Regret Analysis of Learning-Based MPC with Partially-Unknown Cost Function

Ilgin Dogan, Zuo-Jun Max Shen *Member, IEEE*, and Anil Aswani *Member, IEEE*

***Abstract*— The exploration/exploitation trade-off is an inherent challenge in data-driven adaptive control. Though this trade-off has been studied for multi-armed bandits (MAB's) and reinforcement learning for linear systems; it is less well-studied for learning-based control of nonlinear systems. A significant theoretical challenge in the nonlinear setting is that there is no explicit characterization of an optimal controller for a given set of cost and system parameters. We propose the use of a finite-horizon oracle controller with full knowledge of parameters as a reasonable surrogate to optimal controller. This allows us to develop policies in the context of learning-based MPC and MAB's and conduct a control-theoretic analysis using techniques from MPC- and optimization-theory to show these policies achieve low regret with respect to this finite-horizon oracle. Our simulations exhibit the low regret of our policy on a heating, ventilation, and air-conditioning model with partially-unknown cost function.**

***Index Terms*— Non-myopic Exploitation, Learning-Based Control, Model Predictive Control, Restless Bandits**

I. Introduction

Reinforcement learning (RL) research [1]–[3] focuses on regret analysis for primarily unconstrained, linear systems. On the other hand, adaptive model predictive control (MPC), including learning-based MPC (LBMPC), seeks to ensure constraint satisfaction in the presence of models that are updated as more data becomes available [4]–[7]. The relationship between MPC and RL has not yet been fully explored.

Our paper aims to better connect these two areas. We make two main contributions: First, we discuss how comparing finite-horizon policies with different horizon lengths leads to ambiguous regret notions in evaluation of learning-based control policies. Thus we propose a regret notion that compares a finite-horizon learning-based policy with a finite-horizon oracle controller as the benchmark. Second, we bound this regret notion for a class of learning-based control policies for which we prove constraint satisfaction. An important aspect of our regret analysis is that we have to consider the stability of our policy when bounding the regret. In this sense, our analysis draws a connection between stability of the nonlinear control system and regret performance of the learning policy.

A. Partially-Unknown Cost Function

MPC usually assumes the system dynamics and a cost function are exactly known. However, these may be partially-unknown in real-world systems that motivate our setup.

1) Heating, Ventilation, Air-Conditioning (HVAC) Systems: Since HVAC uses a large part of total building energy, improving HVAC energy-efficiency using MPC has been studied [8]–[11]. However, past works typically assume perfect knowledge of a cost function that characterizes the trade-off between energy-efficiency and occupant comfort. In practice, the quantity of trade-off is different for each occupant and is *a priori* unknown to the controller. It makes sense to learn an ideal trade-off from occupant-reported data [12] and then adapt the MPC operation in response, which is an example of MPC with a partially-unknown cost function.

2) Clinical-Inventory Management: Inventory management in hospitals involves periodically restocking drugs and medical supplies, and MPC for inventory management [13]–[16] is powerful as it naturally captures the dynamics of consuming and purchasing drugs and supplies. Although past work typically assumes that consumption dynamics are completely characterized, it is not realistic for the demand in hospitals due to unforeseeable medical emergencies. It then makes sense from a practical standpoint to learn about the demand from such events and then adapt the MPC operation in response, which is an example of MPC with learning for the dynamics.

B. Exploration/Exploitation Trade-Off

A challenge in LBMPC is to jointly optimize the control to minimize a cost function and to steer the system to get more information about unknown system or cost parameters [17]. This *exploration/exploitation* trade-off and has been formally studied in the setting of MAB's [18]–[20], RL for finite Markov chains [21]–[23] and for linear systems [24]–[27].

Most work on MAB's assumes (weak-)stationarity because computing the optimal policy with non-stationary is PSPACE-hard [28]. In RL of control systems, past work on nonlinear systems is limited [29]–[34] because the optimal controller for linear systems with a quadratic cost is completely characterized by the Algebraic Ricatti Equation: This allows one to convert the RL problem into simply a parameter estimation problem. However, extending these ideas to nonlinear systems is nontrivial as there is no such simple characterization of the optimal controller, and so alternative approaches are needed. We design a learning-based controller for nonlinear and non-stationary systems where the policy explores to improve the estimation methodology embedded in the learning mechanism.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. This material is based upon work supported by the National Science Foundation under Grant CMMI-1847666.

I. Dogan, Z. M. Shen, and A. Aswani are with Industrial Engineering and Operations Research, University of California, Berkeley 94720 (email: {ilgindogan,maxshen,aaswani}@berkeley.edu)

Z. M. Shen is with Faculty of Engineering and Faculty of Business and Economics, University of Hong Kong, China (email: maxshen@hku.hk).

C. Outline

Sect. II covers preliminaries. Sect. III defines our setup and and proves safety properties for a class of control policies. Sect. IV introduces *N-step dynamic regret*, and Sect. V and VI present a finite sample analysis for the parameter estimation and regret analysis for the *non-myopic ϵ-greedy algorithm*. Lastly, numerical experiments are done in Sect. VII.

II. Preliminaries

A polytope $\mathcal{U}$ in $\mathbb{R}^n$ can be represented as intersection of a set of half-spaces [35]: $\mathcal{U} = \{x : P_i x \leq q_i, i = 1, \dots, d\}$, $P_i \in \mathbb{R}^{d \times n}$, $q_i \in \mathbb{R}^d$. Let $\mathcal{U}, \mathcal{V}$ be two sets. The linear transformation of $\mathcal{U}$ by a matrix $\mathcal{R}$ is $\mathcal{RU} = \{\mathcal{R}u : u \in \mathcal{U}\}$. Their Minkowski sum [36] is defined as $\mathcal{U} \oplus \mathcal{V} = \{u + v : u \in \mathcal{U}; v \in \mathcal{V}\}$ and Pontryagin set difference [37] is defined as $\mathcal{U} \ominus \mathcal{V} = \{u : u + \mathcal{V} \subseteq \mathcal{U}\}$. Note $\mathcal{R}(\mathcal{U} \ominus \mathcal{V}) \subseteq \mathcal{RU} \ominus \mathcal{RV}$ and $(\mathcal{U} \ominus \mathcal{V}) \oplus \mathcal{V} \subseteq \mathcal{U}$.

III. Problem Formulation

Let $x_t \in \mathbb{R}^n$ be states and $u_t \in \mathbb{R}^q$ be inputs. We assume $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ are constrained by (compact) polytopes $\mathcal{X}, \mathcal{U}$. The true system dynamics are $x_{t+1} = f(x_t, u_t, \theta_0) = Ax_t + Bu_t + g(x_t, u_t, \theta_0)$, where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times q}$, $\theta_0 \in \Theta$ for some compact set $\Theta \subseteq \mathbb{R}^p$, and the nonlinear function $g(\cdot, \cdot, \theta) : \mathbb{R}^n \times \mathbb{R}^q \to \mathbb{R}^n$ is parameterized by $\theta \in \Theta$. We assume $\{g(x, u, \theta_0) : x \in \mathcal{X}, u \in \mathcal{U}\} \subseteq \mathcal{W}$ for some bounded polytope $\mathcal{W}$, and $A, B, g, \mathcal{W}, \Theta$ are known but θ_0 is not known to the controller. Define $w_t = g(x_t, u_t, \theta_0)$, and note $w_t \in \mathcal{W}$ by assumption. The intuition is we have a nominal linear model and a partially-unknown, nonlinear correction.

At each time t, the controller receives a stochastic reward r_t from distribution $\mathbb{P}_{x_t, u_t, \theta_0}$ with probability density function $p(r | x_t, u_t, \theta_0)$ and expectation $\mathbb{E}\, r_t = h(x_t, u_t, \theta_0)$. We assume h is parametrically unknown (θ_0 is unknown). This setup can handle stochastic costs c_t (as opposed to rewards) by setting $r_t = -c_t$. We standardize our notation for rewards.

The control problem is to sequentially choose inputs to maximize expected total reward at the end of a finite time horizon $\mathcal{T} = \{0, \dots, T\}$. At time t, the controller has access to past rewards, inputs, and states. Hence, any policy $u_t = \Lambda_t(\mathcal{F}_t)$ will be a sequence (with respect to t) of functions of

$$\mathcal{F}_t = \{r_0, \dots, r_{t-1}, u_0, \dots, u_{t-1}, x_0, \dots, x_t\}. \tag{1}$$

We distinguish between different policies by using superscripts for the sequence of functions Λ_t characterizing the policy.

A. Learning-Based MPC Formulation

LBMPC uses two models: a learned model to enhance performance and a nominal model to provide robustness [5]. Because A, B are known in our setup, the controller uses as its nominal model $\bar{x}_{t+k+1|t} = A\bar{x}_{t+k|t} + Bu_{t+k|t}$, where $\bar{x} \in \mathbb{R}^n$ is system state of the nominal model. The "$|t$" notation denotes the initial condition is taken to be $\bar{x}_{t|t} = x_t$, where x_t is the true state at time t. Because $g(\cdot, \cdot, \theta)$ is also known, the controller uses as its learned model $\tilde{x}_{t+k+1|t} = A\tilde{x}_{t+k|t} + Bu_{t+k|t} + g(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t)$, where $\tilde{x}$ is the system state of the learned model and $\widehat{\theta}_t$ is the controller's estimate of θ_0 at time t. Here, LBMPC learns the true dynamics by updating its estimate of θ_0 as more state measurements become available.

