o$, with the initial state $\hat{x}_{T_0+1} = y_{T_0+1}$. Let

$$\hat{u}_{0:M-1}^t = \arg \min_{u_{t:t+M-1}} \sum_{k=t}^{t+M-1} c_k(x_k, u_k), \text{ s.t. } x_{k+1} = A(\hat{\theta}_{bs})x_k + B(\hat{\theta}_{bs})u_k, \ x_t = \hat{x}_t.$$

Recall from Eq. (4) that the optimal value of the above problem is $V_t(\hat{x}_t; \hat{\theta}_{bs})$. Also, please note that according to the CE-RHC algorithm, $u_t^\pi_o = \hat{u}_t$. Now,

$$V_t(\hat{x}_t; \hat{\theta}_{bs}) = \sum_{k=0}^{M-1} c_t+k(\phi_t(k, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_k) + \sum_{k=0}^{M-2} c_t+k+1(\phi_t(k+1, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{k+1}),$$

$$= c_t(\hat{x}_t, u^\pi_o) + \sum_{k=0}^{j-2} c_t+k+1(\phi_t(k, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{k+1})$$

$$= c_t(\hat{x}_t, u^\pi_o) + \sum_{k=0}^{j-2} c_t+k+1(\phi_t(k, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{k+1}) + \sum_{k=j-1}^{M-2} c_t+k+1(\phi_t(k, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{k+1})$$

$$= c_t(\hat{x}_t, u^\pi_o) + \sum_{k=0}^{j-2} c_t+k+1(\phi_t(k, \hat{x}_t, \hat{u}^\pi_o_{0:k}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{k+1}) + \sum_{k=0}^{M-j-1} c_t+j+k(\phi_t(j + k - 1, \hat{x}_t, \hat{u}^\pi_o_{0:j+k-1}; \hat{\theta}_{bs}), \hat{u}^\pi_o_{j+k}).$$
\[ (c) \quad \sum_{k=0}^{M} c_{t+k+1}(\phi_t + 1(k, \hat{x}_{t+1}, \hat{u}_{t+1}; \hat{\theta}_b)), \hat{u}_{t+k+1} + \sum_{k=0}^{M-j-1} c_{t+j+k}(\phi_t + j(k, \hat{x}_{t+1}, \hat{u}_{t+k}; \hat{\theta}_b), \hat{u}_{j+k}) \]

\[ (31) \]

Here, we get (a) by using the fact that \( \hat{u}_1 = u_{t+1} \), (b) by using the fact that \( \hat{x}_{t+1} = \phi_t(1, \hat{x}_t, \hat{u}_{0,0}; \hat{\theta}_b) \) and considering the summation from \( t+1 \) with initialization \( \hat{x}_{t+1} \), and (c) by denoting \( \hat{x}_{t+1} = \phi_t(j - 1, \hat{x}_t, \hat{u}_{j-1}; \hat{\theta}_b) \) and considering the sequence from time step \( t+j \).

Similarly,

\[ V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_b) = \sum_{j=0}^{M-1} c_{t+1+k}(\phi_t + 1(k, \hat{x}_{t+1}, \hat{u}_{t+1}; \hat{\theta}_b), \hat{u}_{t+k}) + \sum_{k=0}^{M-j-1} c_{t+j+k}(\phi_t + j(k, \hat{x}_{t+1}, \hat{u}_{t+k}; \hat{\theta}_b), \hat{u}_{j+k}) \]

\[ (32) \]

Here, we get (d) by changing the optimal sequence \( \hat{u}_{t+1}^{t+1} \) to the control sequence \( \hat{u}_{t+1}^{t+1} \) in the first \( j-1 \) steps starting from time \( t+1 \) and finding the minimizing sequence for the remaining steps. For this, we denote \( \hat{x}_{t+1} = \phi_{t+1}(j - 1, \hat{x}_t, \hat{u}_{j-1}; \hat{\theta}) \) as the initial state for the remaining summation. Please note that \( \hat{x}_{t+1} \) we introduced in (c) and (d) above are indeed identical. We get (e) from the definition \( V_{t+j} \) and (f) from the premise of the proposition that \( \hat{\theta}_b \) satisfies the conditions given in Assumption 2 (in particular, Assumption 2 .(ii))

