Logarithmic Regret Bound in Partially Observable Linear Dynamical Systems

Sahin Lale[1], Kamyar Azizzadenesheli[2], Babak Hassibi[1], Anima Anandkumar[2]
[1] Department of Electrical Engineering
[2] Department of Computing and Mathematical Sciences
California Institute of Technology, Pasadena
{alale,kazizzad,hassibi,anima}@caltech.edu

Abstract

We study the problem of adaptive control in partially observable linear dynamical systems. We propose a novel algorithm, adaptive control online learning algorithm (AdaptOn), which efficiently explores the environment, estimates the system dynamics episodically and exploits these estimates to design effective controllers to minimize the cumulative costs. Through interaction with the environment, AdaptOn deploys online convex optimization to optimize the controller while simultaneously learning the system dynamics to improve the accuracy of controller updates. We show that when the cost functions are strongly convex, after T times step of agent-environment interaction, AdaptOn achieves regret upper bound of polylog (T). To the best of our knowledge, AdaptOn is the first algorithm which achieves polylog (T) regret in adaptive control of *unknown* partially observable linear dynamical systems which includes linear quadratic Gaussian (LQG) control.

1 Introduction

Reinforcement learning (RL) is the study of sequential decision making under uncertainty. One of the main and challenging problems in the field of RL is the design of algorithms to maximize/minimize given notions of reward/cost in a priori unknown environments [Bertsekas, 1995, Sutton and Barto, 2018]. Given an environment, the learning agent interacts with the environment, *explores* it to learn the environment behavior, and *exploits* the gathered experiences to improve the future performance [LaValle, 2006]. In order to assess the performance of an agent, we deploy a notion of *regret*, which is how much more cost the agent receives compared to the cost of an optimal policy [Cesa-Bianchi and Lugosi, 2006, Lai et al., 1982].

Adaptive control is one of the core problems in control theory and studies the problem of controlling unknown dynamical systems, [Stengel, 1994]. It has a long and extensive history of research from a variety of viewpoints. Robust control analyses controllers which are robust to the worst-case events under uncertainty, mainly in terms of $\mathcal{H}_2$ and $\mathcal{H}_\infty$ theory [Zhou et al., 1996, Hassibi et al., 1999]. In the case of learning through interaction, asymptotic optimality has been the topic of study for decades in order to improve performance [Lai et al., 1982, Lai and Wei, 1987].

Recent developments in the statistical learning theory [Peña et al., 2009], propose a set of tools to study finite time sample complexity of estimation methods. These methods have been levered to adaptive control to study the problem of sample complexity in fully observable linear systems [Fiechter, 1997, Abbasi-Yadkori and Szepesvári, 2011]. In the setting where the observations of the system's state evolution are partial and noisy, learning the dynamics of linear systems brings a series of challenges due to lack of direct access to the underlying events. For partially observable systems, a variety of methods have been proposed to learn the open-loop system dynamics via exciting the system with random and uncorrelated noise for long enough that the regression methods provide efficient estimations of the model parameters [Oymak and Ozay, 2018, Sarkar et al., 2019, Tsiamis and Pappas, 2019, Simchowitz et al., 2019].

Following these developments, a series of recent works advance these estimation procedures and propose explore-then-commit based methods, with guaranteed regret upper bounds of order $\tilde{\mathcal{O}}(T^{2/3})$ [Lale et al., 2020a, Simchowitz et al., 2020]. Here $\tilde{\mathcal{O}}(\cdot)$ presents the dominant time dependency. The model estimation procedures in these methods rely on an initial and long period of plain exploration using open-loop uncorrelated noise excitation. However, these methods do not generalize to the setting where the agents use the past observations and model estimations to derive better controllers, or even when the agents deploy plain closed-loop controllers. Recently, Lale et al. [2020b] proposed a novel model estimation method which lifts the mentioned limitation, does not rely on the initial long plain open-loop exploration, and can be used in interactive and adaptive learning paradigms. Lale et al. [2020b] deploy this estimator, along with the optimism in the face of uncertainty principle, and propose an interactive RL algorithm which achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T})$. One of the key elements of their algorithm, which allows such regret bound, is the possibility of continuously updating estimation of model parameters, yielding more and more accurate models, therefore, a better controller. While the mentioned work does not make strong convexity assumptions on the cost function, Simchowitz et al. [2020] shows that under this additional assumption, an explore-then-commit based approach can achieve a regret of $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$, even when the disturbances are semi-adversarial. The authors propose to deploy a random excitement open-loop controller for the long plain exploration phase, estimate the model parameters, and then exploit these estimations to run online convex optimization for regret minimization.

In this paper, we propose AdaptOn, **adapt**ive control **on**line learning algorithm that efficiently learns the model dynamics of the environment and optimizes for the controller to reduce the cumulative cost. An agent employing AdaptOn, adaptively learns the model dynamics through interaction with the environment, and deploys online convex optimization on a convex set of persistently exciting linear controllers, to gradually update the controller. We consider a general case where the learning agent need not have access to the cost function until committing its action. We show that when the cost functions are strongly convex, AdaptOn achieves a regret of $\operatorname{polylog}(T)$ after T time steps of environment-agent interaction. The regret analysis in this work is built on the top of the analyses in Lale et al. [2020b], Simchowitz et al. [2020], and Anava et al. [2015]. The proposed regret bound improves the prior work Lale et al. [2020b] when the cost functions are strongly convex, and advances the $\tilde{\mathcal{O}}\left(\sqrt{T}\right)$ regret in Simchowitz et al. [2020] in stochastic setting. To the best of our knowledge, this is the first logarithmic regret bound for partially observable linear dynamical systems when the dynamics are unknown a priori.

2 Preliminaries

We denote the Euclidean norm of a vector x as $\|x\|_2$. For a given matrix A, $\|A\|_2$ denotes the spectral norm, $\|A\|_F$ denotes the Frobenius norm, $A^\top$ is its transpose, $A^\dagger$ is the Moore-Penrose inverse, and $\mathrm{Tr}(A)$ gives the trace of matrix A. The j-th singular value of a rank-n matrix A is denoted by $\sigma_j(A)$, where $\sigma_{\max}(A) := \sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{\min}(A) := \sigma_n(A) > 0$. I is the identity matrix with an appropriate dimension. In the following, $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ.

Suppose we have an unknown discrete time linear time-invariant system $\Theta = (A, B, C)$ characterized as,

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + z_t, \end{aligned} \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the (latent) state of the system, $u_t \in \mathbb{R}^p$ is the control input, and the observation $y_t \in \mathbb{R}^m$ is the output of the system. Let $(\mathcal{F}_t; t \geq 0)$ be the corresponding filtration. For any t, w_t and z_t are σ_w^2-Gaussian and σ_z^2-Gaussian $\mathcal{F}_{t-1}$ measurable random vectors, respectively. In this paper, in contrast to the standard assumptions that the algorithm is given the knowledge of both σ_w^2 and σ_z^2 apriori, we assume that we only have the knowledge of their upper and lower bounds, i.e., $\overline{\sigma}_w^2, \underline{\sigma}_w^2, \overline{\sigma}_z^2$, and $\underline{\sigma}_z^2$, such that, $0 < \underline{\sigma}_w^2 \leq \sigma_w^2 \leq \overline{\sigma}_w^2$ and $0 < \underline{\sigma}_z^2 \leq \sigma_z^2 \leq \overline{\sigma}_z^2$, for some finite $\overline{\sigma}_w^2, \overline{\sigma}_z^2$. For the given system Θ, let Σ be the unique positive semidefinite solution to the following (Discrete Algebraic Riccati Equation) DARE:

$$\Sigma = A\Sigma A^\top - A\Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1} C\Sigma A^\top + \sigma_w^2 I. \tag{2}$$

Σ can be interpreted as the steady state error covariance matrix of state estimation under Θ. Let L denote the Kalman filter for the given system, where $L = \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}$.

The system characterization depicted in (1) is called state-space form of the system Θ. There are several ways to represent the same discrete time linear time-invariant system [Kailath et al., 2000, Tsiamis et al., 2019, Lale et al., 2020b]. One of the most common form is the predictor form[1] of the system characterized as

$$\begin{aligned} x_{t+1} &= \bar{A}x_t + Bu_t + Fy_t \\ y_t &= Cx_t + e_t \end{aligned} \tag{3}$$

where $F = AL$ is the Kalman gain in the observer form, e_t is the zero mean white innovation process and $\bar{A} = A - FC$. In this equivalent representation of system, the state x_t can be seen as the estimate of the state in the state space representation. In the steady state, $e(t) \sim \mathcal{N}\left(0, C\Sigma C^\top + \sigma_z^2 I\right)$. Notice that at the steady state, the current output y_t can be described by the history of inputs and outputs with an i.i.d. Gaussian disturbance e_t. Recall that the Kalman filter converges exponentially fast to the steady-state. Thus, without loss of generality, we assume that $x_0 \sim \mathcal{N}(0, \Sigma)$, *i.e.*, the system starts at the steady-state.

At each time step t, the system is at state x_t and the agent observes y_t, *i.e.*, an imperfect state information. Then, the agent applies a control input u_t, the agent pays the cost c_t, and the system evolves to a new x_{t+1} at time step $t+1$. At each time step t, the cost $c_t = \ell_t(y_t, u_t)$ where ℓ_t is smooth and strongly convex loss for all t, i.e., $0 \prec \underline{\alpha}_{loss} I \preceq \nabla^2 \ell_t(\cdot, \cdot) \preceq \overline{\alpha}_{loss} I$ for a finite constant $\overline{\alpha}_{loss}$. Note that the standard cases of regulatory costs of $\ell_t(y_t, u_t) = y_t^\top Q_t y_t + u_t^\top R_t u_t^\top$ with bounded positive definite matrices Q_t and R_t are special cases of the mentioned setting.

[1]For simplicity, the predictor form of system representation is presented for the steady-state of the system.

Assumption 2.1. *The unknown cost function of each times step t, $c_t = \ell_t(\cdot,\cdot)$, is non-negative strongly convex and associated with a parameter L, such that for any R with $\|u\|, \|u'\| \le R$, and $\|y\|, \|y'\| \le R$, we have,*

$$|\ell_t(y,u) - \ell_t(y',u')| \le LR(\|y-y'\| + \|u-u'\|) \quad \text{and} \quad |\ell_t(y,u)| \le LR^2.$$

Definition 2.1. *A linear system $\Theta = (A, B, C)$ is controllable if the controllability matrix*

$$[B \ \ AB \ \ A^2B \dots A^{n-1}B]$$

has full row rank. Similarly, a linear system $\Theta = (A, B, C)$ is observable if the observability matrix

$$[C^\top \ \ (CA)^\top \ \ (CA^2)^\top \dots (CA^{n-1})^\top]^\top$$

has full column rank.