We must first discuss the terminal set used for the MPC. Assuming that (A, B) is stabilizable, there exists a constant state-feedback matrix $K \in \mathbb{R}^{q \times n}$ such that $(A + BK)$ is Schur stable. We assume $\Omega \subseteq \mathcal{X}$ is a maximal output admissible disturbance invariant set [37] meaning that for some stabilizing K it satisfies: a) $\Omega \subseteq \{\overline{x} : \overline{x} \in \mathcal{X} : K\overline{x} \in \mathcal{U}\}$ (*constraint satisfaction*) and b) $(A + BK)\Omega \oplus \mathcal{W} \subseteq \Omega$ (*disturbance invariance*). The intuition is that Ω is a set of states satisfying the constraints $\mathcal{X}$ for which there exists a feasible action keeping the true state within Ω despite the uncertainty of the nominal model. Several algorithms [37]–[40] can compute this set, and so we assume Ω is available to the controller.

With the set Ω, we consider an (simplified) LBMPC variant that maximizes the expected N-step reward. Our results can be generalized straightforwardly to the full formulation [5], but we do not consider this as it adds substantial notational complexity that hinders showcasing the stochastic aspects of our setting. The LBMPC formulation of a finite-horizon N is

$$\begin{aligned} V_N(x_t, \theta, t) = \max & \textstyle\sum_{k=0}^{N} h(\tilde{x}_{t+k|t}, u_{t+k|t}, \theta) \\ \text{s.t. } & \bar{x}_{t+k+1|t} = A\bar{x}_{t+k|t} + Bu_{t+k|t} && k \in \langle N-1 \rangle \\ & \tilde{x}_{t+k+1|t} = A\tilde{x}_{t+k|t} + Bu_{t+k|t} \\ & \qquad + g(\tilde{x}_{t+k|t}, u_{t+k|t}, \theta) && k \in \langle N-1 \rangle \\ & \bar{x}_{t+k|t} \in \mathcal{X} && k \in [N] \\ & u_{t+k|t} \in \mathcal{U} && k \in \langle N \rangle \\ & \bar{x}_{t+1|t} \in \Omega \ominus \mathcal{W}, \ \bar{x}_{t|t} = \tilde{x}_{t|t} = x_t \end{aligned} \tag{2}$$

where $\langle k \rangle = \{0, \dots, k\}$ and $[k] = \{1, \dots, k\}$. The difference between this simplified variant and the full formulation is that here we apply the invariant set Ω at the first time step, an idea previously used in [41], whereas the full formulation uses a robust tube framework to apply Ω at the N-th time point. Our results apply to the above LBMPC formulation and may generalize to the similar variants, but it is unclear if they would generalize to other LBMPC forms without further study.

B. Safety of Learning-Based MPC Variant

Because applying the invariant set to the first time point in an MPC formulation is nonstandard, we first formally prove that this LBMPC variant ensures recursive properties of robust constraint satisfaction and robust feasibility.

Theorem 1: Suppose $\{u_{t|t}, \dots, u_{t+N|t}\}$ are feasible for $V_N(x_t, \theta, t)$ for any θ. If Ω is a maximal output admissible disturbance invariant set, then choosing $u_t = u_{t|t}$ ensures that we have: a) $x_{t+1} \in \mathcal{X}$ (robust constraint satisfaction) and b) there exist values $\{u_{t+1|t+1}, \dots, u_{t+N|t+1}\}$ that are feasible for $V_N(x_{t+1}, \theta', t+1)$ for any θ' (robust feasibility).

Proof: Since $\{u_{t|t}, \dots, u_{t+N|t}\}$ are feasible for $V_N(x_t, \theta, t)$, then $\overline{x}_{t+1|t} = Ax_t + Bu_{t|t} \in \Omega \ominus \mathcal{W}$ by (2). By relating the true dynamics to the nominal model, the true next state is $x_{t+1} = \overline{x}_{t+1|t} + w_t$ for some $w_t \in \mathcal{W}$. This means $x_{t+1} \in (\Omega \ominus \mathcal{W}) \oplus \mathcal{W} \subseteq \Omega \subseteq \mathcal{X}$ where the last set inclusion follows from the constraint satisfaction property in the definition of Ω. By the definition (2) of $V_N(x_{t+1}, \theta', t+1)$, we have that $\overline{x}_{t+1|t+1} = x_{t+1}$. However, we just showed that

$x_{t+1} \in \Omega$. Hence $\bar{x}_{t+1|t+1} \in \Omega$. Now set $u_{t+1|t+1} = Kx_{t+1}$, and note that the constraint satisfaction property of Ω means $u_{t+1|t+1} \in \mathcal{U}$. Since $\overline{x}_{t+2|t+1} = A\overline{x}_{t+1|t+1} + Bu_{t+1|t+1} = (A + BK)\overline{x}_{t+1|t+1}$, we have $\overline{x}_{t+2|t+1} \in (A + BK)\Omega \subseteq ((A + BK)\Omega \oplus \mathcal{W}) \ominus \mathcal{W} \subseteq \Omega \ominus \mathcal{W}$ where the last set inclusion follows by the disturbance invariance property of Ω. So $\bar{x}_{t+2|t+1} \in \Omega \ominus \mathcal{W} \subseteq \Omega \subseteq \mathcal{X}$ by the constraint satisfaction property of Ω. We can sequentially repeat this argument with $u_{t+k+1|t+1} = Kx_{t+k+1|t+1}$ to show this choice results in $u_{t+k+1|t+1} \in \mathcal{U}$ and $x_{t+k+1|t+1} \in \mathcal{X}$ for $k \in [N-1]$. Thus $\{u_{t+1|t+1}, \ldots, u_{t+N|t+1}\}$ are feasible for $V_N(x_{t+1}, \theta', t+1)$. ■

Remark 1: An important feature of the above result is that there is no required relationship between the θ and θ'. Since estimates of the θ are updated through learning, this shows that the safety properties of this LBMPC variant are decoupled from the design of the learning-process.

C. Technical Assumptions

Our learning-based control problem is well-posed under certain regularity assumptions described below.

Assumption 1: The rewards r_t are conditionally independent given θ_0 and x_0, or equivalently, given θ_0 and the complete sequence of $\{u_0, \ldots, u_t, x_0, \ldots, x_t\}$.

Similar to the independent rewards of the stationary MABs, we have independence of $r_t|\{x_t, \theta_0\}$ and $r_{t'}|\{x_{t'}, \theta_0\}$ for $t \neq t'$.

Assumption 2: The log-likelihood ratio $\ell(r, x, u; \theta, \theta') = \log \frac{p(r|x,u,\theta)}{p(r|x,u,\theta')}$ of $\mathbb{P}_{x,u,\theta}$ is locally $L_{\ell,x}$-Lipschitz continuous with respect to x on the compact set $\mathcal{X}$ for $\theta, \theta' \in \Theta$, $u \in \mathcal{U}$.

This ensures continuity of the reward distribution with respect to the parameters. If two parameter sets are close to each other in value, then the resulting distributions will also be similar.

Assumption 3: The distribution $\mathbb{P}_{x,u,\theta}$ for all $x \in \mathcal{X}, u \in \mathcal{U}$, and $\theta \in \Theta$ is sub-Gaussian with parameter σ, and either $p(r|x, u, \theta)$ has a finite support or $\ell(r, u; x, \theta, x', \theta')$ is locally $L_{\ell,r}$-Lipschitz with respect to r.

This assumption ensures sample averages are close to their means and is satisfied by many distributions (e.g., Gaussian with known variance). Our last condition ensures the dynamics and the expectation function are well-behaved.

Assumption 4: Repeated composition of the true dynamics with itself up to $N - 1$ times, $f^{t+k}(x_t, u_{t|t}, \ldots, u_{t+k|t}, \theta)$, is Lipschitz continuous with respect to $x_t \in \mathcal{X}$ and $u_{t+k|t} \in \mathcal{U}$ with constants $L_{f,x}$ and $L_{f,u}$, respectively. Besides, the expectation $h(x_t, u_t, \theta)$, for $u_t = \Lambda_t(\mathcal{F}_t)$ in (1), is Lipschitz continuous with respect to $x_t \in \mathcal{X}$ and $u_t \in \mathcal{U}$ with constants $L_{h,x}$ and $L_{h,u}$, respectively, for all $\theta \in \Theta$.