Now, using (31) and (32),

\[ V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_b) - V_t(\hat{x}_t; \hat{\theta}_b) \leq \alpha \sigma(\hat{x}_{t+1}) - c_t(\hat{x}_t, u_t) \leq \alpha \sigma(\hat{x}_{t+1}) - \alpha \sigma(\hat{x}_t), \]

\[ (33) \]

where the first inequality is obtained by canceling the common terms and using the fact that \( c_t \)s are positive functions, and the second inequality by using Assumption 2 (i)

For any \( j \in [1, M-1] \), using the fact that \( \hat{x}_{t+1} = \phi_{t+1}(j - 1, \hat{x}_t, \hat{u}_{j-1}; \hat{\theta}_b) = \phi_t(j, \hat{x}_t, \hat{u}_{j-1}; \hat{\theta}_b) \), Assumption 2 .(i), and Assumption 2 .(ii), we get

\[ \alpha \sum_{j=1}^{M-1} \sigma(\hat{x}_{t+1}) \leq \sum_{k=0}^{M-1} c_{t+k}(\phi_t(k, \hat{x}_t, \hat{u}_{0,k}; \hat{\theta}_b), \hat{u}_{k}) = V_t(\hat{x}_t; \hat{\theta}_b) \leq \alpha \sigma(\hat{x}_t). \]

Hence, there exists \( j^* \in [1, M-1] \) such that

\[ \sigma(\hat{x}_{t+1}) \leq \frac{\sigma(\hat{x}_t)}{(M-1)}. \]

Using (31) in (33), for \( j = j^* \), we get

\[ V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_b) - V_t(\hat{x}_t; \hat{\theta}_b) \leq \frac{(\sigma(\hat{x}_t) \leq (\sigma(\hat{x}_t))}{(M-1) - 1}) \alpha \sigma(\hat{x}_t) \]

\[ (35) \]
Let $\gamma = \frac{(\alpha/\pi)^2}{4}$. Note that since $M > (\pi/\alpha)^2 + 1$ according to the premise of the proposition, we have $\gamma < 1$. Using this, we get

$$V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_n) - V_t(\hat{x}_t; \hat{\theta}_n) \leq -(1-\gamma)\alpha \sigma(\hat{x}_t) = -(1-\gamma)(\alpha/\pi)\sigma(\hat{x}_t) \leq -(1-\gamma)(\alpha/\pi)V_t(\hat{x}_t; \hat{\theta}_n),$$

where we used Assumption $[2]$, $(ii)$ to get the last inequality. This will yield,

$$V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_n) \leq \gamma V_t(\hat{x}_t; \hat{\theta}_n),$$

where $\gamma = 1 - (1-\gamma)(\alpha/\pi)$. Applying the above inequality repeatedly, we get

$$c_t(\hat{x}_t, u_{T}^{ce}) \leq V_1(\hat{x}_1; \hat{\theta}_n) \leq \gamma V_{t-1}(\hat{x}_{t-1}; \hat{\theta}_n) \leq \ldots \leq \gamma^{t-T_0}V_{T_0+1}(\hat{x}_{T_0+1}; \hat{\theta}_n) \leq \gamma^{t-T_0}\alpha \sigma(\hat{x}_{T_0+1}).$$

Taking summation on both sides,

$$\sum_{t=T_0+1}^{T} c_t(\hat{x}_t, u_{T}^{ce}) \leq \frac{\alpha}{(1-\gamma)}\sigma(\hat{x}_{T_0+1}) = \mathcal{O}(1).$$

This implies that $J_{T_0+1:T}(u_{T_0+1:T}^{ce}; \hat{\theta}_n) = \sum_{t=T_0+1}^{T} c_t(\hat{x}_t, u_{T}^{ce}) = \mathcal{O}(1)$

This will also imply that $J_{T_0+1:T}(u_{T_0+1:T}^{ce}; \hat{\theta}_n) = J_{T_0+1:T}(u_{T_0+1:T}^{*}; \theta^*) \leq \mathcal{O}(1)$, which concludes the proof. \hfill \Box

### C.4 Proof of the Proposition [3]

Proof. Given the system with parameter $\theta = [A(\theta), B(\theta)]$, initial time $t$, initial state $x$, and control sequence $u_{t:T}$, consider the system evolution $x_{t+1} = A(\theta)x_t + B(\theta)u_t$, for $t \in [t, T]$ with $x_t = x$. Let $\phi_t(k, x_{t:k}, \theta)$ denotes $x_{t+k}$ (the state at time $t+k$) for any $k \geq 0$. We will also denote this as $\phi_t(k, x_{t:k-1}, \theta)$ since $u_{t+k:T}$ does not affect $x_{t+k}$.