Definition 2.2. *For any positive integer H, the H-length Markov parameters matrix is given as*

$$\mathbf{G}(H) = [G^{[0]} \ \ G^{[1]} \dots \ \ G^{[H-1]}] \in \mathbb{R}^{m \times Hp}.$$

$\mathbf{G}(H)$ is the length H impulse response of the system Θ. Moreover, the Markov parameters operator of the system Θ is defined using the set $\mathbf{G} = \{G^{[i]}\}_{i\ge 0}$ with $G^{[0]} = 0_{m\times p}$, and $\forall i > 0$, $G^{[i]} = CA^{i-1}B$.

Utilizing the definition of Markov parameters operator, we rewrite the observation at each time step t as follows;

$$y_t = z_t + \sum_{i=1}^{t-1} CA^{t-i-1}w_i + \sum_{i}^{t-1} G^{[i]}u_{t-i}. \tag{4}$$

Subtracting the controller contributing parts of y_t, we derive the Nature's y vector [Youla et al., 1976, Simchowitz et al., 2020].

Definition 2.3 (Nature's y)**.** *For a linear dynamical system Θ and Markov parameters operator $\mathbf{G}$, given a sequence of disturbances (w_i, z_i), $i \in \{1 \dots t\}$, Nature's y, $b_t(\mathbf{G})$, is defined as follows,*

$$b_t(\mathbf{G}) := y_t - \sum_{i}^{t-1} G^{[i]}u_{t-i} = z_t + \sum_{i=1}^{t-1} CA^{t-i-1}w_i.$$

It is the output of the system without the inputs applied until time step t.

We study the setting where the matrices A, B, and C, therefore the set of Markov parameters operator $\mathbf{G}$ of the system, are unknown. The agent interacts with this environment for T time steps and aims to minimize its cumulative cost $\sum_{t=1}^T c_t$. We consider the following problem setup.

Assumption 2.2. *The system is order n and stable, i.e. $\rho(A) < 1$, where $\rho(\cdot)$ denotes the spectral radius of a matrix which is the largest absolute value of its eigenvalues. Define $\Phi(A) = \sup_{\tau\ge 0} \frac{\|A^\tau\|}{\rho(A)^\tau}$. We assume that $\Phi(A) < \infty$ for the given system.*

The assumption regarding $\Phi(A)$ is a mild condition, *e.g.* if A is diagonalizable, $\Phi(A)$ is finite. Similar settings of study are also the main topic of interest in the recent literature [Oymak and Ozay, 2018, Sarkar et al., 2019, Tsiamis and Pappas, 2019, Simchowitz et al., 2020].

Assumption 2.3. *For the unknown system $\Theta = (A, B, C, F)$, (A, B) is controllable, (A, C) is observable and (A, F) is controllable. $\mathrm{Tr}(\mathbf{G}(H)^\top \mathbf{G}(H)) \le \kappa^2$ for some $\kappa \ge 0$, where $\mathbf{G}(H) \in \mathbb{R}^{m\times Hp}$ is the H-length Markov parameters matrix of system Θ. $\kappa_{\mathbf{G}}$ is an upper bound on the Markov parameters operator of the system Θ, i.e., $\sum_{i\ge 0} \|G^{[i]}\| \le \kappa_{\mathbf{G}}$ for $\kappa_{\mathbf{G}} \ge 1$.*

3 Linear Controller

A linear dynamical controller (LDC), π, is a s dimensional linear controller on a state $s_t^\pi \in \mathbb{R}^s$ of a linear dynamical system $(A_\pi, B_\pi, C_\pi, D_\pi)$, with input y_t^π and output u_t^π, where the state dynamics evolves as follows,

$$s_{t+1}^\pi = A_\pi s_t^\pi + B_\pi y_t^\pi,$$

with the controller given as

$$u_t^\pi = C_\pi s_t^\pi + D_\pi y_t^\pi.$$

Deploying a LDC policy π on the environment characterized with $\Theta = (A, B, C)$ induces the following joint dynamics of the x_t^π, s_t^π and the observation-action process:

$$\begin{bmatrix} x_{t+1}^\pi \\ s_{t+1}^\pi \end{bmatrix} = \underbrace{\begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}}_{A'_\pi} \begin{bmatrix} x_t^\pi \\ s_t^\pi \end{bmatrix} + \underbrace{\begin{bmatrix} I_{n\times n} & BD_\pi \\ 0_{s\times n} & B_\pi \end{bmatrix}}_{B'_\pi} \begin{bmatrix} w_t \\ z_t \end{bmatrix}$$

$$\begin{bmatrix} y_t^\pi \\ u_t^\pi \end{bmatrix} = \underbrace{\begin{bmatrix} C & 0_{s\times d} \\ D_\pi C & C_\pi \end{bmatrix}}_{C'_\pi} \begin{bmatrix} x_t^\pi \\ s_t^\pi \end{bmatrix} + \underbrace{\begin{bmatrix} 0_{d\times n} & I_{d\times d} \\ 0_{m\times n} & D_\pi \end{bmatrix}}_{D'_\pi} \begin{bmatrix} w_t \\ z_t \end{bmatrix}, \tag{5}$$

where $(A'_\pi, B'_\pi, C'_\pi, D'_\pi)$ are the associated parameters of induced closed loop system. Consider the Markov parameters operator of the system $(A'_\pi, B'_\pi, C'_\pi, D'_\pi)$, $\{{G'_\pi}^{[i]}\}_{i=0}$, as ${G'_\pi}^{[0]} = D'_\pi$, and $\forall i > 0$, ${G'_\pi}^{[i]} = C'_\pi {A'_\pi}^{i-1} B'_\pi$.

Let $B'_{\pi,w} := \begin{bmatrix} I_{n\times n} \\ 0_{s\times n} \end{bmatrix}$, and $B'_{\pi,z} := \begin{bmatrix} BD_\pi \\ B_\pi \end{bmatrix}$, the columns of B'_π applied on process noise, and measurement noise respectively. Similarly $C'_{\pi,y} := \begin{bmatrix} C & 0_{s\times d} \end{bmatrix}$ and $C'_{\pi,u} := \begin{bmatrix} D_\pi C & C_\pi \end{bmatrix}$ are rows of C'_π generating the observation and action.

Let $\psi : \mathbb{N} \to \mathbb{R}_{\geq 0}$ be a proper decay function, such that ψ is non-increasing and $\lim_{h'\to\infty} \psi(h') = 0$. For a Markov operator $\mathbf{G}$, $\psi_{\mathbf{G}}(h)$ defines the induced decay function on $\mathbf{G}$, *i.e.*, $\psi_{\mathbf{G}}(h) := \sum_{i\geq h} \|G^{[i]}\|$. $\Pi(\psi)$ denotes the class of LDC policies associated with a proper decay function ψ, such that for all $\pi \in \Pi(\psi)$, and all $h \geq 0$, $\sum_{i\geq h} \|{G'_\pi}^{[i]}\| \leq \psi(h)$. Let $\kappa_\psi := \psi(0)$.

In this work, we adopt disturbance feedback controllers (DFCs), truncated approximations to LDC policies. A DFC policy of depth H' is defined with a set of parameters $\mathbf{M}(H') := \{M^{[i]}\}_{i=0}^{H'-1}$ and Nature's y, which prescribe the control input of

$$u_t^{\mathbf{M}} = \sum_{i=0}^{H'-1} M^{[i]} b_{t-i}(\mathbf{G}).$$

and results in state $x_{t+1}^{\mathbf{M}}$ and observation $y_{t+1}^{\mathbf{M}}$. The following lemma shows that Nature's y, $b_t(\mathbf{G})$, is uniformly bounded throughout the interaction with the system.

Lemma 3.1 (Bounded Nature's y)**.** *For all $t \in [T]$, the following holds with probability at least $1 - \delta$,*

$$\|b_t(\mathbf{G})\| \leq \kappa_b := \overline{\sigma}_z \sqrt{2m \log \frac{4mT}{\delta}} + \frac{\overline{\sigma}_w \Phi(A) \|C\| \sqrt{2n \log \frac{4nT}{\delta}}}{1 - \rho(A)}. \tag{6}$$

The proof is given in the Appendix A.1. In the following, directly using the analysis in Simchowitz et al. [2020], we show that for any $\pi \in \Pi(\psi)$ and any input u_t^π at time step t, there is a parameters set $\mathbf{M}$, such that $u_t^{\mathbf{M}}$ is sufficiently close to u_t^π, and the resulting y_t^π is sufficiently close to $y_t^{\mathbf{M}}$.

Lemma 3.2. *For any policy* LDC *policy* $\pi \in \Pi(\psi)$*, there exist a* H' *length* DFC *policy* $\mathbf{M}(H')$ *such that, with probability at least* $1-\delta$*,*

$$\begin{aligned}
\|u_t^\pi - u_t^{\mathbf{M}}\| &\leq \left\| \sum_{i=H'}^{t} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i} \right\| \leq \psi(H')\kappa_b \\
\|y_t^\pi - y_t^{\mathbf{M}}\| &\leq \psi(H')\kappa_{\mathbf{G}}\kappa_b
\end{aligned}$$

and one of the DFC *policies that satisfies these conditions is* $M^{[0]} = D_\pi$*, and* $M^{[i]} = C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z}$ *for all* $0 < i < H'$*.*

The proof is provided in Appendix A.2. Lemma 3.2 further entails that any stabilizing LDC can be well approximated by a DFC that belongs to the following class of DFCs

$$\mathcal{M}\left(H', \kappa_\psi\right) = \left\{ \mathbf{M}(H') := \{M^{[i]}\}_{i=0}^{H'-1} : \sum_{i\geq 0}^{H'-1} \|M^{[i]}\| \leq \kappa_\psi \right\}.$$

This observation indicates that using the set of DFC for the policy design does not, considerably, alter the performance of the LDC controllers.

4 Regret Analysis

We evaluate the agent's performance by its regret with respect to $\pi_\star$, which is the optimal policy for infinite horizon control problem,

$$\pi_\star(\Theta) = \lim_{T\to\infty} \arg\min_{\pi\in\Pi(\psi)} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(y_t, u_t^\pi)\right].$$

After T time step of agent-environment interaction, we consider agent's regret $\textsc{Regret}(T)$ as follows,

$$\textsc{Regret}(T) = \sum_{t=1}^{T} c_t - \sum_{t=1}^{T} \ell(y^{\pi_\star}, u^{\pi_\star}).$$

Throughout the interaction with the system, the agent has access to a convex set of DFCs, $\mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$, such that $\kappa_{\mathcal{M}} \geq 2\kappa_\psi$ and all controllers $\mathbf{M} \in \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$ is persistently exciting on the system Θ. The precise definition of persistence of excitation condition is given in Appendix C.3. Furthermore, we assume that $\mathbf{M}_\star(H') := \{{M_\star}^{[i]}\}_{i=0}^{H'-1}$, a DFC approximation of $\pi_\star$ with ${M_\star}^{[0]} = D_{\pi_\star}$, and ${M_\star}^{[i]} = C'_{\pi_\star,u} {A'_{\pi_\star}}^{i-1} B'_{\pi_\star,z}$ for all $0 < i < H'$, is contained in $\mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$.