IV. The N-Step Dynamic Regret

Our interest is in evaluating the performance of an LBMPC *exploitation* policy for a given $N \leq T$ that is $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ for the corresponding value from the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$ where $\widehat{\theta}_t$ are the control policy's estimates of the unknown θ_0. Data-driven policies are often evaluated by comparing performance to a benchmark policy, and it is typical to benchmark using the optimal policy [42]–[44]. In our setting, the optimal policy is a sequence of functions $\Lambda_t^*(\mathcal{F}_t)_{t=0}^T$ maximizing $\sum_{t=0}^T h(x_t, u_t, \theta_0)$ subject to the knowledge available to the control policy (which does not include θ_0). However, computing optimal policies for the problems we consider is PSPACE-hard [28]. Even their structure is not known for our setup, including for the special case of linear dynamics and quadratic cost function with unknown coefficients.

An alternative benchmark is an oracle policy that has perfect knowledge of θ_0. Specifically, we will use the LBMPC *oracle* policy that is $\Lambda_t^{O,N}(\mathcal{F}_t) = u^*_{t|t}(\theta_0)$ for the corresponding value from the maximizer of $V_N(x_t, \theta_0, t)$ as defined in (2). However, there are two subtleties that have to be discussed.

The first subtlety is that the horizon length of the LBMPC oracle policy could potentially be different than the horizon length of the LBMPC policy. However, using different control horizon lengths can lead to different sums of expected rewards over the entire control horizon $\mathcal{T}$. Though this behavior is well known within the MPC community, its implication on evaluating learning-based control policies has not been previously appreciated. The implication is that comparing policies with different horizon lengths leads to a poorly-defined regret notion, and that we should compare oracle policies and learning-based policies with the same finite-horizon.

The second subtlety is that the presence of nonlinear dynamics in our setup means the state trajectory of a system always controlled by a benchmark policy can be very different than that of a system always controlled by a learning policy, even if the learning policy converges towards the benchmark policy. For this reason, we define a regret notion to compare a finite-horizon benchmark policy to a finite-horizon learning-based policy. We consider an ϵ-greedy policy $\Lambda_t^{\epsilon,N}$ that uses the LBMPC policy $\Lambda_t^{E,N}$ at each greedy exploitation step. Let x_t, u_t be the state and input for the system as controlled by the oracle policy $\Lambda_t^{O,N}$, and let x'_t, u'_t be the state and input for the system as controlled by the ϵ-greedy policy $\Lambda_t^{\epsilon,N}$. Then, the expected N*-step dynamic regret* is defined as

$$R_{N,T} = \sum_{t=0}^T h(x_t, \Lambda_t^{O,N}(\mathcal{F}_t), \theta_0) - h(x'_t, \Lambda_t^{\epsilon,N}(\mathcal{F}'_t), \theta_0) \tag{3}$$

where $\mathcal{F}_t$ is as defined in (1) and $\mathcal{F}'_t$ is as defined in (1) with x', u' replacing x, u. This definition is closely related to the traditional dynamic regret [45], [46], and the novel aspect of ours is that it compares two N-step finite-horizon policies.

V. Parameter Estimation

Let the variables $\{r_i\}_{i=0}^{t-1}$ be the actual observed values of the rewards up to time t. Using Assumption 1, the joint likelihood $p(\{r_i\}_{i=0}^{t-1}|x_0, \ldots, x_t, u_0, \ldots, u_{t-1}, \theta)$ can be expressed as $\prod_{i=0}^{t-1} p(r_i|x_i, u_i, \theta)P(x_i|x_{i-1}, \theta)$. Here, the one step transition likelihood $P(x_i|x_{i-1}, \theta)$ is a degenerate distribution with all probability mass at x_i, by perpetuation of the dynamics $f(x_i, u_i, \theta)$ with initial conditions x_{i-1}. Thus, the maximum likelihood estimator (MLE) for θ is

$$\begin{aligned} \widehat{\theta}_t \in \arg\min_{\theta\in\Theta} & -\textstyle\sum_{i=0}^{t-1} \log p(r_i|x_i, u_i, \theta) \\ \text{s.t. } & x_{i+1} = f(x_i, u_i, \theta)\ \forall i \in \{0, \ldots, t-1\} \end{aligned} \tag{4}$$

This MLE problem can be computed using optimization, dynamic programming, or various filtering techniques for different problem structures. The Kalman Filter (KF) is a recursive estimator for linear-quadratic discrete-time systems. In more complex systems with non-Gaussian distributions and nonlinear dynamics, the Extended KF and Particle Filter are well-known estimators [47]–[49]. For practical purposes, these efficient approaches motivate the use of MLE in our policy. Further, if the controller did not have perfect state measurements, we could use the noisy state data to estimate the dynamics in the constraints of (4) [50], [51], which would also alleviate any potential infeasibility issues of the MLE.

We further analyze the concentration properties of the solution to (4) and take an approach to the theoretical analysis that generalizes that of [20]. We begin by introducing the notion of trajectory Kullback–Leibler (KL) divergence. Since this problem includes the joint distribution of a trajectory of values, the concentration bound for the parameter estimates is computed with regards to the trajectory KL divergence.

Definition 1: The *trajectory Kullback–Leibler (KL) divergence* between the parameter trajectories $\theta, \theta' \in \Theta$ with the same input sequence $\Pi_T = \{u_t\}_{t=0}^T$ is $D_{\Pi_T}(\theta||\theta') = \sum_{i=0}^T D_{KL}(\mathbb{P}_{f^i(x_0,\Pi_i,\theta),u_i,\theta}||\mathbb{P}_{f^i(x_0,\Pi_i,\theta'),u_i,\theta'})$, where Π_i is the given sequence of control inputs from time 0 to i, f^i is the repeated composition of the dynamics f with itself i times subject to Π_i, and D_{KL} is the standard KL-Divergence.
We have an *observability* assumption with the implication that the distance between two different parameters $\theta, \theta' \in \Theta$ is bounded proportional to their trajectory KL divergence.

Assumption 5: For a given input sequence Π_T and parameters $\theta \neq \theta'$, if $D_{\Pi_T}(\theta||\theta') \leq \delta$, then $\|\theta - \theta'\| \leq C\delta$ for $C > 0$.
We next reformulate the MLE problem (4) by removing the state dynamics constraints through repeated composition of f, that is $\widehat{\theta}_t \in \arg\min_{\theta\in\Theta} \frac{1}{t-1}\sum_{i=0}^{t-1} \log \frac{p(r_i|f^i(x_0,\Pi_i,\theta_0),u_i,\theta_0)}{p(r_i|f^i(x_0,\Pi_i,\theta),u_i,\theta)}$. This reformulation is helpful for our theoretical analysis since for fixed θ, the expected value of the above objective function under $\mathbb{P}_{x_0,\Pi_T,\theta_0}$ is simply $\frac{1}{t-1}D_{\Pi_T}(\theta_0||\theta)$. Hence, we can interpret the MLE problem as minimizing the trajectory KL divergence between the distribution of potential sets of parameters and that of the true parameter set. This interpretation is helpful for us to derive our concentration inequalities. For conciseness of our analysis in this paper, we present the final concentration bound for $\widehat{\theta}_t$ and do not include its proof since it largely follows by the theoretical arguments in [20].

Theorem 2: For any constant $\zeta > 0$, we have the bound that $P(\frac{1}{t-1}D_{\Pi_t}(\theta_0||\widehat{\theta}_t) \leq \zeta + \frac{c_f(d_x,d_\theta)}{\sqrt{t-1}}) \geq 1 - \exp(-\frac{\zeta^2(t-1)}{2L_{\ell,r}^2\sigma^2})$ where the constant $c_f(d_x, d_\theta) = 8L_{f,x}L_{\ell,x}\text{diam}(\mathcal{X})\sqrt{\pi} + 48\sqrt{2}(2)^{\frac{1}{d_x+d_\theta}}L_{f,x}L_{\ell,x}\text{diam}(\Theta \times \mathcal{X})\sqrt{\pi(d_x + d_\theta)}$ depends upon d_x and d_θ (dimensionalities of $\mathcal{X}$ and Θ), and $\text{diam}(\mathcal{X}) = \max_{x,y\in\mathcal{X}}\|x - y\|_2$.

Proof: Omitted. Refer to Section 3 of [20]. ■

We will use this concentration inequality to prove the regret bound of our non-myopic ϵ-greedy policy that we present next.

VI. Proposed Approach

We develop a *non-myopic ϵ-greedy algorithm* that can achieve effective regret bounds for the non-stationary and nonlinear LBMPC introduced in Section III. Our choice of algorithm aims to draw a connection between the control and MAB literature. A possible alternative could be adding additive noise to the control inputs which we leave as a future work. When compared with the other well-known MAB strategies, Thompson Sampling (TS) and Upper Confidence Bound (UCB), ϵ-greedy is significantly easier from a computational standpoint for combining with the LBMPC formulation of our non-myopic exploitation problem. TS requires characterization of the posterior distribution which is indeed not possible under the general dynamics considered. Similarly, UCB requires being able to compute the confidence bounds which is not feasible in this framework. Hence, those strategies are not practical for the kinds of applications we are interested in.