We will first show that $J_{T_0+1:T}(u_{T_0+1:T}^{*}; \hat{\theta}_n)$ is $\mathcal{O}(T/\sqrt{T})$ under the stated assumptions.

For any $t \in [T_0 + 1, T]$, consider the system evolution $\hat{x}_{t+1} = A(\hat{\theta})\hat{x}_t + B(\hat{\theta})u_{T}^{ce}$, with the initial state $\hat{x}_{T_0+1} = y_{T_0+1}$. Let

$$\hat{x}_{t+1} = \min_{\hat{\theta}_n} A(\hat{\theta})\hat{x}_t + B(\hat{\theta})u_{T}^{ce}, \quad \hat{x}_{T_0+1} = \min_{\hat{\theta}_n} A(\hat{\theta})\hat{x}_{T_0+1} + B(\hat{\theta})u_{T}^{ce}.$$

Let $\hat{x}_{t+1} = \phi_t(j, \hat{x}_t, \hat{u}_{t,j-1}(\hat{\theta}); \theta^*)$, and $\hat{x}_{T_0+1} = \phi_t(j - 1, \hat{x}_{T_0+1}, \hat{u}_{t,j-1}(\hat{\theta}); \theta^*)$ for all $j \in [1, M]$. Also, define

$$\hat{u}_{t,j-1}(\hat{\theta}) = \min_{\hat{\theta}_n} \sum_{k=0}^{M-1} c_{t+j+k}(\phi_{t+j}(k, \hat{x}_{t+j}, \hat{u}_{t,j+k}; \theta^*), \hat{u}_{k}).$$

Now, by definition

$$\hat{x}_{t+1} = A(\hat{\theta})(\hat{x}_t + B(\hat{\theta})u_{T}^{ce}, \quad \hat{x}_{T_0+1} = A(\hat{\theta})\hat{x}_{T_0+1} + B(\hat{\theta})u_{T}^{ce}.$$

Hence, by definition, $\forall \ k, \ k \in [0, M - 1],$

$$\phi_{t+j}(k, \hat{x}_{t+j}, \hat{u}_{t,j+k}(\hat{\theta}); \theta^*) - \phi_{t+j}(k, \hat{x}_{t+j}, \hat{u}_{t,j+k}(\hat{\theta}); \theta^*) = -A(\theta)^k \delta(\theta) \hat{x}_t.$$

Consider $x_{k+1} = A(\theta)x_k + B(\theta)u_{k-1}(\hat{\theta})$, $k \in [t, t + M - 1]$, $x_t = \hat{x}_t$, and $x_{k+1} = A(\hat{\theta})\hat{x}_t + B(\hat{\theta})u_{T}^{ce}(\hat{\theta})$, $\hat{x}_t = \hat{x}_t$. Then, applying the same argument from Eq. (24) to (25) in the proof of Proposition [3] we get that

$$x_k = \hat{x}_k + x_k^{\delta(\theta)}, \quad \text{where} \quad x_k^{\delta(\theta)} = \sum_{j=t+1}^{k} (A(\theta)^{k-j}(\delta(\theta))\hat{x}_{j-1}, \hat{x}_{j}^{\top} = [\hat{x}_{j}^{\top}, u_{j-1}(\hat{\theta})], \quad x_k^{\delta(\theta)} = 0, \quad \delta(\theta) = \theta^* - \hat{\theta}.$$