4.1 AdaptOn

We propose AdaptOn, a sample efficient **adapt**ive control **on**line learning algorithm which learns the model dynamics through interaction with the environment and simultaneously deploys online convex optimization approach to optimize the control policy. AdaptOn is illustrated in Algorithm 1. AdaptOn uses $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ to excite the system for a fixed warm-up period of $T_{burn} \geq T_{\max}$, where

$$T_{\max} \coloneqq \max\{H, H', T_o, T_A, T_B, T_c, T_{\epsilon_G}, T_{cl}, T_{cx}\}. \tag{7}$$

The duration of the warm-up period is chosen to guarantee an accountable first estimate of the underlying system (T_o), the stability of the online learning algorithm on the underlying system (T_A, T_B), the stability of the inputs and outputs ($T_{\epsilon_\mathbf{G}}$), the persistence of excitation during the adaptive control period (T_{cl}), an accountable estimate at the first epoch of adaptive control (T_c), and the conditional strong convexity of expected counterfactual losses (T_{cx}). The precise expressions are given in the Appendix.

In the adaptive control period, AdaptOn operates in epochs with doubling length. We set the base period T_{base} to an initial value $T_{base} = T_{burn}$. After the warm-up period, in the first epoch of adaptive control, the agent runs for T_{base} time steps. For each remaining epoch i, the AdaptOn runs for $2^{i-1}T_{base}$ time steps. At the beginning of epoch i, i.e., at time step t_i, AdaptOn exploits the past experiences up to the i'th epoch and estimates $\widehat{\mathbf{G}}_i(H)$, the first H Markov parameters of the environment. AdaptOn utilizes these estimates to approximate $b_t(\mathbf{G})$ using $\widehat{\mathbf{G}}_i(H)$. During the i'th epoch, at any time step $t \in [t_i, \ldots, t_{i+1} - 1]$, AdaptOn computes

$$b_t(\widehat{\mathbf{G}}_i) = y_t - \sum_{j=1}^{H} \widehat{G}_i^{[j]} u_{t-j}, \tag{8}$$

and using the approximations of Nature's y, AdaptOn executes a DFC policy $\mathbf{M}_t \in \mathcal{M}(H', \kappa_\mathcal{M})$, $u_t^{\mathbf{M}_t} = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\widehat{\mathbf{G}}_i)$ and observes the loss function ℓ_t. Using the estimate $\widehat{\mathbf{G}}_i(H)$, we define the counterfactual loss at time step t as follows,

$$f_t\left(\mathbf{M}_t, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right) = \ell_t\left(b_t(\widehat{\mathbf{G}}_i) + \sum_{j=1}^{H} \widehat{G}_i^{[j]} \sum_{l=0}^{H'-1} M^{[l]} b_{t-j-l}(\widehat{\mathbf{G}}_i), \sum_{l=0}^{H'-1} M^{[l]} b_{t-l}(\widehat{\mathbf{G}}_i)\right).$$

Note that the counterfactual loss is convex in $\mathbf{G}$. During the time steps in epoch i, $t \in [t_i, \ldots, t_{i+1} - 1]$, AdaptOn applies steps of online learning on this cost function while keeping updates in the set $\mathcal{M}(H', \kappa_\mathcal{M})$ via projection.

4.2 Dynamics Learning

In order to estimate the Markov parameters, AdaptOn follows the estimation process provided in Lale et al. [2020b] which uses the predictor form of the state-space representation given in (3). Using the generated input-output sequence $\mathcal{D}_\tau = \{y_t, u_t\}_{t=1}^{\tau}$, the agent constructs N subsequences of H_{est} input-output pairs, ϕ_t for $H_{est} \leq t \leq \tau$, where $\tau = H_{est} + N - 1$,

$$\phi_t = \left[y_{t-1}^\top \ldots y_{t-H_{est}}^\top \;\; u_{t-1}^\top \ldots \;\; u_{t-H_{est}}^\top\right]^\top \in \mathbb{R}^{(m+p)H_{est}}.$$

Algorithm 1 AdaptOn

Input: T, H, H', T_{burn}, $\mathcal{M} = \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$

—— Warm-Up ——

for $t = 1, \ldots, T_{burn}$ **do**

  Deploy $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$

end for

Store $\mathcal{D}_{T_{burn}} = \{y_t, u_t\}_{t=1}^{T_{burn}}$,

Set $T_{base} = T_{burn}$, $t = T_{base} + 1$, and , $t_1 = T_{base}$

—— Adaptive Control ——

Set $\mathbf{M}_t$ as any member of $\mathcal{M}$

for $i = 1, 2, \ldots$ **do**

  Solve (11) using $\mathcal{D}_t$, estimate $\hat{A}_i, \hat{B}_i, \hat{C}_i$ using SysId and construct $\widehat{\mathbf{G}}_i(H)$

  Compute $b_\tau(\widehat{\mathbf{G}}_i) := y_\tau - \sum_{j=1}^{H} \widehat{G}_i^{[j]} u_{\tau-j}$, $\forall \tau \leq t$

  while $t \leq t_i + 2^{i-1} T_{base}$ *and* $t \leq T$ **do**

    Observe y_t, and compute $b_t(\widehat{\mathbf{G}}_i) := y_t - \sum_{j=1}^{H} \widehat{G}_i^{[j]} u_{t-j}$

    Commit to $u_t := u_t^{\mathbf{M}_t} = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\widehat{\mathbf{G}}_i)$, observes ℓ_t, and pays a cost of $\ell_t(y_t, u_t)$

    Update $\mathbf{M}_{t+1} = proj_{\mathcal{M}} \left(\mathbf{M}_t - \eta_t \nabla f_t \left(\mathbf{M}_t, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i) \right) \right)$

    $\mathcal{D}_{t+1} = \mathcal{D}_t \cup \{y_t, u_t\}$

    $t = t + 1$

  end while

  $t_{i+1} = t_i + 2^{i-1} T_{base}$

end for

One can write the following truncated autoregressive exogenous (ARX) model for the given system Θ,

$$y_t = \mathcal{G}_{\mathbf{yu}} \phi_t + e_t + C \bar{A}^H x_{t-H_{est}}$$

where $\mathcal{G}_{\mathbf{yu}} \in \mathbb{R}^{m \times (m+p) H_{est}}$ defined as

$$\mathcal{G}_{\mathbf{yu}} = \left[CF,\ C\bar{A}F,\ \ldots,\ C\bar{A}^{H_{est}-1}F,\ CB,\ C\bar{A}B,\ \ldots,\ C\bar{A}^{H_{est}-1}B \right]. \tag{9}$$

Using this, we have the following form for any input-output trajectory $\{y_i, u_t\}_{t=1}^{\tau}$ can be represented as

$$Y_\tau = \Phi_\tau {\mathcal{G}_{\mathbf{yu}}}^\top + E_\tau + N_\tau \tag{10}$$

where

$$Y_\tau = \left[y_{H_{est}}, y_{H_{est}+1}, \ldots, y_\tau \right]^\top \in \mathbb{R}^{N \times m} \quad \Phi_\tau = \left[\phi_{H_{est}}, \phi_{H_{est}+1}, \ldots, \phi_\tau \right]^\top \in \mathbb{R}^{N \times (m+p) H_{est}}$$

$$E_\tau = \left[e_{H_{est}}, e_{H_{est}+1}, \ldots, e_\tau \right]^\top \in \mathbb{R}^{N \times m} \quad N_\tau = \left[C\bar{A}^{H_{est}} x_0, C\bar{A}^{H_{est}} x_1, \ldots, C\bar{A}^{H_{est}} x_{\tau - H_{est}} \right]^\top \in \mathbb{R}^{N \times m}$$

for $N = \tau - H_{est} + 1$. After the warm-up period and before the first epoch, AdaptOn obtains the first estimate of the unknown truncated ARX model $\mathcal{G}_{\mathbf{yu}}$ by solving the following regularized least square problem for $i = 1$,

$$\widehat{\mathcal{G}}_{\mathbf{yu,i}} = \arg\min_X \| Y_{t_i} - \Phi_{t_i} X^\top \|_F^2 + \lambda \| X \|_F^2 \tag{11}$$

where the solution

$$\widehat{\mathcal{G}}_{\mathbf{yu,i}}^{\top} = (\Phi_{t_i}^{\top}\Phi_{t_i} + \lambda I)^{-1}\Phi_{t_i}^{\top} Y_{t_i}.$$

Using this solution, AdaptOn deploys SysId, a system-identification algorithm given in Appendix B. SysId uses the blocks of the estimate $\widehat{\mathcal{G}}_{\mathbf{yu,1}}$ to form two Hankel matrices and concatenate them to construct, $\hat{\mathcal{H}}_1$. Using Definition 2.1, if the input to the SysId was $\mathcal{G}_{\mathbf{yu}}$ then the constructed matrix, $\mathcal{H}$, would be rank n, where $\|\mathcal{H}\|$ denotes the spectral norm and $\sigma_n(\mathcal{H}) > 0$ denotes n'th singular value of $\mathcal{H}$. From $\hat{\mathcal{H}}_1$, SysId obtains the estimates of the system parameters $\hat{A}_1, \hat{B}_1, \hat{C}_1$. For more details of SysId refer to Lale et al. [2020b]. Finally, from these estimates, AdaptOn forms the estimate $\widehat{\mathbf{G}}_1(H)$. As explained in the previous section, this estimation process is repeated in the beginning of each epoch by using all the data gathered.