Our Algorithm 1 explores randomly according to a non-stationary stochastic process. The initial state x_0 is an arbitrary point from the $\mathcal{X}$. At each time $t \in \mathcal{T}$, the algorithm samples a Bernoulli variable s_t based on the exploration probability ϵ_t. If $s_t = 1$, it performs pure exploration. To ensure robust constraint satisfaction and feasibility after exploration, it chooses an input $u_{t|t}$ uniform randomly from $\overline{\mathcal{U}}(x_t) = \{u : Ax_t + Bu \in \Omega \ominus \mathcal{W}, u \in \mathcal{U}\}$. If $s_t = 0$, the algorithm performs a greedy exploitation step by solving the non-myopic exploitation problem $V_N(x_t, \widehat{\theta}_t, t)$ to select the sequence of inputs with the highest MLE-estimated N-step reward. Finally, the algorithm observes the updated state x_{t+1} and reward r_t after applying the chosen input $\Lambda_t^{\epsilon,N}(\mathcal{F}_t)$ to the system.

Algorithm 1 Non-myopic ϵ-Greedy Algorithm

```
1:  Set: $c > 0$ and $x_0 \in \mathcal{X}$
2:  for $t \in \mathcal{T}$ do
3:      Set: $\epsilon_t = \min\{1, c/t\}$
4:      Sample: $s_t \sim$ Bernoulli($\epsilon_t$)
5:      if $s_t = 1$ then
6:          Randomly select: $u_{t|t} \in \overline{\mathcal{U}}(x_t)$
7:          Set: $\Lambda_t^{\epsilon,N}(\mathcal{F}_t) = u_{t|t}$
8:      else
9:          Compute: $\widehat{\theta}_t$ from (4)
10:         Compute: $u^*_{t|t}(\widehat{\theta}_t)$ from $V_N(x_t, \widehat{\theta}_t, t)$ (2)
11:         Set: $\Lambda_t^{\epsilon,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$
12:     end if
13:     Observe: $r_t$ and $x_{t+1}$
14: end for
```

Remark 2: If $\mathcal{W}, \mathcal{X}, \mathcal{U}$ are all polytopes, then Ω can be approximated by a polytope arbitrarily well. Then, $\Omega \ominus \mathcal{W}$ is also a polytope. As a result, line 6 involves randomly picking an element from a polytope that can be done in a computationally efficient way using standard algorithms.
For clarity, we consider a randomization at the initial system state, and then assume noise-free transitions for the subsequent states which is common in the line of RL for finite sample analysis [52]–[55]. Our analysis here provides a strong ground for generalization of our policy to the setting of imperfect state measurements as an important direction for future work.
Note that the exploration probability ϵ_t decays over time. This reduces the cost of exploration by ensuring the algorithm makes fewer unnecessary explorations as more data collected and the estimates of our policy improve.

A. Lipschitzian Stability of Non-myopic Exploitation

We prove Lipschitzian stability, with respect to perturbations of parameter values, of optimal solutions of the non-myopic exploitation policy $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ by proving a second order growth condition and Lipschitz continuity of the difference of the perturbed and unperturbed objective functions.

Lemma 1: Suppose $U_{N,t} = \{u_{t|t}, \dots, u_{t+N|t}\}$ is a feasible input sequence for $V_N(x_t, \widehat{\theta}_t, t)$. Let $J_N(x_t, U_{N,t}, \widehat{\theta}_t, t)$ be the estimated N-step reward of this input sequence at time t, i.e.,

$$J_N(x_t, U_{N,t}, \widehat{\theta}_t, t) = \textstyle\sum_{k=0}^N h(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t) \tag{5}$$

where $\tilde{x}_{t+k+1|t} = f(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t)$ for $k \in \langle N-1\rangle$ as given in (2). Then, $J_N(x_t, U_{N,t}, \widehat{\theta}_t, t)$ is $(L_{f,u} {\cdot} L_{h,u})$-Lipschitz continuous with respect to $U_{N,t}$ on the compact set $\mathcal{U}^{N+1}$ for any feasible input sequence $U'_{N,t} = \{u'_{t|t}, \dots, u'_{t+N|t}\}$.

Proof: By Assumption 4, $f^{t+k}(x_t, u_{t|t}, \dots, u_{t+k|t}, \widehat{\theta})$ is $L_{f,u}$-Lipschitz continuous and $h(\tilde{x}_{t+k|t}, u_{t+k|t}, \widehat{\theta}_t)$ is $L_{h,u}$-Lipschitz continuous with respect to $u_{t+k|t} \in \mathcal{U}$. Then, by preservation of Lipschitz continuity across functional compositions and addition, we have the desired condition. ■

Lemma 1 implies the second order growth condition for $V_N(x_t, \widehat{\theta}_t, t)$ since it shows J_N increases at least linearly over a compact set. We next present the second condition required for the Lipschitzian stability of the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$.

Assumption 6: Let $U^*_{N,t}(\widehat{\theta}_t) = \{u^*_{t|t}(\widehat{\theta}_t), \dots, u^*_{t+N|t}(\widehat{\theta}_t)\}$ and $U^*_{N,t}(\theta) = \{u^*_{t|t}(\theta), \dots, u^*_{t+N|t}(\theta)\}$ be maximizers of $V_N(x_t, \widehat{\theta}_t, t)$ and $V_N(x_t, \theta, t)$. Then, for $\kappa \geq 0$, we have $\|[J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)] - [J_N(x_t, U^*_{N,t}(\theta), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\theta), \theta, t)]\| \leq \kappa \|\widehat{\theta}_t - \theta\| {\cdot} \|U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)\|$.

We now give a sufficient condition for Assumption 6.

Proposition 1: For any $\theta \in \Theta$ and real constant $L_J \geq 0$, if $\|\nabla_u J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - \nabla_u J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)\|_\infty \leq L_J \|\widehat{\theta}_t - \theta\|$ holds, then Assumption 6 is satisfied.

Proof: Let $s(\tau) = U^*_{N,t}(\widehat{\theta}_t) + \tau \cdot (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t))$. This implies $s(0) = U^*_{N,t}(\widehat{\theta}_t)$ and $s(1) = U^*_{N,t}(\theta)$. Then,

$$\begin{aligned}
&[J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\widehat{\theta}_t), \theta, t)] \\
&\quad - [J_N(x_t, U^*_{N,t}(\theta), \widehat{\theta}_t, t) - J_N(x_t, U^*_{N,t}(\theta), \theta, t)] \\
&= \textstyle\int_0^1 \nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t)^T (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau \\
&\quad - \textstyle\int_0^1 \nabla_U J(x_t, s(\tau), \theta, t)^T (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau
\end{aligned} \tag{6}$$

where the last equality follows by the Fundamental Theorem of Calculus for Line Integrals. Then, we continue as

$$\begin{aligned}
&= |\textstyle\int_0^1 [\nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t) - \nabla_U J(x_t, s(\tau), \theta, t)]^T \\
&\qquad\qquad (U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)) d\tau| && (7)\\
&\leq \textstyle\int_0^1 \|\nabla_U J(x_t, s(\tau), \widehat{\theta}_t, t) - \nabla_U J(x_t, s(\tau), \theta, t)\|_\infty \\
&\qquad\qquad ||U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)||_1 d\tau && (8)\\
&\leq L_J \|\widehat{\theta}_t - \theta\| {\cdot} ||U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)\|_1 && (9)\\
&\leq \sqrt{N} L_J \|\widehat{\theta}_t - \theta\| {\cdot} ||U^*_{N,t}(\theta) - U^*_{N,t}(\widehat{\theta}_t)\|_2 && (10)
\end{aligned}$$

where (8) follows by Hölder's inequality, and (9) follows by the assumed property in Proposition 1. This gives us the desired result in Assumption 6 by setting $\kappa = \sqrt{N} L_J$. ■

Lemma 2: If the state dynamics $f(x, u, \theta)$ and the expectation function $h(x, u, \theta)$ are polynomial functions, then the sufficient condition given in Proposition 1 holds.

Proof: Since (5) is the average of compositions of two polynomials f and h, it is polynomial. Then, $\nabla_u J_N(x, U, \theta, t)$ is polynomial on the bounded domain $\mathcal{X} \times \mathcal{U}^{N+1} \times \Theta$. Hence, by Corollary 8.2 in [56], $\nabla_u J_N(x, U, \theta, t)$ is locally Lipschitz with respect to $\theta \in \Theta$ for any $x \in \mathcal{X}, U \in \mathcal{U}^{N+1}, t \in \mathcal{T}$. ■

A specific example where Lemma 2 holds is a discrete-time linear time-invariant system with $f(x, u, \theta) = Ax + Bu$ and $h(x, u, \theta) = x^T Q x + u^T R u$ where $\theta = [Q, R, A, B]$.

Lemma 3: If Assumption 6 and Lemma 1 hold, then the Lipschitzian stability property follows by Proposition 4.32 in [57], i.e., $||U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)|| \leq c_u^{-1} \kappa \|\widehat{\theta}_t - \theta\|$ for $c_u > 0$. Since $||u^*_{t|t}(\widehat{\theta}_t) - u^*_{t|t}(\theta)|| \leq ||U^*_{N,t}(\widehat{\theta}_t) - U^*_{N,t}(\theta)||$, we conclude that the non-myopic exploitation policy $\Lambda_t^{E,N}(\mathcal{F}_t) = u^*_{t|t}(\widehat{\theta}_t)$ corresponding from the maximizer of $V_N(x_t, \widehat{\theta}_t, t)$ is $c_u^{-1}\kappa$-Lipschitz continuous with respect to $\hat{\theta}_t \in \Theta$.