Deepan Muthirayan, Jianjun Yuan, Dileep Kalathil, Pramod P. Khargonekar
Since the solution to Eq. (36) is continuous in \( \hat{x}_t \) (for proof, see assertion (c) of [Rockafellar and Wets, 2009, Theorem 1.17]), \( \hat{u}_{0:M-1}^t \) and \( \hat{\theta}_t \) are continuous functions of \( \hat{x}_t \). Then, it follows that \( x_{k, \hat{x}_t, \hat{x}_t^{k+1-t}, \hat{x}_t^{k+1-t}} \) and \( x_{k}^{\delta \theta} \) in Eq. (38) are also continuous functions of \( \hat{x}_t \) for all \( k \in [t, t + M - 1] \). Similarly, \( \hat{u}_{0:M-1}^{k+1-t} \) is a continuous function of \( \hat{x}_t \) for all \( k \in [t, t + M - 1] \), because it is a continuous function of \( \hat{x}_t^{k+1-t} \). Therefore \( \phi_{k+1}(l, \hat{x}_t^{k+1-t}, \hat{u}_{0:M-1}^{k+1-t}, \hat{\theta}_t) \) is a continuous function of \( \hat{x}_t \) for all \( l \in [0, M - 1] \), \( k \in [t, t + M - 1] \).

Next, we prove by induction that, under the event \( \theta^* \in \hat{\Theta} \), \( \hat{x}_t \) is bounded by a constant that does not increase with \( T_0 \) and \( T \). Let’s assume that the bound on \( \hat{x}_t \) under the event \( \theta^* \in \hat{\Theta} \) to be the constant \( b \). Then, there exist functions \( \hat{b} \) and \( d \) with \( \text{dom}(\hat{b}) = \text{dom}(d) = \mathbb{R} \) such that, the bound on \( x_{k, \hat{x}_t, \hat{x}_t^{k+1-t}, \hat{x}_t^{k+1-t}, \hat{x}_t^{\delta \theta}} \) and \( \phi_{k+1}(l, \hat{x}_t^{k+1-t}, \hat{u}_{0:M-1}^{k+1-t}, \hat{\theta}_t) \) for all \( k \in [t, t + M - 1] \) is \( \hat{b}(b) \) and the bound on \( \hat{u}_{k+1}^{k+1-t} \) and \( \hat{u}_{k}^{k+1-t} \) for all \( k \in [t, t + M - 1] \) is \( d(b) \). Let \( \alpha_0 \) be the local Lipschitz constant for \( c_{0,8} \) in the compact set \( \{(x, u) : \|x\| \leq \hat{b}(b), \|u\| \leq d(b)\} \).

Recall from (4) that the optimal value of the above problem is \( V_t(\hat{x}_t; \hat{\theta}_t) \). Since \( \rho(A^*) < 1 \), there exist constants \( c_\rho > 0 \) and \( \lambda_\rho < 1 \) such that \( \|A^k\| \leq c_\rho \lambda_\rho^k \). Also, please note that according to the O-RHC algorithm, \( u_t^\infty = \hat{u}_0^k \). Now, under the event \( \theta^* \in \hat{\Theta} \),

\[
V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_{t+1}) = \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}); \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1}))
\]

(a) \[
\leq \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}); \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\theta^*))
\]

(b) \[
\leq \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}); \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\theta^*)) + \sum_{k=0}^{M-j} c_{t+j+k}(\phi_{t+j}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}((\theta^*))
\]

(c) \[
= \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}); \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1}))
\]

(d) \[
\leq \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{V}_{t+1}(\hat{x}_{t+1}; \theta^*) + \alpha_0 \sum_{k=0}^{M-1} \|\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}((\theta^*))
\]

(e) \[
\leq \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{\theta}_{t+1}), \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1}) + \alpha_0 \sum_{k=0}^{M-1} \|\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}((\theta^*)), \hat{u}_{k}^{t+1}((\theta^*))
\]

(f) \[
\leq \sum_{k=0}^{M-1} c_{t+1+k}(\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1})), \hat{\theta}_{t+1}), \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1}) + \alpha_0 \sum_{k=0}^{M-1} \|\phi_{t+1}(k, \hat{x}_{t+1}, \hat{u}_{0:k-1}^{t+1}(\hat{\theta}_{t+1}; \hat{\theta}_{t+1}), \hat{u}_{k}^{t+1}(\hat{\theta}_{t+1}))
\]