First, consider the effect of truncation bias term, N_{t_i}. From Assumption 2.3, there exists a similarity transformation that gives $\|\bar{A}\| \leq \upsilon < 1$. Thus, each term in N_{t_i} is order of υ^H. In order to get consistent estimation, for some problem dependent constant c_H, AdaptOn sets $H_{est} \geq \max\{2n+1, \frac{\log(c_H T^2\sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}\}$, resulting in a negligible bias term of order $1/T^2$. Now, consider H_{est}-length truncated open-loop noise evolution parameters, $\mathcal{G}^{ol}$ as defined in Appendix A of Lale et al. [2020b]. Let σ_o denote a lower bound on $\sigma_{\min}(\mathcal{G}^{ol})$, *i.e.*, $\sigma_{\min}(\mathcal{G}^{ol}) > \sigma_o > 0$. Similarly, due to persistence of excitation of all $\mathbf{M} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$, defined in detail in Appendix C.3, during the adaptive control period we have the lower bound, σ_c, on the smallest singular value of the matrix that generates ϕ_t from system disturbances $w_{1:t}, z_{1:t}$. Let

$$\sigma_\star^2 := \min\left\{\frac{\sigma_o^2\underline{\sigma}_w^2}{2}, \frac{\sigma_o^2\underline{\sigma}_z^2}{2}, \frac{\sigma_o^2\sigma_y^2}{2}, \frac{\sigma_c^2\underline{\sigma}_w^2}{16}, \frac{\sigma_c^2\underline{\sigma}_z^2}{16}\right\}. \tag{12}$$

The following is an adaptation of Theorem 3.3 of Lale et al. [2020b] to the given setting using the problem dependent parameters $\Upsilon_w(\delta) = poly\left(\|C\|, \overline{\sigma}_z, \overline{\sigma}_w, \sigma_u, \sqrt{m}, \sqrt{n}, \sqrt{p}, \sqrt{\log(T_{burn}/\delta)}\right)$ and $\Upsilon_c(\delta) = poly\left(\kappa_{\mathcal{M}}, \kappa_b, \kappa_{\mathbf{G}}, \sqrt{m}\right)$.

Theorem 4.1. *[Estimation of Truncated ARX Model After Warm-up] Let $\delta \in (0,1)$. During the warm-up period, for all $t \leq T_{burn}$, $\|\phi_t\| \leq \Upsilon_w(\delta)\sqrt{H_{est}}$ and $\|u_t\| \leq \kappa_{u_{burn}} := \sigma_u\sqrt{2p\log(2pT_{burn}/\delta)}$ with probability at least $1-\delta$. Let $\|\mathcal{G}_{\mathbf{yu}}\|_F \leq S$. After the warm-up period of $T_{burn} \geq T_{\max}$, the first estimate of truncated ARX model, $\widehat{\mathcal{G}}_{\mathbf{yu,1}}$, obeys the following with probability at least $1-2\delta$,*

$$\|\widehat{\mathcal{G}}_{\mathbf{yu,1}} - \mathcal{G}_{\mathbf{yu}}\| \leq \frac{\kappa_{est}}{\sigma_\star\sqrt{T_{burn}}},$$

where $\kappa_{est} = \sqrt{m\|C\Sigma C^\top + \overline{\sigma}_z^2 I\|\left(\log(\frac{1}{\delta}) + \frac{H_{est}(m+p)}{2}\log\left(1 + \frac{T\max\{\Upsilon_w^2(\delta), \Upsilon_c^2(\delta)\}}{\lambda(m+p)}\right)\right)} + S\sqrt{\lambda} + \frac{\sqrt{H_{est}}}{T}$.

The proof is given in Appendix C.1, where we show the persistence of excitation of the inputs in the warm-up period and use this to derive the presented bound via Theorem E.2. Define the following quantities,

$$\gamma_{\mathbf{G}} = (\|B\|+\|C\|+1)\left(1+\frac{\Phi(A)}{1-\rho(A)}+\frac{2\Phi(A)}{(1-\rho(A))^2}\right)+\frac{2\Phi(A)}{(1-\rho(A))^2}\|C\|\|B\|$$

$$\gamma_{\mathcal{H}} = \frac{\sqrt{nH_{est}}(\|\mathcal{H}\|+\sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})}$$

$$\alpha \leq \underline{\alpha}_{loss}\left(\sigma_z^2+\sigma_w^2\left(\frac{\sigma_{\min}(C)}{1+\|A\|^2}\right)^2\right)$$

where α is a lower bound to the strong convexity dependent parameter. The following lemma shows that warm-up duration was long enough to obtain good initial Markov parameter estimates.

Lemma 4.1. *[Estimation Error in First Epoch] Let $\delta \in (0,1)$, $T > T_{burn} \geq T_{\max}$ and $\psi_{\mathbf{G}}(H+1) \leq 1/10T$. In the first epoch of the adaptive control period, at any time step $t \in [T_{base}, \dots, 2T_{base}-1]$, with probability at least $1-2\delta$, Markov parameter estimation error of* AdaptOn *is bounded as*

$$\sum_{j\geq 1}\|\widehat{G}_1^{[j]} - G^{[j]}\| \leq \epsilon_{\mathbf{G}}(1,\delta) \leq \max\left\{\frac{1}{4\kappa_b\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}}\sqrt{\frac{\alpha}{H'\underline{\alpha}_{loss}}}, \frac{1}{2\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}}\right\},$$

where $\epsilon_{\mathbf{G}}(1,\delta) := \frac{2c_1\gamma_{\mathbf{G}}\gamma_{\mathcal{H}}\kappa_{est}}{\sigma_\star\sqrt{T_{burn}}}$ for some problem dependent constant c_1.

The proof is given in Appendix C.2. To give an overview, we combine Theorem 4.1 with Theorem E.3, which translates $\|\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}\|$ to estimation error bounds on the system parameter estimates obtained by SysId, to get the guarantee for the Markov parameter estimates that hold during the first epoch of adaptive control period. Notice that for the first epoch, $\epsilon_{\mathbf{G}}(1,\delta) = \mathcal{O}\left(\frac{polylog(T)}{\sqrt{T_{burn}}}\right)$. Using Lemma 4.1, we show that with the given T_{burn} duration, the Markov parameter estimates are well-refined such that the inputs, outputs and Nature's y estimates of AdaptOn are bounded uniformly, *i.e.*, the system remains stable.

Lemma 4.2 (Lemma 6.1 in [Simchowitz et al., 2020])**.** *Let $T > T_{burn} \geq T_{\max}$ and $\psi_{\mathbf{G}}(H+1) \leq 1/10T$ for $\delta \in (0,1)$. Then, for all $t \in [T]$, with probability at least $1-2\delta$, the following holds,*

$$\|u_t\| \leq \kappa_u := 2\max\{\kappa_{u_{burn}},\ \kappa_{\mathcal{M}}\kappa_b\}$$
$$\|b_t(\widehat{\mathbf{G}})\| \leq 2\kappa_b$$
$$\|y_t\| \leq \kappa_y := \kappa_b + \kappa_{\mathbf{G}}\kappa_u$$

Next, we consider the concentration of truncated ARX Model estimates during the adaptive control period. The following shows that the estimation error of closed-loop truncated ARX Model estimates have the same characteristics (decay) with the open-loop estimates, *i.e.* Theorem 4.1.

Theorem 4.2. *[Estimation of Truncated ARX Model During Adaptive Control] During the adaptive control period for all $t > T_{burn} \geq T_{\max}$, $\|\phi_t\| \leq \Upsilon_c(\delta)\sqrt{H_{est}}$ with probability $1-2\delta$. Let $\|\mathcal{G}_{\mathbf{yu}}\|_F \leq S$. Then for all $i = \{1,2,\dots\}$, at the beginning of i'th epoch the estimate of the truncated ARX model, $\widehat{\mathcal{G}}_{\mathbf{yu,i}}$, obeys the following with $1-4\delta$,*

$$\|\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}\| \leq \frac{\kappa_{est}}{\sigma_\star\sqrt{t_i}}.$$

The proof is given in Appendix C.4, but here we provide a proof sketch. We first show that the initial estimation error given in Lemma 4.1 is small enough to provide persistence of excitation in expectation. Then using Theorem E.1 and a standard covering argument, we show that AdaptOn has persistence of excitation during the adaptive control period with high probability. Combining this result with Theorem E.2, we derive Theorem 4.2. Similar to Lemma 4.1, we have the following lemma which extends $\|\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}\|$ to estimation error in Markov parameters.

Lemma 4.3. *[Estimation Error in All Epochs] Let $\delta \in (0,1)$, $T > T_{burn} \geq T_{\max}$ and $\psi_{\mathbf{G}}(H+1) \leq 1/10T$. During the i'th epoch of adaptive control period, at any time step $t \in [t_i, \ \ldots, \ t_{i+1}-1]$, with probability at least $1-4\delta$, for all i, Markov parameter estimation error of* AdaptOn *is bounded as*

$$\sum_{j\geq 1} \|\widehat{G}_i^{[j]} - G^{[j]}\| \leq \epsilon_{\mathbf{G}}(i,\delta) \coloneqq \frac{2c_1 \gamma_{\mathbf{G}} \gamma_{\mathcal{H}} \kappa_{est}}{\sigma_\star \sqrt{t_i}},$$

for some problem dependent constant c_1.

The proof of Lemma 4.3 is identical with Lemma 4.1. Notice that for the i'th epoch, $\epsilon_{\mathbf{G}}(i,\delta) = \mathcal{O}\left(\frac{polylog(T)}{\sqrt{t_i}}\right)$. This observation will be key in proving the main result of this work, Theorem 4.4.

4.3 Regret Bound

Using the guarantees in learning the system dynamics, we obtain the following regret upper bound of AdaptOn.

Theorem 4.3. *Let the decision makers memory H' satisfy $H' \geq 3H \geq 1$, $\psi(\lfloor H'/2 \rfloor - H) \leq \kappa_{\mathcal{M}}/T$ and $\psi(H+1) \leq 1/10T$. Under the Assumptions 2.1-2.3, after a warm-up period time $T_{burn} \geq T_{\max}$, if* AdaptOn *runs with step size $\eta_t = \frac{12}{\alpha t}$, then with probability at least $1-5\delta$, the regret of* AdaptOn *is bounded as follows*

$$\textsc{Regret}(T) \lesssim \underbrace{T_{burn} L \kappa_y^2}_{R_1} + \underbrace{\frac{L^2 {H'}^3 \min\{m,p\} \kappa_b^4 \kappa_{\mathbf{G}}^4 \kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2 \kappa_{\mathbf{G}}^2\}} \left(1 + \frac{\overline{\alpha}_{loss}}{\min\{m,p\} L \kappa_{\mathcal{M}}}\right) \log\left(\frac{T}{\delta}\right)}_{R_2}$$
$$+ \underbrace{\sum_{t=T_{burn}+1}^{T} \epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right) H' \kappa_b^2 \kappa_{\mathcal{M}}^2 \left(\frac{\kappa_{\mathbf{G}}^2 \kappa_b^2 \left(\overline{\alpha}_{loss} + L\right)^2}{\alpha} + \kappa_y^2 \max\left\{L, \frac{L^2}{\alpha}\right\}\right)}_{R_3}.$$

The proof is given in Appendix D. The proof follows and adapts Theorem 5 of Simchowitz et al. [2020] to the setting of AdaptOn. The key difference is that the Markov parameter estimation errors are not fixed during the adaptive control period and they decay in each epoch due to closed-loop model estimation that AdaptOn runs. Note that regret upper bound is composed of fairly accessible terms. The first term is the regret obtained during the warm-up period and the second term is the regret of online learning controller. Finally, the last term is due to Markov parameter estimation errors. However, notice that the regret scales quadratically with the estimation error, due to strong convexity of the loss function.