B. Regret Analysis

We next characterize the N-step dynamic regret $R_{N,T}$ (3) of Algorithm 1. By definition, $R_{N,T}$ compares the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}_t)$ for the system x_t, u_t as controlled by the oracle policy to our non-myopic ϵ-greedy policy $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$ for the system x'_t, u'_t as controlled by the learning-policy that uses the LBMPC policy $\Lambda_t^{E,N}(\mathcal{F}'_t)$ at greedy exploitation steps. We start by bounding a weaker notion that compares the actions chosen under the states x'_t achieved by $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$.

Theorem 3: The non-myopic ϵ-greedy policy $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$ and the LBMPC oracle policy $\Lambda_t^{O,N}(\mathcal{F}'_t)$ satisfy the following result for the system states x'_t that are achieved by $\Lambda_t^{\epsilon,N}(\mathcal{F}'_t)$:

$$\begin{aligned}
&\textstyle\sum_{t=0}^T h(x'_t, \Lambda^{O,N}(\mathcal{F}'_t), \theta_0) - \sum_{t=0}^T h(x'_t, \Lambda^{\epsilon,N}(\mathcal{F}'_t), \theta_0) \\
&\leq \mathcal{M} \exp(\tfrac{c_f^2(d_x, d_\theta)}{2L_{\ell,r}^2 \sigma^2})(\mathcal{C} + \log T) \\
&\quad + \mathcal{M} c(1 - \log(c+1) + \log T) + \tfrac{L_{h,u} \kappa C \sqrt{4L_{\ell,r}^2 \sigma^2}}{c_u} \sqrt{T} \log T
\end{aligned} \tag{11}$$

where $C > 0$, $c_f(d_x, d_\theta)$ is the constant in Theorem 2, and $\mathcal{C}$ is a bound on the finite summation $\sum_{t=1}^9 \exp(-(\log t)^2)$.

Proof: For notational convenience, let $\mathbb{E}[M_t] = h(x'_t, \Lambda^{O,N}(\mathcal{F}'_t), \theta_0) - h(x'_t, \Lambda^{\epsilon,N}(\mathcal{F}'_t), \theta_0)$. Let $\mathcal{T}^{\text{xit}} \in \mathcal{T}$ and $\mathcal{T}^{\text{xre}} \in \mathcal{T}$ be the set of random time points that Algorithm (1) performs exploitation and exploration, respectively. Noticing the cardinalities $\#\mathcal{T}^{\text{xit}}$, $\#\mathcal{T}^{\text{xre}}$ are random variables, we have $\sum_{t=0}^T \mathbb{E}[M_t] = \sum_{t \in \mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) + \sum_{t \in \mathcal{T}^{\text{xre}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u_{t|t}, \theta_0)$. We note that $\mathbb{E}[M_t]$ is a bounded value since $\mathcal{X}, \Theta, \mathcal{U}$ are all compact sets and $h(x, u, \theta)$ is a bounded continuous function on this domain. Then, assuming $\mathbb{E}[M_t] \leq \mathcal{M}$, we obtain

$$\begin{aligned}
&[\textstyle\sum_{t=0}^T \mathbb{E}[M_t] | \mathcal{T}^{\text{xit}}] \leq \mathcal{M}\mathbb{E}[\#\mathcal{T}^{\text{xre}}] \\
&\quad + \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0)
\end{aligned} \tag{12}$$

We can rewrite each term inside the summation above as

$$\begin{aligned}
&h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) \\
&= \mathbb{E}[M_t | D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}, x_t, \theta_0, \widehat{\theta}_t] P(D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}) \\
&\quad + \mathbb{E}[M_t | D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \geq \delta_{\widehat{\theta}_t}, x_t, \theta_0, \widehat{\theta}_t] P(D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \geq \delta_{\widehat{\theta}_t}) \\
&= (13, a) + (13, b) \qquad (13)
\end{aligned}$$

Let $\varepsilon(\delta_{\widehat{\theta}_t}) = \max\{\|\theta_0 - \widehat{\theta}_t\| \colon D_{\Pi_t}(\theta_0 || \widehat{\theta}_t) \leq \delta_{\widehat{\theta}_t}\}, \forall t \in \mathcal{T}^{\text{xit}}$.

$$\begin{aligned}
&\textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} (13, a) \\
&= \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} h(x'_t, u^*_{t|t}(\theta_0), \theta_0) - h(x'_t, u^*_{t|t}(\widehat{\theta}_t), \theta_0) && (14) \\
&\leq \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} L_{h,u} \|u^*_{t|t}(\theta_0) - u^*_{t|t}(\widehat{\theta}_t)\| && (15) \\
&\leq \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} L_{h,u} \|U^*_{N|t}(\theta_0) - U^*_{N|t}(\widehat{\theta}_t)\| && (16) \\
&\leq \tfrac{L_{h,u}\kappa}{c_u} \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} \|\theta_0 - \widehat{\theta}_t\| && (17) \\
&\leq \tfrac{L_{h,u}\kappa}{c_u} \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} \varepsilon(\delta_{\widehat{\theta}_t}) && (18)
\end{aligned}$$

where (15) follows by Assumption 4 and (17) follows by Lemma 3. Now, we have $\varepsilon(\delta_{\widehat{\theta}_t}) = C\delta_{\widehat{\theta}_t}$ for a constant $C > 0$ by Assumption 5 and let $\eta(t) = |\{s \in \mathcal{T}^{\text{xit}} : s \leq t\}|$. Then, for $\delta_{\widehat{\theta}_t} = \sqrt{4L^2_{\ell,r}\sigma^2 \log \eta(t)} / \sqrt{\eta(t)}$, we obtain

$$\begin{aligned}
&\leq \tfrac{L_{h,u}\kappa C \sqrt{4L^2_{\ell,r}\sigma^2}}{c_u} \sqrt{\#\mathcal{T}^{\text{xit}}} \log \#\mathcal{T}^{\text{xit}} && (19) \\
&\leq \tfrac{L_{h,u}\kappa C \sqrt{4L^2_{\ell,r}\sigma^2}}{c_u} \sqrt{T} \log T && (20)
\end{aligned}$$

To bound the second term in (13), recall $\mathbb{E}[M_t] \leq \mathcal{M}$. Then,

$$\textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} (13, b) \leq \mathcal{M} \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} \exp\Big(\frac{-(\delta_{\widehat{\theta}_t}\sqrt{t-1} - c_f(d_x, d_\theta))^2}{2L^2_{\ell,r}\sigma^2}\Big) \tag{21}$$

$$\begin{aligned}
&\leq \mathcal{M} \textstyle\sum_{t \in \mathcal{T}^{\text{xit}}} \exp\Big(\frac{-\delta^2_{\widehat{\theta}_t}(t-1)/2 + c^2_f(d_x, d_\theta)}{2L^2_{\ell,r}\sigma^2}\Big) && (22) \\
&\leq \mathcal{M} \exp\Big(\tfrac{c^2_f(d_x, d_\theta)}{2L^2_{\ell,r}\sigma^2}\Big) \\
&\big(\textstyle\sum_{t=1}^{9} \exp(-(\log t)^2) + \sum_{t \in \mathcal{T}^{\text{xit}}, t \geq 10} \exp(-\log t)\big) && (23) \\
&\leq \mathcal{M} \exp\Big(\tfrac{c^2_f(d_x, d_\theta)}{2L^2_{\ell,r}\sigma^2}\Big)(\mathcal{C} + \log T) && (24)
\end{aligned}$$

where (21) follows by Theorem 2 and $\mathcal{C}$ can be approximated as 2.2232. Lastly, we bound the first term in (12): $\mathcal{M}\mathbb{E}[\#\mathcal{T}^{\text{xre}}] = \mathcal{M}\sum_{t=0}^{T} \min\{1, \frac{c}{t}\} \leq \mathcal{M}(c + \sum_{t=c+1}^{T} \frac{c}{t}) \leq \mathcal{M}c(1 - \log(c+1) + \log T)$. Substituting these into (12):

$$\begin{aligned}
&[\textstyle\sum_{t=0}^{T} \mathbb{E}[M_t] | \mathcal{T}^{\text{xit}}] \leq \mathcal{M} \exp\Big(\tfrac{c^2_f(d_x, d_\theta)}{2L^2_{\ell,r}\sigma^2}\Big)(\mathcal{C} + \log T) \\
&\quad + \mathcal{M}c(1 - \log(c+1) + \log T) + \tfrac{L_{h,u}\kappa C \sqrt{4L^2_{\ell,r}\sigma^2}}{c_u} \sqrt{T} \log T
\end{aligned}$$

and taking the expectation gives us the desired result. ■

We analyze regret of $\Lambda^{\epsilon,N}(\mathcal{F}'_t)$ by assuming stability of $\Lambda^{O,N}_t(\mathcal{F}_t)$. If the LBMPC from Sect. III does not provide stability, the full LBMPC formulation [5] can achieve stability. Our results in this paper generalize to the full formulation but at the expense of substantial notational complexity.