Here, we get (a) by the fact that \( \hat{\theta}_t \) is the optimal model at \( t \), (b) by the fact that the sequence of actions \( \hat{a}^{t+1-j}(\hat{\theta}_t) \) for the first \( j - 1 \) steps followed by \( \hat{a}^{t+1}_{0:M-1}(\theta) \), (c) by using Eq. (37) and Lipschitz condition on \( c_{0,8} \), (d) by Eq. (4), (e) by Assumption 2(ii) and (f) by applying Cauchy-Schwarz to the last term. Similarly for \( V_t(\hat{x}_t; \hat{\theta}_t) \),

\[
V_t(\hat{x}_t; \hat{\theta}_t) = \sum_{k=0}^{M-1} c_{t+k}(\phi_{t+k}(k, \hat{x}_{t+k}, \hat{u}_{0:k-1}^{t+k}(\hat{\theta}_{t+k}), \hat{u}_{k}^{t+k}(\hat{\theta}_{t+k}))
\]

\[
= \sum_{k=0}^{M-1} c_{t+k}(\hat{x}_{t+k}, \hat{u}_{k}^{t+k}(\hat{\theta}_{t+k})) = \sum_{k=0}^{M-1} c_{t+k}(x_{t+k} - x_{t+k}^{\delta \theta}, \hat{u}_{k}^{t+k}(\hat{\theta}_{t+k})) \geq \sum_{k=0}^{M-1} c_{t+k}(x_{t+k}, \hat{u}_{k}^{t+k}(\hat{\theta}_{t+k})) - \alpha_0 \sum_{k=0}^{M-1} \|x_{t+k}^{\delta \theta} \|
\]

\[
= \sum_{k=0}^{M-1} c_{t+k}(x_{t+k}, \hat{u}_{k}^{t+k}(\hat{\theta}_{t+k})) - \alpha_0 \sum_{j=t+1}^{t+k} \|A^*(t-k-j)(\delta \theta)\| \hat{z}_{j-1} \|
\]
Then, setting $j$ and then applying Cauchy-Schwarz inequality, (i) by using
Here, we get (g) by applying the local Lipschitz condition for $\phi$
By Assumption 2. (such that $\theta$
$\hat{c}$ is the optimal model at $t+1$ and $\hat{x}_t^0 = \hat{u}_0^0(\hat{\theta}_t)$, and the definition of $\hat{x}_t^1$ and $x_{t+1} = \phi_{t+1}(k, \hat{x}_t^1, \hat{a}_{t+1}^k(\hat{\theta}_t); \theta^*)$

Now taking the difference between $V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_{t+1})$ and $V_t(\hat{x}_t; \hat{\theta}_t)$,

By Assumption 2(i), for all $j \in [1, M - 1]$, $\alpha\sigma(\hat{x}_t^j) \leq c_t(\hat{x}_t, \hat{a}_t^0(\hat{\theta}_t))$. Then, summing over $j$ and adding $c_t(\hat{x}_t, \hat{a}_t^0(\hat{\theta}_t))$ to the right, we get

\[
\sum_{j=1}^{M-1} c_t(\hat{x}_t, \hat{a}_t^0(\hat{\theta}_t)) + \sum_{j=0}^{M-1} c_t(\phi_t(j, \hat{x}_t, \hat{a}_{t+1}^j(\hat{\theta}_t); \theta^*), \hat{a}_j^0(\hat{\theta}_t))
\]

Next, we get (k) by (g), (l) by computing the sum $\alpha_0 \sum_{k=0}^{M-1} \|x^\theta_{t+k}\|$ as in steps (g) to (i), (m) by the fact that $\hat{\theta}_t$ is the optimal model at $t$ and finally (n) by Assumption 2(ii). This implies that, there exists $j^* \in [1, M - 1]$ such that

\[
\sigma(\hat{x}_t^{j^*}) \leq \frac{\alpha_0 \sigma(\hat{x}_t)}{\alpha(M - 1)} + \frac{\alpha_0(\bar{b}(b) + d(b))c_p \|\delta\theta_t\| M}{\alpha(M - 1)(1 - \lambda_p)}.
\]