Following the doubling update rule of AdaptOn, the length of each epoch grows as T_{base}, $2T_{base}$, $4T_{base}$, ... Therefore after T times steps of agent-environment interaction, the number of epochs, *i.e.* the number of times AdaptOn estimates the first H Markov parameters is $\mathcal{O}(\log T)$.

As indicated in Lemma 4.1 and Lemma 4.3, during the i'th epoch of the adaptive control period, at any time step $t \in [t_i, \dots, t_i - 1]$, $\epsilon^2_{\mathbf{G}}\Big(\Big\lceil \log_2\left(\frac{t}{T_{burn}}\right)\Big\rceil, \delta\Big)$ is $\mathcal{O}\left(\frac{polylog(T)}{t_i}\right)$. Following the update rule of AdaptOn, we have

$$\sum_{t=T_{base}+1}^{T} \epsilon^2_{\mathbf{G}}\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right) = \sum_{i=1}^{\mathcal{O}(\log T)} 2^{i-1} T_{base} \epsilon^2_{\mathbf{G}}(i, \delta)$$

$$\leq \sum_{i=1}^{\mathcal{O}(\log T)} 2^{i-1} T_{base}\ \mathcal{O}\left(\frac{polylog(T)}{2^{i-1}T_{base}}\right) = \mathcal{O}\left(polylog(T)\right) \tag{13}$$

Using the result of (13), we can upper bound the R_3 of the regret upper bound in Theorem 4.3 with a $polylog(T)$ bound which entails the following polylogarithmic regret:

Theorem 4.4. *Under the conditions of the Theorem 4.3, the regret of* AdaptOn *is bounded as follows:*

$$\textsc{Regret}(T) = polylog(T).$$

Note that without any estimation updates during the adaptive control, AdaptOn reduces to a variant of the algorithm given in Simchowitz et al. [2020]. While the update rule in AdaptOn results in $\mathcal{O}(\log(T))$ updates in adaptive control period, one can follow different update schemes as long as AdaptOn obtains enough samples in the beginning of the adaptive control period to obtain persistence of excitation. The following is an immediate corollary of Theorem 4.4 which considers the case when number of epochs or estimations are limited during the adaptive control period.

Corollary 4.1. *As long as enough samples are gathered in the adaptive control period, with any update scheme less than* $\log(T)$ *updates during adaptive control period, the regret of* AdaptOn *is bounded as follows:*

$$\textsc{Regret}(T) \in \left[polylog(T), \tilde{O}(\sqrt{T})\right].$$

Now consider the case where the condition on persistence of excitation of $\mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$ does not hold. In order to efficiently learn the model parameters and minimize the regret, one can add an additional independent Gaussian excitation to the control input u_t for each time step t. This guarantees the concentration of Markov parameter estimates, but it also results in an extra regret term in the bound of Theorem 4.3.

If the variance of the added Gaussian vector is set to be $\widetilde{\sigma}^2$, exploiting the Lipschitzness of the loss functions, the additive regret of the random excitation is $\tilde{\mathcal{O}}(T\widetilde{\sigma})$. Following the results in Lemma 4.3, the additional random excitation helps in parameter estimation and concentration of Markov parameters up to the error of $\mathcal{O}(polylog(T)/\sqrt{\widetilde{\sigma}^2 t})$. Since the contribution of the error in the Markov parameter estimates in the Theorem 4.3 is quadratic, the contribution of this error in the regret through R_3 will be $\mathcal{O}(polylog(T)/\sigma^2)$.

Corollary 4.2. *When the condition on persistent excitation of all* $\mathbf{M} \in \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$ *is not fulfilled, adding independent Gaussian vectors with variance of* $\mathcal{O}(1/T^{1/3})$ *to the inputs in adaptive control period results in the regret upper bound of* $\tilde{\mathcal{O}}(T^{2/3})$.

5 Related Works

Adaptive control arises when there is uncertainty in the system model. In order to achieve good performance, the agent needs to learn the dynamics by interacting with the system and adapt accordingly based on its observations. For fully observable linear systems, Lai et al. [1982], Chen and Guo [1987] provide asymptotic analysis of consistency in the model estimation, which is based on pure exploration.

In order to capture the control objective while learning the system dynamics in finite time, Abbasi-Yadkori and Szepesvári [2011] used regret as the performance metric to provide $\tilde{\mathcal{O}}(\sqrt{T})$ regret in the adaptive control of linear quadratic regulator (LQR). Their method builds upon optimism in the face of uncertainty (OFU) principle and self-normalized estimations [Abbasi-Yadkori et al., 2011a,b]. This work sparked the flurry of research with different directions in the regret analysis of controlling unknown LQR [Faradonbeh et al., 2017, Abeille and Lazaric, 2017, 2018, Dean et al., 2018, Faradonbeh et al., 2018, Cohen et al., 2019, Mania et al., 2019, Abbasi-Yadkori et al., 2019]. These works consider the systems with stochastic noise. Recently, Cassel et al. [2020] show that logarithmic regret is achievable if only A or B is unknown in LQR. Moreover, Simchowitz and Foster [2020] recently provide $\tilde{\mathcal{O}}(\sqrt{T})$ regret lower bound for LQR setting with the fully unknown system. They show that a slight deviation in the input matrix causes in an ambiguity in learning the model for the agent that competes against an oracle and this ambiguity prevents logarithmic regret. Note that, due to the persistent noise in the observations of the hidden states in partially observable linear dynamical systems, the mentioned lower bound does not carry to the provided guarantee of AdaptOn.

In the adversarial noise setting, most of the works consider full information of the underlying system and aim to control the system under adversarial noise [Agarwal et al., 2019a, Cohen et al., 2018, Agarwal et al., 2019b, Foster and Simchowitz, 2020]. Recent efforts extend to adaptive control in the adversarial setting for the unknown system model [Hazan et al., 2019, Simchowitz et al., 2020].

In the partially observable linear systems, similar to the trend in fully observable counterparts, most of the prior works focus on the system identification aspects [Ljung, 1999, Chen et al., 1992, Juang et al., 1993, Phan et al., 1994, Lee and Zhang, 2019, Oymak and Ozay, 2018, Sarkar et al., 2019, Simchowitz et al., 2019, Lee and Lamperski, 2019, Tsiamis and Pappas, 2019, Tsiamis et al., 2019, Umenberger et al., 2019, Tsiamis and Pappas, 2020]. A body of work aimed to extend the problem of estimation and prediction to online convex optimization where a set of strong theoretical guarantees on cumulative prediction errors are provided [Abbasi-Yadkori et al., 2014, Hazan et al., 2017, Arora et al., 2018, Hazan et al., 2018, Lee and Zhang, 2019, Ghai et al., 2020]

Building upon the system identification algorithms, Lale et al. [2020a] provides $\tilde{\mathcal{O}}(T^{2/3})$ regret upper bound in system with stochastic noise using OFU. Similarly building upon the system identification algorithms, Simchowitz et al. [2020] uses online convex optimization [Anava et al., 2015] to achieve $\tilde{\mathcal{O}}(T^{2/3})$ for convex cost functions and $\tilde{\mathcal{O}}(\sqrt{T})$ for strongly convex cost functions. Exploiting the different model representations of partially observable linear models, Lale et al. [2020b] devise a new system identification method and provide $\tilde{\mathcal{O}}(\sqrt{T})$ regret in stochastic setting without strong convexity assumption.

6 Conclusion

In this paper, we propose AdaptOn, a novel adaptive control algorithm that efficiently learns the truncated Markov parameters of the underlying dynamical system and deploys projected online gradient descent to design a controller. The design of AdaptOn is developed based on the recent novel studies on RL in partially observable dynamical systems [Simchowitz et al., 2020, Lale et al., 2020b]. We show that in the presence of convex set of persistently exciting linear controllers and strongly convex loss functions, AdaptOn achieves a regret upper bound of polylogarithmic in number of agent-environment interactions.

In this work, we relaxed the requirement in a priori knowledge of the variance of the Gaussian process noise, and measurement noise to just their upper and lower bounds. For the future work, we plan to extend the study of AdaptOn to more general sub-Gaussian noise, and potentially to adversarial perturbations [Simchowitz et al., 2020].

Acknowledgements

S. Lale is supported in part by DARPA PAI. K. Azizzadenesheli gratefully acknowledge the financial support of Raytheon and Amazon Web Services. B. Hassibi is supported in part by the National Science Foundation under grants CNS-0932428, CCF-1018927, CCF-1423663 and CCF-1409204, by a grant from Qualcomm Inc., by NASA's Jet Propulsion Laboratory through the President and Director's Fund, and by King Abdullah University of Science and Technology. A. Anandkumar is supported in part by Bren endowed chair, DARPA PAIHR00111890035 and LwLL grants, Raytheon, Microsoft, Google, and Adobe faculty fellowships.

References

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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011b.

Yasin Abbasi-Yadkori, Peter Bartlett, and Varun Kanade. Tracking adversarial targets. In *International Conference on Machine Learning*, pages 369–377, 2014.

Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3108–3117, 2019.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. *arXiv preprint arXiv:1703.08972*, 2017.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham M Kakade, and Karan Singh. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019a.

Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. In *Advances in Neural Information Processing Systems*, pages 10175–10184, 2019b.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: price of past mistakes. In *Advances in Neural Information Processing Systems*, pages 784–792, 2015.

Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems. 2018.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena scientific Belmont, MA, 1995.

Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. *arXiv preprint arXiv:2002.08095*, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Chung-Wen Chen, Jen-Kuang Huang, Minh Phan, and Jer-Nan Juang. Integrated system identification and state estimation for control offlexible space structures. *Journal of Guidance, Control, and Dynamics*, 15(1):88–95, 1992.

Han-Fu Chen and Lei Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. *SIAM Journal on Control and Optimization*, 25(4):845–867, 1987.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. *arXiv preprint arXiv:1806.07104*, 2018.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *arXiv preprint arXiv:1902.06223*, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Optimism-based adaptive regulation of linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive regulation and learning. *arXiv preprint arXiv:1811.04258*, 2018.

Claude-Nicolas Fiechter. Pac adaptive control of linear systems. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 72–80. Citeseer, 1997.