Assumption 7: Let $x_{eq} \in \Omega$ be an equilibrium for the LBMPC system in Sect. III. For $\alpha \in [0, 2/3]$ and $\mathcal{F}_t$ as in (1), the LBMPC oracle policy $\Lambda^{O,N}_t(\mathcal{F}_t)$ satisfies $\|Ax_t + B\Lambda^{O,N}_t(\mathcal{F}_t) + g(x_t, \Lambda^{O,N}_t(\mathcal{F}_t), \theta_0) - x_{eq}\| \leq \alpha\|x_t - x_{eq}\| \ \forall t$. Exponential stability of the nonlinear LBMPC implied by this assumption can be ensured under certain sufficient conditions established in the literature [58], [59]. Generalizing the results with less restrictive stability notions poses future research.

Theorem 4: For $4 \leq c \leq \sqrt[4]{T}/\sqrt{3}$, the expected N-step dynamic regret $R_{N,T}$ (3) for a policy $\Lambda^{\epsilon,N}(\mathcal{F}'_t)$ computed by Algorithm 1 satisfies $R_{N,T} \leq 2L_{h,x}\sqrt{T}\text{diam}(\mathcal{X}) + \frac{2L_{h,x}c(3-\alpha)}{1-\alpha}\text{diam}(\mathcal{X}) \log T + \frac{4L_{h,x}\overline{C}c^2}{\alpha}\sqrt{T}(\log T)^3 + \mathcal{M}\exp(\frac{c^2_f(d_x, d_\theta)}{2L^2_{\ell,r}\sigma^2})(\mathcal{C} + \log T) + \mathcal{M}c(1 - \log(c+1) + \log T) + \frac{L_{h,u}\kappa C\sqrt{4L^2_{\ell,r}\sigma^2}}{c_u}\sqrt{T}\log T$ with probability at least $[1 - (T - 2\sqrt{T})\exp\big(-\frac{4c^2\left(\log \frac{e(2\sqrt{T}+2}{c+1}\right)^2}{2c\log(2\sqrt{T}+1) + \frac{2c^2}{2\sqrt{T}+1} + \frac{4c^2}{3}\log\frac{e(2\sqrt{T}+2)}{c+1}}\big) - \exp\big(-\frac{c^2\left(\log\frac{T}{2\sqrt{T}+1}\right)^2}{(4+\frac{2}{3}c^2)\log T}\big)]$ where $\overline{C} = c_u^{-1}(\|B\| + L_{f,u})\kappa C\sqrt{4L^2_{\ell,r}\sigma^2}$.

Proof: By Assumption 4 and the upper bound in (11),

$$\begin{aligned}
R_{N,T} &= \textstyle\sum_{t=0}^{T} h(x_t, \Lambda^{O,N}_t(\mathcal{F}_t), \theta_0) - h(x'_t, \Lambda^{O,N}_t(\mathcal{F}'_t), \theta_0) \\
&\quad + \textstyle\sum_{t=0}^{T} h(x'_t, \Lambda^{O,N}_t(\mathcal{F}'_t), \theta_0) - h(x'_t, \Lambda^{\epsilon,N}_t(\mathcal{F}'_t), \theta_0) \\
&\leq L_{h,x} \textstyle\sum_{t=0}^{T} \|x_t - x'_t\| + (11) \qquad (25)
\end{aligned}$$

Algorithm (1) performs exploration at random times according to a non-stationary stochastic process over $\mathcal{T}$. We divide $\mathcal{T}$ into "inter-explore intervals" composed of an exploration and the subsequent exploitations until the next one is reached. Let $I_k = [\underline{I}_k, \overline{I}_k]$ be the k^{th} sub-interval such that $I_{-1} = [0, 2\lceil\sqrt{T}\rceil]$, $I_0 = [2\lceil\sqrt{T}\rceil + 1, t^{\text{xre}}_1 - 1]$, $I_k = [t^{\text{xre}}_k, t^{\text{xre}}_{k+1} - 1]$ for $k \in [1, K-1]$ where t^{xre}_k is the k^{th} exploration step after time $2\lceil\sqrt{T}\rceil$, and $I_K = [t^{\text{xre}}_K, T]$ where $K = \sum_{t=2\lceil\sqrt{T}\rceil+1}^{T} s_t$ and $s_t \sim \text{Bernoulli}\,(\min\{1, c/t\})$. Then, $\sum_{t=0}^{T}\|x_t - x'_t\| = \sum_{k=-1}^{K}\sum_{t \in I_k}\|x_t - x'_t\|$. The key idea is that regret over each $I_k, k \in [0, K]$ is bounded above by the regret over $S_k = [\underline{S}_k, \overline{S}_k] = [\underline{I}_k, T]$ that includes a single exploration at time $\underline{I}_k$ followed by exploitation steps thereafter up to T.

Suppose Algorithm 1 uses $\Lambda^{O,N}_t(\mathcal{F}_t)$ at all greedy exploitation steps of $S_k, k \in [0, K]$. Since $x_t \in \mathcal{X}$ for $t \in \mathcal{T}^{\text{xre}}$ and $\mathcal{X}$ is compact, $\|x_t - x_{eq}\| \leq \text{diam}(\mathcal{X})$, $t \in \mathcal{T}^{\text{xre}}$. Then, by Assumption 7, $\sum_{t \in I_k}\|x_t - x_{eq}\| \leq \sum_{t \in S_k}\|x_t - x_{eq}\| = \|x_{t^{\text{xre}}_k} - x_{eq}\| + \sum_{t=\underline{S}_k+1}^{\overline{S}_k}\|x_t - x_{eq}\| \leq \text{diam}(\mathcal{X}) + \sum_{t=\underline{S}_k+1}^{\overline{S}_k}\alpha^{t-\underline{S}_k}\text{diam}(\mathcal{X}) \leq \text{diam}(\mathcal{X})/(1-\alpha)$. Next, suppose instead $\Lambda^{E,N}_t(\mathcal{F}'_t)$ is used at all greedy exploitation steps of $S_k, k \in [0, K]$. Observe the convergence of $\Lambda^{E,N}(\mathcal{F}'_t)$:

$$\begin{aligned}
&\|Ax'_t + B\Lambda^{E,N}_t(\mathcal{F}'_t) + g(x'_t, \Lambda^{E,N}_t(\mathcal{F}'_t), \theta_0) - x_{eq}\| \\
&\leq \|Ax'_t + B\Lambda^{O,N}_t(\mathcal{F}'_t) + g(x'_t, \Lambda^{O,N}_t(\mathcal{F}'_t), \theta^o) - x_{eq}\| \\
&\quad + \|B\Lambda^{E,N}_t(\mathcal{F}'_t) + g(x'_t, \Lambda^{E,N}_t(\mathcal{F}'_t), \theta_0) \\
&\qquad - B\Lambda^{O,N}_t(\mathcal{F}'_t) - g(x'_t, \Lambda^{O,N}_t(\mathcal{F}'_t), \theta_0)\| && (26) \\
&\leq \alpha\|x'_t - x_{eq}\| + \|B\Lambda^{E,N}_t(\mathcal{F}'_t) + g(x'_t, \Lambda^{E,N}_t(\mathcal{F}'_t), \theta_0) \\
&\qquad - B\Lambda^{O,N}_t(\mathcal{F}'_t) - g(x'_t, \Lambda^{O,N}_t(\mathcal{F}'_t), \theta_0)\| && (27)
\end{aligned}$$

where (26) follows by the triangle inequality and (27) follows by Assumption 7. Recall that $\|\Lambda^{E,N}_t(\mathcal{F}'_t) - \Lambda^{O,N}_t(\mathcal{F}'_t)\| \leq \frac{\kappa C\sqrt{4L^2_{\ell,r}\sigma^2}}{c_u}\frac{\log\eta(t)}{\sqrt{\eta(t)}}$ as followed from (15) to (19). By Assumption 4, we have (27) $\leq \alpha\|x'_t - x_{eq}\| + \overline{C}\frac{\log\eta(t)}{\sqrt{\eta(t)}}$, where

$\overline{C} = \frac{(\|B\|+L_{f,u})\kappa C\sqrt{4L_{\ell,r}^2\sigma^2}}{c_u}$. Then, for $k \in [0, K]$,

$$\begin{aligned} &\textstyle\sum_{t\in I_k}\|x_t' - x_{eq}\| \leq \sum_{t\in S_k}\|x_t' - x_{eq}\| \\ &= \|x'_{t_k^{\text{xre}}} - x_{eq}\| + \textstyle\sum_{t=\underline{S}_k+1}^{\overline{S}_k}\|x_t' - x_{eq}\| \leq \operatorname{diam}(\mathcal{X}) + \\ &\quad \textstyle\sum_{t=t_k^{\text{xre}}+1}^{T}[\alpha^{t-t_k^{\text{xre}}}\operatorname{diam}(\mathcal{X}) + \overline{C}\sum_{i=t_k^{\text{xre}}+1}^{t-1}\frac{\alpha^{t-1-i}\log\eta(i)}{\sqrt{\eta(i)}}] \\ &\leq \tfrac{2-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \overline{C}\textstyle\sum_{t=t_k^{\text{xre}}+1}^{T}\sum_{i=t_k^{\text{xre}}+1}^{t-1}\alpha^{t-1-i}\frac{\log\eta(i)}{\sqrt{\eta(i)}} \end{aligned}$$