Then, setting $j = j^*$ in Eq. (41), we get

\[
V_{t+1}(\hat{x}_{t+1}; \hat{\theta}_{t+1}) - V_t(\hat{x}_t; \hat{\theta}_t) \leq \frac{\alpha_0(\bar{b}(b) + d(b))c_p \|\delta\theta_t\| M}{\alpha(M - 1)(1 - \lambda_p)}.
\]
\[
\begin{align*}
&\leq \frac{\omega^2 \sigma(\tilde{x}_t)}{\omega(M-1)} - c_t(\tilde{x}_t, u^\tau_t) + \left( \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M + 1}{1 - \lambda^\alpha} + \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M}{\omega(M-1)(1 - \lambda^\alpha)} \right) \| \delta_{t+1} \\
&\leq \frac{\omega^2 \sigma(\tilde{x}_t)}{\omega(M-1)} - c_t(\tilde{x}_t, u^\tau_t) + \left( \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M + 1}{1 - \lambda^\alpha} + \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M}{\omega(M-1)(1 - \lambda^\alpha)} \right) \| \delta_{t+1} \\
&\leq \left( \frac{\omega^2}{\omega^2(M-1)} - 1 \right) \alpha \sigma(\tilde{x}_t) + \left( \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M + 1}{1 - \lambda^\alpha} + \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M}{\omega(M-1)(1 - \lambda^\alpha)} \right) \| \delta_{t+1} .
\end{align*}
\]

Here, we get (o) by Eq. (43), (p) by the fact that the factors accompanying \( \| \delta_{t+1} \| \) are all constants, (q) by Assumption 2 (i).

Let \( \gamma = \frac{\omega^2}{\omega^2(M-1)} \). Since \( M > (\omega/\omega)^2 + 1 \), we have \( \gamma < 1 \). Let \( g(b, M) = \left( \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M + 1}{1 - \lambda^\alpha} + \frac{\alpha \sigma(\tilde{x}_t) + g(b, M) M}{\omega(M-1)(1 - \lambda^\alpha)} \right) \).

Using these observations, we get

\[
V_{t+1}(\tilde{x}_{t+1}; \tilde{\theta}_{t+1}) - V_t(\tilde{x}_t; \tilde{\theta}_t) \leq - \left( 1 - \gamma \right) \frac{\omega^2}{\omega^2(M-1)} \| \delta_{t+1} \| \leq \leq - \left( 1 - \gamma \right) \frac{\omega^2}{\omega^2(M-1)} V_t(\tilde{x}_t; \tilde{\theta}_t) + g(b, M) \| \delta_{t+1} \| ,
\]

where we used Assumption 2 (ii) to get the last inequality. This yields

\[
V_{t+1}(\tilde{x}_{t+1}; \tilde{\theta}_{t+1}) \leq \gamma V_t(\tilde{x}_t; \tilde{\theta}_t) + g(b, M) \| \delta_{t+1} \| , \quad \gamma = 1 - \gamma \frac{\omega}{\omega^2(M-1)} .
\]

Under the event \( \theta^* \in \tilde{\Theta} \), \( \| \delta_{t+1} \| \leq 2 \beta(\delta) \) for all \( t \geq T_0 + 1 \). Therefore,

\[
V_{t+1}(\tilde{x}_{t+1}; \tilde{\theta}_{t+1}) \leq \gamma V_t(\tilde{x}_t; \tilde{\theta}_t) + \mathcal{O}(\frac{g(b, M)}{\sqrt{T_0}}) . \tag{44}
\]

Now, by Assumption 2 (ii), and that \( \tilde{\theta}_t \) is optimistic, \( V_t(\tilde{x}_t; \tilde{\theta}_t) \leq \omega \sigma(\tilde{x}_t) \), for all \( t \). Also, by Assumption 2 (i), \( V_t(x; \theta) \geq \omega \sigma(x) \) for any \( \theta \). Hence, from Eq. (44), we get

\[
\sigma(\tilde{x}_{t+1}) \leq \gamma \omega \sigma(\tilde{x}_t) + \mathcal{O}(\frac{g(b, M)}{\sqrt{T_0}}) .
\]

Given the expression for \( \gamma \),

\[
\frac{\omega}{\omega^2(M-1)} = \frac{\omega}{\omega} - 1 + \gamma .
\]

Let \( \delta_3 > 0 \) be such that \( \frac{\omega}{\omega^2(M-1)} = 2 - \delta_3 \). Then, there exists a \( M > 0 \) sufficiently large such that \( \gamma = \frac{\omega}{\omega^2(M-1)} = \delta_2 < \delta_3 \). Let \( \delta_3 = \delta_1 - \delta_2 \). Then, \( \frac{\omega}{\omega^2(M-1)} = \frac{\omega}{\omega} - 1 + \gamma = 1 - \delta_3 < 1 \).