Dylan J Foster and Max Simchowitz. Logarithmic regret for adversarial online control. *arXiv preprint arXiv:2003.00189*, 2020.

Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. *arXiv preprint arXiv:2002.02064*, 2020.

Babak Hassibi, Ali H Sayed, and Thomas Kailath. *Indefinite-Quadratic Estimation and Control: A Unified Approach to H2 and H-infinity Theories*, volume 16. SIAM, 1999.

Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems*, pages 6702–6712, 2017.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

Elad Hazan, Sham M Kakade, and Karan Singh. The nonstochastic control problem. *arXiv preprint arXiv:1911.12178*, 2019.

Jer-Nan Juang, Minh Phan, Lucas G Horta, and Richard W Longman. Identification of observer/kalman filter markov parameters-theory and experiments. *Journal of Guidance, Control, and Dynamics*, 16(2):320–329, 1993.

Thomas Kailath, Ali H Sayed, and Babak Hassibi. Linear estimation, 2000.

Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. *SIAM Journal on Control and Optimization*, 25(2):466–481, 1987.

Tze Leung Lai, Ching Zong Wei, et al. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable linear quadratic control. *arXiv preprint arXiv:2002.00082*, 2020a.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret bound of adaptive control in linear quadratic gaussian (lqg) systems. *arXiv preprint arXiv:2003.05999*, 2020b.

Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

Bruce Lee and Andrew Lamperski. Non-asymptotic closed-loop system identification using autoregressive processes and hankel model reduction. *arXiv preprint arXiv:1909.02192*, 2019.

Holden Lee and Cyril Zhang. Robust guarantees for learning an autoregressive filter. *arXiv preprint arXiv:1905.09897*, 2019.

Lennart Ljung. System identification. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–19, 1999.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media, 2009.

Minh Phan, Jer-Nan Juang, Lucas G Horta, and Richard W Longman. System identification from closed-loop data with known output feedback dynamics. *Journal of guidance, control, and dynamics*, 17(4):661–669, 1994.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Max Simchowitz and Dylan J Foster. Naive exploration is optimal for online lqr. *arXiv preprint arXiv:2001.09576*, 2020.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. *arXiv preprint arXiv:2001.09254*, 2020.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Anastasios Tsiamis and George Pappas. Online learning of the kalman filter with logarithmic regret. *arXiv preprint arXiv:2002.05141*, 2020.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

Anastasios Tsiamis, Nikolai Matni, and George J Pappas. Sample complexity of kalman filtering for unknown systems. *arXiv preprint arXiv:1912.12309*, 2019.

Jack Umenberger, Mina Ferizbegovic, Thomas B Schön, and Håkan Hjalmarsson. Robust exploration in linear quadratic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 15310–15320, 2019.

Dante Youla, Hamid Jabr, and Jr Bongiorno. Modern wiener-hopf design of optimal controllers–part ii: The multivariable case. *IEEE Transactions on Automatic Control*, 21(3):319–338, 1976.

Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

Appendix

A Proofs of Section 3

A.1 Proof of Lemma 3.1

Using Lemma E.1, the following hold for all $t \in [T]$, with probability at least $1-\delta$,

$$\|w_t\| \le \overline{\sigma}_w \sqrt{2n \log \frac{4nT}{\delta}}, \qquad \|z_t\| \le \overline{\sigma}_z \sqrt{2m \log \frac{4mT}{\delta}}. \tag{14}$$

Thus we have,

$$\|b_t(\mathbf{G})\| = \|z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i\| \le \|z_t\| + \|C\| \|w_t\| \sum_{i=1}^{\infty} \|A^{t-i-1}\|. \tag{15}$$

Combining (14) and (15) gives the advertised bound. □

A.2 Proof of Lemma 3.2

Rolling out the dynamical system defining a policy π in Eq. 5 we can restate the action u_t^π as follows,

$$\begin{aligned} u_t^\pi &= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i} \\ &= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + C'_{\pi,u} B'_{\pi,w} w_{t-1} + \sum_{i=2}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i} \\ &= D_\pi z_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + D_\pi C w_{t-1} + \sum_{i=2}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,w} w_{t-i} \end{aligned}$$

Note that ${A'_\pi}^{i-1} B'_{\pi,w}$ is equal to $\begin{bmatrix} A + BD_\pi C \\ B_\pi C \end{bmatrix}$. Based on the definition of A'_π in Eq. 5, we restate A'_π as follows,

$$A'_\pi = \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} = \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} + \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix}$$

For any given bounded matrices A'_π and A, and any integer $i > 0$, we have

$$
\begin{aligned}
{A'_\pi}^{i} = \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^i &= \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&= \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \\
&\quad + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-2} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&\quad + \begin{bmatrix} A + BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix}^{i-2} \begin{bmatrix} A^2 & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} \\
&\vdots \\
&= \begin{bmatrix} A^i & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix} + \sum_{j=1}^{i} {A'_\pi}^{j-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A^{i-j} & 0_{n\times s} \\ 0_{s\times n} & 0_{s\times s} \end{bmatrix}
\end{aligned}
$$

We use this decomposition to relate u_t^π and $u_t^{\mathbf{M}}$. Now considering ${A'_\pi}^{i-1} B'_{\pi,w}$, for $i-1>0$ we have

$$
{A'_\pi}^{i-1} B'_{\pi,w} = \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} + \sum_{j=1}^{i-1} {A'_\pi}^{j-1} \begin{bmatrix} BD_\pi C & BC_\pi \\ B_\pi C & A_\pi \end{bmatrix} \begin{bmatrix} A^{i-1-j} \\ 0_{s\times n} \end{bmatrix} = \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} + \sum_{j=1}^{i-1} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j}
$$

Using this equality in the derivation of u_t^π we derive,

$$
\begin{aligned}
u_t^\pi = D_\pi z_t &+ \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + D_\pi C w_{t-1} \\
&\quad + \sum_{i=2}^{t-1} \begin{bmatrix} D_\pi C & C_\pi \end{bmatrix} \begin{bmatrix} A^{i-1} \\ 0_{s\times n} \end{bmatrix} w_{t-i} + \sum_{i=2}^{t-1} C'_{\pi,u} \sum_{j=1}^{i-1} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} w_{t-i} \\
= D_\pi z_t &+ \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=1}^{t-1} D_\pi C A^{i-1} w_{t-i} + \sum_{i=2}^{t-1} \sum_{j=1}^{i-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} C A^{i-1-j} w_{t-i}
\end{aligned}
$$

Note that $b_t(\mathbf{G}) = z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i = z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i}$. Inspired by this expression, we rearrange the previous sum as follows:

$$
\begin{aligned}
u_t^\pi &= D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{i=2}^{t-1} \sum_{j=1}^{i-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} CA^{i-1-j} w_{t-i} \\
&= D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{j=1}^{t-2} \sum_{i=j+1}^{t-1} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} CA^{i-1-j} w_{t-i} \\
&= D_\pi \left(z_t + \sum_{i=1}^{t-1} CA^{i-1} w_{t-i} \right) + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} z_{t-i} + \sum_{j=1}^{t-2} C'_{\pi,u} {A'_\pi}^{j-1} B'_{\pi,z} \sum_{i=1}^{t-j-1} CA^{t-j-i-1} w_i \\
&= D_\pi b_t + \sum_{i=1}^{t-1} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i}
\end{aligned}
$$

Now setting $M^{[0]} = D_\pi$, and $M^{[i]} = C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z}$ for all $0 < i < H'$, we conclude that for any LDC policy $\pi \in \Pi$, there exists at least one length H' DFC policy $\mathbf{M}(H')$ such that

$$
u_t^\pi - u_t^{\mathbf{M}} = \sum_{i=H'}^{t} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i}
$$

Using Cauchy Schwarz inequality we have

$$
\|u_t^\pi - u_t^{\mathbf{M}}\| \leq \left\| \sum_{i=H'}^{t} C'_{\pi,u} {A'_\pi}^{i-1} B'_{\pi,z} b_{t-i} \right\| \leq \psi(H')\kappa_b
$$

which states the first half of the Lemma.

Using the definition of y_t^π Eq. 4, we have

$$
y_t^\pi = z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i + \sum_{i}^{t-1} G^{[i]} u_{t-i}^\pi.
$$

Similarly for $y_t^{\mathbf{M}}$ we have,

$$
y_t^{\mathbf{M}} = z_t + \sum_{i=1}^{t-1} CA^{t-i-1} w_i + \sum_{i}^{t-1} G^{[i]} u_{t-i}^\pi.
$$

Subtracting these two equations, we derive,

$$
y_t^\pi - y_t^{\mathbf{M}} = \sum_{i}^{t-1} G^{[i]} u_{t-i}^\pi - \sum_{i}^{t-1} G^{[i]} u_{t-i}^{\mathbf{M}} = \sum_{i}^{t-1} G^{[i]} (u_{t-i}^\pi - u_{t-i}^{\mathbf{M}})
$$

resulting in

$$
\|y_t^\pi - y_t^{\mathbf{M}}\| \leq \psi(H')\kappa_{\mathbf{G}}\kappa_b
$$

which states the second half of the Lemma. □

B System Identification Algorithm

Algorithm 2 gives the system-identification algorithm, SysId, that is called in the beginning of each epoch. For further discussion of the algorithm please refer to Lale et al. [2020b].