Recall $\eta(i) = i - \sum_{j=1}^{i} s_j$ where $s_j \sim \text{Bernoulli}(\min\{1, c/j\})$ and $\mathbb{E}\sum_{j=1}^{i} s_j \leq c + \int_{j=c}^{i}\frac{c}{j}dj = c\log\frac{ei}{c}$. By conditioning on the event $\mathcal{E}_i = \{\sum_{j=1}^{i} s_j \leq 3\mathbb{E}\sum_{j=1}^{i} s_j\}$, we get $\eta(i) \geq i - 3\mathbb{E}\sum_{j=1}^{i} s_j \geq i - 3c\log\frac{ei}{c} \geq \frac{i}{c^2}$ where the last inequality holds for all $i \geq 2\lceil\sqrt{T}\rceil + 1 \geq 6c^2 + 1$. Then, for $\alpha \in [0, 2/3]$,

$$\begin{aligned} &\leq \tfrac{2-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \overline{C}c\textstyle\sum_{t=t_k^{\text{xre}}+1}^{T}\sum_{i=t_k^{\text{xre}}+1}^{t-1}\alpha^{t-i-1}\frac{\log i}{\sqrt{i}} \\ &\leq \tfrac{2-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \tfrac{\overline{C}c}{\alpha}\textstyle\sum_{t=t_k^{\text{xre}}+1}^{T}\sum_{i=t_k^{\text{xre}}+1}^{t-1}\frac{\log i}{(t-i)\sqrt{i}} \\ &\leq \tfrac{2-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \tfrac{2\overline{C}c}{\alpha}\textstyle\sum_{t=t_k^{\text{xre}}+1}^{T}\frac{(\log t)^2}{\sqrt{t}} \\ &\leq \tfrac{2-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \tfrac{2\overline{C}c}{\alpha}\sqrt{T}(\log T)^2 \qquad (28) \end{aligned}$$

Note $\mathbb{E}\sum_{j=1}^{i} s_j \geq c\log\frac{e(i+1)}{c+1}$ and $\text{Var}(\sum_{j=1}^{i} s_j) = \sum_{j=1}^{i}\frac{c}{j}\cdot\frac{j-c}{j} \leq c\log i + \frac{c^2}{i}$. Then, by Bernstein's inequality (Corollary 2.11 in [60]), it follows that (28) holds with probability $\mathbb{P}(\cap_{i=t_k^{\text{xre}}+1}^{T}\mathcal{E}_i) \geq 1 - \sum_{i=t_k^{\text{xre}}+1}^{T}P(\overline{\mathcal{E}}_i) \geq 1 - \sum_{i=t_k^{\text{xre}}+1}^{T}\exp(-\frac{4c^2(\log\frac{e(i+1)}{c+1})^2}{2c\log i + \frac{2c^2}{i} + \frac{4c^2}{3}\log\frac{e(i+1)}{c+1}}) \geq 1 - (T - 2\sqrt{T})\exp(-\frac{4c^2(\log\frac{e(2\sqrt{T}+2}{c+1})^2}{2c\log(2\sqrt{T}+1) + \frac{2c^2}{2\sqrt{T}+1} + \frac{4c^2}{3}\log\frac{e(2\sqrt{T}+2)}{c+1}})$. The above bounds for $\Lambda^{O,N}(\mathcal{F}_t)$ and (28) for $\Lambda^{\epsilon,N}(\mathcal{F}_t')$ allow us to bound the deviation of the system trajectory under the learning policy from the one under the oracle policy over I_{-1} as $\sum_{t\in I_{-1}}\|x_t - x_t'\| \leq 2\sqrt{T}\operatorname{diam}(\mathcal{X})$ and over $I_k, k \geq 0$ as $\sum_{t\in I_k}\|x_t - x_t'\| \leq \sum_{t\in S_k}\|x_t - x_{eq}\| + \sum_{t\in S_k}\|x_t' - x_{eq}\| \leq \frac{3-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \frac{2\overline{C}c}{\alpha}\sqrt{T}(\log T)^2$. Combining this with (25), we obtain $R_{N,T} \leq 2L_{h,x}\sqrt{T}\operatorname{diam}(\mathcal{X}) + L_{h,x}K(\frac{3-\alpha}{1-\alpha}\operatorname{diam}(\mathcal{X}) + \frac{2\overline{C}\sqrt{c}}{\alpha}\sqrt{T}(\log T)^2) + (11)$, and it remains to bound K. Note $\mathbb{E}K \geq c\log\frac{T+1}{2\sqrt{T}+1}$, $\text{Var}(K) \leq 2\log T$, and Bernstein's inequality yields $\mathbb{P}(K \leq 2\mathbb{E}K) \geq 1 - \exp(-\frac{c^2(\log\frac{T}{2\sqrt{T}+1})^2}{(4+\frac{2}{3}c^2)\log T})$. Bounding K by $2\mathbb{E}K \leq 2c\log T$ gives the desired result. ■

This instantaneous bound implies asymptotic N-step dynamic regret of order $O(\sqrt{T}(\log T)^3)$ for Algorithm 1.

VII. Numerical Experiments

We conduct experiments using Python 3.7.4 and Anaconda on a laptop with 2.3 GHz 8-Core Intel Core i9 processor and 16GB DDR4 RAM. We use MOSEK [61] for optimization. We simulate an HVAC system (see Sect. I-A.1), using a discrete time model from [8] with 15 minutes sampling interval and dynamics $x_{t+1} = k_r x_t - k_c u_t + k_v v_t + q_t$, where $x_t \in [20, 24]$ in $^\circ C$, $u_t \in [0, 0.5]$ is AC duty cycle, v_t is outside temperature in $^\circ C$, and q_t is heating load due to occupants. We assume $r_t = -c_t \sim \mathcal{N}\left(h(x_t, u_t, \theta_0), \sigma^2\right)$ for $h(x_t, u_t, \theta_0) = \gamma_1 p_t u_t + (x_t - \gamma_2 - v_t)^2$ where p_t is the electricity price assumed to follow a peak-pricing plan between 12-6 p.m. over an 24 hour day. The $\gamma_1 p_t u_t$ accounts for energy use, and $v_t + \gamma_2$ indicates a setpoint preference that adjusts with outside temperature [62]. We suppose $\theta_0 = [q_t, \gamma_1, \gamma_2]$ are unknown to the controller, and use $\sigma = 1, k_r = 0.64, k_c = 2.64, k_v = 0.10$ [8]. We assume v_t and q_t are generated from a sinusoidal distribution with a single peak over 24 hours and average values of 6.98 and 17, respectively. All metrics are averaged across 1000 replicates. Fig. 1 shows regret up to time $T = 100,000$ of the $N = 10$ policy. These results are compatible with our asymptotic regret bound $O(\sqrt{T}(\log T)^3)$. Fig. 2 compares cumulative expected costs of the $N = 1$ and $N = 10$ policies by subtracting the expected cost of $\Lambda_t^{\epsilon,10}(\mathcal{F}_t')$ from that of $\Lambda_t^{\epsilon,1}(\mathcal{F}_t')$. Lower costs are obtained with $N = 10$.

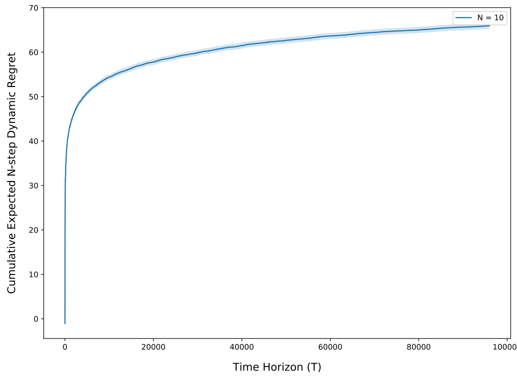

Fig. 1: Expected 10-step dynamic regret.

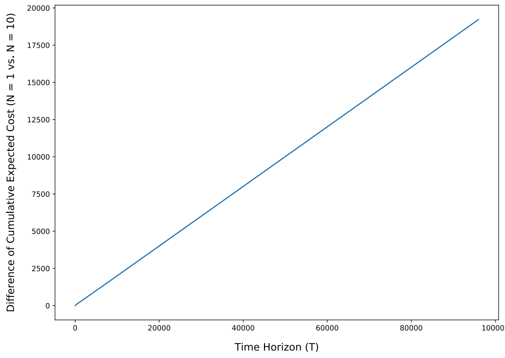

Fig. 2: Difference of cumulative expected costs for $N = 1, 10$.