Let \( \gamma_1 = 1 - \delta_3 \). Then, for every \( b > 0 \) there exists \( T_0 > 0 \) sufficiently large and a constant \( \tilde{b} \geq b \) such that

\[
\sigma(\tilde{x}_{t+1}) \leq \gamma_1 \sigma(\tilde{x}_t) + \mathcal{O}(\frac{\tilde{b}}{\sqrt{T_0}}) \leq \sigma(\tilde{x}_t) \leq \max_{x: \| x \| \leq \tilde{b}} \sigma(x) \Rightarrow \| \tilde{x}_{t+1} \| \leq \tilde{b} .
\]

Now, we can choose \( b \) to be such that \( b = \| y_{T_0+1} \| . \) This \( b \) is a constant that does not increase with \( T_0 \) because \( \| y_{T_0+1} \| \) is bounded by a constant that does not increase with \( T_0 \). This is because \( \rho(A^*) < 1 \) and \( \| u^\tau_t \| \leq (n+1)m \) in the estimation phase. For this \( b \), we can pick \( M, T_0 \) and \( \tilde{b} \) as above. Since \( \sigma(x) \) is a continuous function, \( b \) is a constant that does not increase with \( T_0 \) and \( T \).

Then, by mathematical induction, it follows that under the event \( \theta^* \in \tilde{\Theta} \), \( \| \tilde{x}_t \| \leq \tilde{b} \) for all \( t \geq T_0 + 1 \), where \( \tilde{b} \) is a constant that does not increase with \( T_0 \) and \( T \).

Consequently, Eq. (43) is true for all \( t \geq T_0 + 1 \), under the event \( \theta^* \in \tilde{\Theta} \). Therefore, applying the inequality in Eq. (43) repeatedly, we get under the event \( \theta^* \in \tilde{\Theta} \)

\[
V_t(\tilde{x}_t; \tilde{\theta}_t) \leq \sum_{k=T_0+1}^{T_0} \gamma^{-k+T_0-1} V_{T_0+1}(\tilde{x}_{T_0+1}; \tilde{\theta}_{T_0+1}) + \mathcal{O}(\frac{1}{\sqrt{T_0}}) ,
\]

where \( \gamma = \frac{\omega}{\omega^2(M-1)} \).
where the last equality follows from the fact that $\gamma < 1$. Using $c_t(\hat{x}_t, u_t^{\pi_{ot}}) \leq V_t(\hat{x}_t; \hat{\theta}_t)$, and summing over all $t \geq T_0 + 1$, we get
\[
\sum_{t=T_0+1}^{T} c_t(\hat{x}_t, u_t^{\pi_{ot}}) \leq \sum_{t=T_0+1}^{T} \gamma^{t-T_0-1} V_{T_0+1}(\hat{x}_{T_0+1}; \hat{\theta}_{T_0+1}) + \sum_{t=T_0+1}^{T} \mathcal{O}(\frac{1}{\sqrt{T_0}})
\]
\[
\leq \mathcal{O}(\gamma V_{T_0+1}(\hat{x}_{T_0+1}; \hat{\theta}_{T_0+1})) + \mathcal{O}(\frac{T}{\sqrt{\gamma T_0}}) = \mathcal{O}\left(\frac{T}{\sqrt{T_0}}\right).
\]

Here, we get (r) by the fact that $\gamma < 1$ and (s) by the fact that $\hat{\theta}_{T_0+1}$ is optimistic and by Assumption 2(ii), both of which imply $V_{T_0+1}(\hat{x}_{T_0+1}; \hat{\theta}_{T_0+1}) \leq \overline{\sigma}(\hat{x}_{T_0+1}) = \mathcal{O}(1)$. This also implies that, with probability greater than $1 - \delta$, $J_{T_0+1:T}(u^{\pi_{ot}_{T_0+1:T}; \hat{\theta}_{T_0+1:T}}) - J_{T_0+1:T}(u^{\pi_{t_0+1:T}; \theta^*}) \leq \mathcal{O}(\frac{T}{\sqrt{T_0}})$, which concludes the proof. \qed