Algorithm 2 SysId

Input: $\widehat{\mathcal{G}}_{\mathbf{yu,i}}$, H_{est}, system order n, d_1, d_2 such that $d_1 + d_2 + 1 = H_{est}$
Form two $d_1 \times (d_2+1)$ Hankel matrices $\mathcal{H}_{\hat{\mathbf{F}}_\mathbf{i}}$ and $\mathcal{H}_{\hat{\mathbf{G}}_\mathbf{i}}$ from $\widehat{\mathcal{G}}_{\mathbf{yu,i}} = [\hat{\mathbf{F}}_{\mathbf{i,1}}, \dots, \hat{\mathbf{F}}_{\mathbf{i,H_{est}}}, \hat{\mathbf{G}}_{\mathbf{i,1}}, \dots, \hat{\mathbf{G}}_{\mathbf{i,H_{est}}}]$, and construct $\hat{\mathcal{H}}_i = \left[\mathcal{H}_{\hat{\mathbf{F}}_\mathbf{i}}, \ \mathcal{H}_{\hat{\mathbf{G}}_\mathbf{i}}\right] \in \mathbb{R}^{md_1 \times (m+p)(d_2+1)}$
Obtain $\hat{\mathcal{H}}_i^-$ by discarding (d_2+1)th and $(2d_2+2)$th block columns of $\hat{\mathcal{H}}_i$
Using SVD obtain $\hat{\mathcal{N}}_i \in \mathbb{R}^{md_1 \times (m+p)d_2}$, the best rank-$n$ approximation of $\hat{\mathcal{H}}_i^-$
Obtain $\mathbf{U_i}, \mathbf{\Sigma_i}, \mathbf{V_i} = \mathrm{SVD}(\hat{\mathcal{N}}_i)$
Construct $\hat{\mathbf{O}}_\mathbf{i}(\bar{A}, C, d_1) = \mathbf{U_i}\mathbf{\Sigma_t}^{1/2} \in \mathbb{R}^{md_1 \times n}$
Construct $[\hat{\mathbf{C}}_\mathbf{i}(\bar{A}, F, d_2+1), \ \ \hat{\mathbf{C}}_\mathbf{i}(\bar{A}, B, d_2+1)] = \mathbf{\Sigma_i}^{1/2}\mathbf{V_i} \in \mathbb{R}^{n \times (m+p)d_2}$
Obtain $\hat{C}_i \in \mathbb{R}^{m \times n}$, the first m rows of $\hat{\mathbf{O}}_\mathbf{i}(\bar{A}, C, d_1)$
Obtain $\hat{B}_i \in \mathbb{R}^{n \times p}$, the first p columns of $\hat{\mathbf{C}}_\mathbf{i}(\bar{A}, B, d_2+1)$
Obtain $\hat{F}_i \in \mathbb{R}^{n \times m}$, the first m columns of $\hat{\mathbf{C}}_\mathbf{i}(\bar{A}, F, d_2+1)$
Obtain $\hat{\mathcal{H}}_i^+$ by discarding 1st and (d_2+2)th block columns of $\hat{\mathcal{H}}_i$
Obtain $\hat{\bar{A}}_i = \hat{\mathbf{O}}_\mathbf{i}^\dagger(\bar{A}, C, d_1) \ \ \hat{\mathcal{H}}_i^+ \ \ [\hat{\mathbf{C}}_\mathbf{i}(\bar{A}, F, d_2+1), \quad \hat{\mathbf{C}}_\mathbf{i}(\bar{A}, B, d_2+1)]^\dagger$
Obtain $\hat{A}_i = \hat{\bar{A}}_i + \hat{F}_i\hat{C}_i$
Obtain $\hat{L}_i \in \mathbb{R}^{n \times m}$, as the first $n \times m$ block of $\hat{A}_i^\dagger \hat{\mathbf{O}}_\mathbf{i}^\dagger(\bar{A}, C, d_1)\hat{\mathcal{H}}_i^-$

C Proofs for Dynamics Learning

C.1 Proof of Theorem 4.1

First, we have the following lemma that provides the persistence of excitation of inputs in the warm-up period. Let $T_o = \frac{32\Upsilon_w^4 \log^2\left(\frac{2H_{est}(m+p)}{\delta}\right)}{\sigma_{\min}^4(\mathcal{G}^{ol})\min\{\underline{\sigma}_w^4, \underline{\sigma}_z^4, \sigma_u^4\}}$.

Lemma C.1 (Open-Loop Persistence of Excitation, Lemma A.1 of Lale et al. [2020b])**.** *If the warm-up duration $T_{burn} \geq T_o$, then for $T_o \leq t \leq T_{burn}$, with probability at least $1-\delta$ we have*

$$\sigma_{\min}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top\right) \geq t\frac{\sigma_o^2 \min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2, \sigma_u^2\}}{2}. \tag{16}$$

Combining Lemma C.1 with Theorem E.2 gives

$$\|\widehat{\mathcal{G}}_{\mathbf{yu,1}} - \mathcal{G}_{\mathbf{yu}}\| \leq \frac{\kappa_{est}}{\sigma_o\sqrt{T_{burn}}\sqrt{\frac{\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2, \sigma_u^2\}}{2}}},$$

with probability at least $1-2\delta$. Notice that $\sigma_*^2 \coloneqq \min\left\{\frac{\sigma_o^2\underline{\sigma}_w^2}{2}, \frac{\sigma_o^2\underline{\sigma}_z^2}{2}, \frac{\sigma_o^2\sigma_y^2}{2}, \frac{\sigma_c^2\underline{\sigma}_w^2}{16}, \frac{\sigma_c^2\underline{\sigma}_z^2}{16}\right\} \leq \frac{\sigma_o^2\min\left\{\underline{\sigma}_w^2, \underline{\sigma}_z^2, \sigma_u^2\right\}}{2}$. Thus, the statement of Theorem 4.1 holds for $T_{burn} \geq t \geq T_o$ with probability at least $1-2\delta$. □

C.2 Proof of Lemma 4.1

In the beginning of the first epoch, AdaptOn constructs H-length Markov parameters matrix using the estimates, $\widehat{A}_1, \widehat{B}_1, \widehat{C}_1$, provided by SysId on the estimated $\widehat{\mathcal{G}}_{\mathbf{yu,1}}$. From the assumption that $\psi_{\mathbf{G}}(H+1) \leq 1/10T$, we have that $\sum_{j\geq H+1} \|\widehat{G}_1^{[j]} - G^{[j]}\| \leq \epsilon_{\mathbf{G}}(1,\delta)/2$. Next consider the first H-parameters

$$\sum_{j\geq 1}^{H} \|\widehat{G}_1^{[j]} - G^{[j]}\| = \sum_{j\geq 1}^{H} \|\widehat{C}_1 \widehat{A}_1^{j-1} \widehat{B}_1 - CA^{j-1}B\|. \tag{17}$$

Consider Theorem E.3. For some unitary matrix $\mathbf{T}$, we denote $\Delta A \coloneqq \|\widehat{A}_1 - \mathbf{T}^\top A \mathbf{T}\|$, $\Delta B \coloneqq \|\widehat{B}_1 - \mathbf{T}^\top B\| = \|\widehat{C}_1 - C\mathbf{T}\|$. Define

$$T_{\mathcal{G}_{\mathbf{yu}}} \coloneqq \frac{\kappa_{est}^2}{\sigma_*^2}, \quad T_A \coloneqq T_{\mathcal{G}_{\mathbf{yu}}} \frac{4c_1^2\gamma_{\mathcal{H}}^2}{(1-\rho(A))^2}, \quad T_B \coloneqq T_{\mathcal{G}_{\mathbf{yu}}} \frac{20nH_{est}}{\sigma_n(\mathcal{H})},$$
$$T_{cx} \coloneqq T_{\mathcal{G}_{\mathbf{yu}}} \frac{16c_1^2\kappa_b^2\kappa_{\mathcal{M}}^2\kappa_{\mathbf{G}}^2 H' \gamma_{\mathbf{G}}^2 \gamma_{\mathcal{H}}^2 \alpha_{loss}}{\alpha}, \quad T_{\epsilon_{\mathbf{G}}} \coloneqq 4c_1^2\kappa_{\mathcal{M}}^2\kappa_{\mathbf{G}}^2\gamma_{\mathbf{G}}^2\gamma_{\mathcal{H}}^2 T_{\mathcal{G}_{\mathbf{yu}}}. \tag{18}$$

For $T_{burn} > \max\{T_A, T_B\}$, we have that $\Delta A \leq \frac{1-\rho(A)}{2}$ and $\Delta B \leq 1$. Using this fact, we bound (17):

$$\begin{aligned}
&\sum_{j\geq 1}^{H} \|\widehat{C}_1 \widehat{A}_1^{j-1} \widehat{B}_1 - CA^{j-1}B\| \\
&\leq \Delta B(\|B\|+\|C\|+1) + \sum_{i=1}^{H-1} \Phi(A)\rho^i(A)\Delta B(\|B\|+\|C\|+1) + \|\widehat{A}_1^i - \mathbf{T}^\top A^i \mathbf{T}\|(\|C\|\|B\|+\|B\|+\|C\|+1) \\
&\leq \left(1 + \frac{\Phi(A)}{1-\rho(A)}\right)\Delta B(\|B\|+\|C\|+1) + \Delta A(\|C\|\|B\|+\|B\|+\|C\|+1) \sum_{i=1}^{H-1}\sum_{j=0}^{i-1}\binom{i}{j}\|A^j\|(\Delta A)^{i-1-j} \\
&\leq \left(1 + \frac{\Phi(A)}{1-\rho(A)}\right)\Delta B(\|B\|+\|C\|+1) \\
&\qquad + \Delta A \Phi(A)(\|C\|\|B\|+\|B\|+\|C\|+1) \sum_{i=1}^{H-1}\sum_{j=0}^{i-1}\binom{i}{j}\rho^j(A)\left(\frac{1-\rho(A)}{2}\right)^{i-1-j} \\
&\leq \left(1 + \frac{\Phi(A)}{1-\rho(A)}\right)\Delta B(\|B\|+\|C\|+1) + \frac{2\Delta A\Phi(A)}{1-\rho(A)}(\|C\|\|B\|+\|B\|+\|C\|+1)\sum_{i=1}^{H-1}\left[\left(\frac{1+\rho}{2}\right)^i - \rho^i\right] \\
&\leq \Delta B\left(1 + \frac{\Phi(A)}{1-\rho(A)}\right)(\|B\|+\|C\|+1) + \frac{2\Delta A\Phi(A)}{(1-\rho(A))^2}(\|C\|\|B\|+\|B\|+\|C\|+1)
\end{aligned}$$

Assuming that $\|F\| + \|C\| > 1$ for simplicity, from the exact expressions of Theorem E.3, we have $\Delta A > \Delta B$. For the given $\gamma_{\mathbf{G}}$ and $\gamma_{\mathcal{H}}$, we can upper bound the last expression above as follow,

$$\sum_{j\geq 1}^{H} \|\widehat{C}_1 \widehat{A}_1^{j-1} \widehat{B}_1 - CA^{j-1}B\| \leq \gamma_{\mathbf{G}}\Delta A \leq \frac{c_1\gamma_{\mathbf{G}}\gamma_{\mathcal{H}}\kappa_{est}}{\sigma_\star\sqrt{T_{burn}}}, \tag{19}$$

where (19) follows from Theorem 4.1 and the concentration result for $\|\widehat{A} - \mathbf{T}^\top A \mathbf{T}\|$ in Theorem E.3. The second inequality of the lemma holds since $T_{burn} \geq \max\{T_{\epsilon_{\mathbf{G}}}, T_{cx}\}$. □

C.3 Persistence of Excitation Condition of $\mathbf{M} \in \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$