VIII. Conclusion

This paper studies the intersection of nonlinear MPC and RL. Stability is one of the unique (and not previously well-studied) issues that arises with RL for nonlinear systems. We develop a new class of LBMPC policies that we prove achieves low regret, which is supported by our numerical experiments.

References

[1] Y. Abbasi-Yadkori, N. Lazic, and C. Szepesvari, "Model-free linear quadratic control via reduction to expert prediction," in *AISTATS*, 2019.

[2] C. Chen, H. Modares, K. Xie, F. L. Lewis, Y. Wan, and S. Xie, "Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics," *IEEE TAC*, 2019.

[3] N. Agarwal, N. Brukhim, E. Hazan, and Z. Lu, "Boosting for control of dynamical systems," in *ICML*, 2020.

[4] R. Negenborn, B. De Schutter, M. Wiering, and J. Hellendoorn, "Experience-based model predictive control using reinforcement learning," in *Proceedings of the 8th TRAIL Congress*, 2004.

[5] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica*, vol. 49, 2013.

[6] N. Karnchanachari, M. I. Valls, D. Hoeller, and M. Hutter, "Practical reinforcement learning for mpc: Learning from sparse objectives in under an hour on a real robot," in *L4DC*, 2020.

[7] S. Gros and M. Zanon, "Reinforcement learning for mixed-integer problems based on mpc," *ArXiv*, vol. abs/2004.01430, 2020.

[8] A. Aswani, N. Master, J. Taneja, D. Culler, and C. Tomlin, "Reducing transient and steady state electricity consumption in hvac using learning-based model-predictive control," *Proc. IEEE*, vol. 100, 2011.

[9] A. Afram and F. Janabi-Sharifi, "Theory and applications of hvac control systems–a review of model predictive control (mpc)," *Building and Environment*, vol. 72, pp. 343–355, 2014.

[10] M. Ostadijafari and A. Dubey, "Linear model-predictive controller (lmpc) for building's heating ventilation and air conditioning (hvac) system," in *IEEE CCTA*. IEEE, 2019, pp. 617–623.

[11] J. Fang, R. Ma, and Y. Deng, "Identification of the optimal control strategies for the energy-efficient ventilation under the model predictive control," *Sustainable Cities and Society*, vol. 53, 2020.

[12] A. Aswani, Z.-J. Shen, and A. Siddiq, "Inverse optimization with noisy data," *Operations Research*, vol. 66, no. 3, pp. 870–892, 2018.

[13] P. Velarde, J. Maestre, I. Jurado, I. Fernandez, B. I. Tejera, and J. del Prado, "Application of robust model predictive control to inventory management in hospitalary pharmacy," in *IEEE ETFA*, 2014.

[14] G. Schildbach and M. Morari, "Scenario-based model predictive control for multi-echelon supply chain management," *European Journal of Operational Research*, vol. 252, no. 2, pp. 540–549, 2016.

[15] J. Maestre, M. Fernández, and I. Jurado, "An application of economic model predictive control to inventory management in hospitals," *Control Engineering Practice*, vol. 71, pp. 120–128, 2018.

[16] I. F. Garcia, P. Chanfreut, I. Jurado, and J. M. Maestre, "A data-based model predictive decision support system for inventory management in hospitals," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[17] A. Mesbah, "Stochastic model predictive control with active uncertainty learning: A survey on dual control," *Annu. Rev. Control.*, vol. 45, pp. 107–117, 2018.

[18] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[19] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *ICML*, 2013, pp. 127–135.

[20] Y. Mintz, A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka, "Non-stationary bandits with habituation and recovery dynamics," *Operations Research*, vol. 68, 2017.

[21] M. Heger, "Consideration of risk in reinforcement learning," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 105–111.

[22] E. Biyik, J. Margoliash, S. R. Alimo, and D. Sadigh, "Efficient and safe exploration in deterministic markov decision processes with unknown transition models," in *ACC*, 2019, pp. 1792–1799.

[23] M. Budd, B. Lacerda, P. Duckworth, A. West, B. Lennox, and N. Hawes, "Markov decision processes with unknown state feature values for safe exploration using gaussian processes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[24] S. J. Bradtke, B. E. Ydstie, and A. G. Barto, "Adaptive linear quadratic control using policy iteration," in *ACC*, vol. 3, 1994, pp. 3475–3479.

[25] B. Kiumarsi-Khomartash, F. Lewis, and Z. Jiang, "H∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Autom.*, vol. 78, pp. 144–152, 2017.

[26] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 2019.

[27] M. Simchowitz and D. Foster, "Naive exploration is optimal for online LQR," in *ICML*, vol. 119, 2020.

[28] C. H. Papadimitriou and J. N. Tsitsiklis, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, 1999.

[29] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE conference on decision and control (CDC)*. IEEE, 2018, pp. 6059–6066.

[30] S. Gros and M. Zanon, "Data-driven economic nmpc using reinforcement learning," *IEEE TAC*, vol. 65, no. 2, pp. 636–648, 2019.

[31] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun, "Information theoretic regret bounds for online nonlinear control," *arXiv preprint arXiv:2006.12466*, 2020.

[32] K. P. Wabersich and M. N. Zeilinger, "Performance and safety of bayesian model predictive control: Scalable model-based rl with guarantees," *arXiv preprint arXiv:2006.03483*, 2020.

[33] Y. Fan and Y. Ming, "Efficient exploration for model-based reinforcement learning with continuous states and actions," *arXiv preprint arXiv:2012.09613*, 2020.

[34] N. M. Boffi, S. Tu, and J.-J. E. Slotine, "Regret bounds for adaptive nonlinear control," in *Learning for Dynamics and Control*, 2021.

[35] F. Borelli, A. Bemporad, and M. Morari, "Constrained optimal control and predictive control for linear and hybrid systems," 2009.

[36] R. Schneider, "Convex bodies: The brunn-minkowski theory," 1993.

[37] I. Kolmanovsky and E. G. Gilbert, "Theory and computation of disturbance invariant sets for discrete-time linear systems," *Mathematical Problems in Engineering*, vol. 4, pp. 317–367, 1998.

[38] D. Limon, T. Alamo, D. M. Raimondo, D. M. De La Peña, J. M. Bravo, A. Ferramosca, and E. F. Camacho, "Input-to-state stability: a unifying framework for robust model predictive control," in *Nonlinear model predictive control*, 2009, pp. 1–26.

[39] S. V. Rakovic and M. Baric, "Parameterized robust control invariant sets for linear systems: Theoretical advances and computational remarks," *IEEE Transactions on Automatic Control*, vol. 55, 2010.

[40] Z. Wang, R. M. Jungers, and C. J. Ong, "Computation of the maximal invariant set of discrete-time linear systems subject to a class of non-convex constraints," *Automatica*, vol. 125, 2021.

[41] A. Aswani, P. Bouffard, and C. Tomlin, "Extensions of learning-based model predictive control for real-time application to a quadrotor helicopter," in *2012 American Control Conference (ACC)*, 2012.

[42] A. Garivier and E. Moulines, "On upper-confidence bound policies for non-stationary bandit problems," 2008.

[43] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Advances in neural information processing systems*, 2014, pp. 199–207.

[44] D. Bouneffouf and R. Féraud, "Multi-armed bandit problem with known trend," *Neurocomputing*, vol. 205, pp. 16–21, 2016.

[45] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *ICML*, 2003, pp. 928–936.

[46] E. C. Hall and R. M. Willett, "Dynamical models and tracking regret in online convex programming," *arXiv preprint arXiv:1301.1254*, 2013.

[47] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 1960.

[48] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," *J Comput Graph Stat*, vol. 5, no. 1, 1996.

[49] B. D. Anderson and J. B. Moore, *Optimal filtering*. Dover, 2012.

[50] K. S. Amelin and O. N. Granichin, "Randomized controls for linear plants and confidence regions for parameters under external arbitrary noise," *2012 American Control Conference (ACC)*, 2012.

[51] A. Kalmuk, K. Tyushev, O. N. Granichin, and M. Yuchi, "Online parameter estimation for mpc model uncertainties based on lscr approach," *IEEE CCTA*, 2017.

[52] D. P. Bertsekas and H. Yu, "Distributed asynchronous policy iteration in dynamic programming," in *Allerton*, 2010, pp. 1368–1375.

[53] A. Lazaric, M. Ghavamzadeh, and R. Munos, "Analysis of a classification-based policy iteration algorithm," in *ICML*, 2010.

[54] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, 2014.

[55] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1467–1476.

[56] D. Estep, *Practical Analysis in One Variable*. Springer, 2010.

[57] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[58] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[59] G. Pannocchia, J. B. Rawlings, and S. J. Wright, "Conditions under which suboptimal nonlinear mpc is inherently robust," *Systems & Control Letters*, vol. 60, no. 9, pp. 747–755, 2011.

[60] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[61] M. ApS, *MOSEK Optimizer API for Python 9.2.37*, 2019. [Online]. Available: https://docs.mosek.com/9.2/pythonapi/index.html

[62] ASHRAE, "Ansi/ashrae standard 55-2013: Thermal environmental conditions for human occupancy," 2013.