If the underlying system is fully known, the following are the inputs and outputs of the system:

$$
\begin{aligned}
u_t &= \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G}) \\
y_t &= [G^{[0]} \;\; G^{[1]} \dots \;\; G^{[H]}] \left[u_t^\top \;\; u_{t-1}^\top \dots u_{t-H}^\top\right]^\top + b_t(\mathbf{G}) + \mathbf{r_t}
\end{aligned}
$$

where $\mathbf{r_t} = \sum_{k=H+1}^{t-1} G^{[k]} u_{t-k}$. For H_{est} defined in Section 4.2, $H_{est} \geq \max\{2n+1, \frac{\log(c_H T^2 \sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}\}$, define

$$
\phi_t = \left[y_{t-1}^\top \dots y_{t-H_{est}}^\top \;\; u_{t-1}^\top \dots \;\; u_{t-H_{est}}^\top\right]^\top \in \mathbb{R}^{(m+p)H_{est}}.
$$

We have the following decompositions for ϕ_t:

$$
\phi_t = \underbrace{\begin{bmatrix}
G^{[0]} & G^{[1]} & \dots & \dots & \dots & G^{[H]} & 0_{m\times p} & 0_{m\times p} & \dots & 0_{m\times p} \\
0_{m\times p} & G^{[0]} & \dots & \dots & \dots & G^{[H-1]} & G^{[H]} & 0_{m\times p} & \dots & 0_{m\times p} \\
 & & \ddots & & & & \ddots & & & \\
0_{m\times p} & \dots & 0_{m\times p} & G^{[0]} & G^{[1]} & \dots & \dots & \dots & G^{[H-1]} & G^{[H]} \\
I_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & \dots & \dots & \dots & 0_{p\times p} \\
0_{p\times p} & I_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & 0_{p\times p} & \dots & \dots & \dots & 0_{p\times p} \\
 & & \ddots & & & & & & & \\
0_{p\times p} & 0_{p\times p} & \dots & I_{p\times p} & 0_{p\times p} & \dots & \dots & \dots & \dots & 0_{p\times p}
\end{bmatrix}}_{\mathcal{T}_{\mathbf{G}} \in \mathbb{R}^{H_{est}(m+p)\times(H_{est}+H)p}}
\underbrace{\begin{bmatrix} u_{t-1} \\ \vdots \\ u_{t-H} \\ \vdots \\ u_{t-H-H_{est}} \end{bmatrix}}_{\mathcal{U}_t}
+ \underbrace{\begin{bmatrix} b_{t-1} \\ \vdots \\ b_{t-H_{est}} \\ 0_p \\ \vdots \\ 0_p \end{bmatrix}}_{B_y(\mathbf{G})(t)}
+ \underbrace{\begin{bmatrix} \mathbf{r_{t-1}} \\ \vdots \\ \mathbf{r_{t-H_{est}}} \\ 0_p \\ \vdots \\ 0_p \end{bmatrix}}_{\mathbf{R}_t}
$$

$$
\mathcal{U}_t = \underbrace{\begin{bmatrix}
M_{t-1}^{[0]} & M_{t-1}^{[1]} & \dots & \dots & M_{t-1}^{[H'-1]} & 0_{p\times m} & 0_{p\times m} & \dots & 0_{p\times m} \\
0_{p\times m} & M_{t-2}^{[0]} & \dots & \dots & M_{t-2}^{[H'-2]} & M_{t-2}^{[H'-1]} & 0_{p\times m} & \dots & 0_{p\times m} \\
 & & \ddots & & & & & \ddots & \\
0_{p\times m} & \dots & 0_{p\times m} & M_{t-H_{est}-H}^{[0]} & \dots & \dots & \dots & \dots & M_{t-H_{est}-H}^{[H'-1]}
\end{bmatrix}}_{\mathcal{T}_{\mathbf{M}_t} \in \mathbb{R}^{(H_{est}+H)p\times m(H+H'+H_{est}-1)}}
\underbrace{\begin{bmatrix} b_{t-1}(\mathbf{G}) \\ b_{t-2}(\mathbf{G}) \\ \vdots \\ b_{t-H'+1}(\mathbf{G}) \\ \vdots \\ b_{t-H_{est}-H-H'+1}(\mathbf{G}) \end{bmatrix}}_{B(\mathbf{G})(t)}
$$

$$
B(\mathbf{G})(t) = \underbrace{\begin{bmatrix}
I_m & 0_m & \dots & 0_m & C & CA & \dots & \dots & \dots & CA^{t-3} \\
0_m & I_m & & 0_m & 0_{m\times n} & C & \dots & \dots & \dots & CA^{t-4} \\
 & & \ddots & & & & \ddots & \ddots & & \\
0_m & 0_m & \dots & I_m & 0_{m\times n} & \dots & \dots & C & \dots & CA^{t-H_{est}-H-H'-1}
\end{bmatrix}}_{\mathcal{O}_t}
\underbrace{\begin{bmatrix} z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-H_{est}-H-H'+1} \\ w_{t-2} \\ w_{t-3} \\ \vdots \\ w_1 \end{bmatrix}}_{\boldsymbol{\eta}_t}
$$

and $B_y(\mathbf{G})(t) = \underbrace{\begin{bmatrix} I_m & 0_m & \dots & \dots & 0_m & C & \dots & \dots & \dots & CA^{t-3} \\ & \ddots & & & \vdots & & \ddots & & \ddots & \\ 0_m & \dots & I_m & \dots & 0_m & 0_{m\times n} & \dots & C & \dots & CA^{t-H_{est}-2} \\ & & & & \multicolumn{6}{c}{\mathbf{0}_{(pH_{est})\times((H_{est}+H+H'-1)m+(t-2)n)}} \end{bmatrix}}_{\bar{\mathcal{O}}_t} \boldsymbol{\eta}_t.$

Combining all gives

$$\phi_t = \left(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}\mathcal{O}_t + \bar{\mathcal{O}}_t\right)\boldsymbol{\eta}_t + \mathbf{R}_t.$$

Persistence of Excitation of $\mathbf{M} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$ on System Θ. *For the given system Θ, for $t \geq H + H' + H_{est}$, $\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}\mathcal{O}_t + \bar{\mathcal{O}}_t$ is full row rank for all $\mathbf{M} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$, i.e.,*

$$\sigma_{\min}(\mathcal{T}_{\mathbf{G}}\mathcal{T}_{\mathbf{M}_t}\mathcal{O}_t + \bar{\mathcal{O}}_t) > \sigma_c > 0. \tag{20}$$

C.4 Proof of Theorem 4.2

First we have the following lemma, that shows inputs have persistence of excitation during the adaptive control period. Let $d = \min\{m, p\}$. Using (18), define

$$T_{cl} = \frac{T_{\epsilon_{\mathbf{G}}}}{\left(\frac{3\sigma_c^2 \min\{\sigma_w^2, \sigma_z^2\}}{8\kappa_u^2\kappa_y H_{est}} - \frac{1}{10T}\right)^2}, \quad T_c = \frac{2048\Upsilon_c^4 H_{est}^2 \log\left(\frac{H_{est}(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{d} + \frac{2}{\epsilon}\right)}{\sigma_c^4 \min\{\sigma_w^4, \sigma_z^4\}}.$$

Lemma C.2. *After T_c time steps in the adaptive control period, with probability $1 - 3\delta$, we have persistence of excitation for the remainder of adaptive control period,*

$$\sigma_{\min}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top\right) \geq t\frac{\sigma_c^2 \min\{\sigma_w^2, \sigma_z^2\}}{16}. \tag{21}$$

Proof. During the adaptive control period, at time t, the input of AdaptOn is given by

$$u_t = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G}) + M_t^{[j]}\left(b_{t-j}(\widehat{\mathbf{G}}_i) - b_{t-j}(\mathbf{G})\right)$$

where

$$b_{t-j}(\mathbf{G}) = y_{t-j} - \sum_{k=1}^{t-j-1} G^{[k]} u_{t-j-k} = z_{t-j} + \sum_{k=1}^{t-j-1} CA^{t-j-k-1} w_k \tag{22}$$

$$b_{t-j}(\widehat{\mathbf{G}}_i) = y_{t-j} - \sum_{k=1}^{H} \widehat{G}_i^{[k]} u_{t-j-k} \tag{23}$$

Thus, we obtain the following for u_t and y_t,

$$u_t = \sum_{j=0}^{H'-1} M_t^{[j]} b_{t-j}(\mathbf{G}) + \underbrace{\sum_{j=0}^{H'-1} M_t^{[j]} \left(\sum_{k=1}^{t-j-1} [G^{[k]} - \widehat{G}_i^{[k]}] u_{t-j-k} \right)}_{u_{\Delta b}(t)}$$

$$y_t = [G^{[0]} \;\; G^{[1]} \dots \;\; G^{[H]}] \left[u_t^\top \;\; u_{t-1}^\top \dots u_{t-H}^\top \right]^\top + b_t(\mathbf{G}) + \mathbf{r_t}$$

where $\mathbf{r_t} = \sum_{k=H+1}^{t-1} G^{[k]} u_{t-k}$ and $\sum_{k=H}^{t-1} \|G^{[k]}\| \leq \psi_{\mathbf{G}}(H+1) \leq 1/10T$ which is bounded by the assumption. Notice that $\|u_{\Delta b}(t)\| \leq \kappa_{\mathcal{M}} \kappa_u \epsilon_{\mathbf{G}}(1,\delta)$ for all $t \in T_{burn}$. Using the definitions from Appendix C.3, ϕ_t can be written as,

$$\phi_t = \left(\mathcal{T}_{\mathbf{G}} \mathcal{T}_{\mathbf{M}_t} \mathcal{O}_t + \bar{\mathcal{O}}_t \right) \boldsymbol{\eta}_t + \mathbf{R}_t + \mathcal{T}_{\mathbf{G}} \mathcal{U}_{\Delta b}(t) \tag{24}$$

where

$$\mathcal{U}_{\Delta b}(t) = \begin{bmatrix} u_{\Delta b}(t-1) \\ u_{\Delta b}(t-2) \\ \vdots \\ u_{\Delta b}(t-H_{est}) \\ \vdots \\ u_{\Delta b}(t-H_{est}-H) \end{bmatrix}.$$

Consider the following,

$$\begin{aligned} \mathbb{E}\left[\phi_t \phi_t^\top\right] = \mathbb{E}\Big[& \left(\mathcal{T}_{\mathbf{G}} \mathcal{T}_{\mathbf{M}_t} \mathcal{O}_t + \bar{\mathcal{O}}_t \right) \boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top \left(\mathcal{T}_{\mathbf{G}} \mathcal{T}_{\mathbf{M}_t} \mathcal{O}_t + \bar{\mathcal{O}}_t \right)^\top + \boldsymbol{\eta}_t^\top \left(\mathcal{T}_{\mathbf{G}} \mathcal{T}_{\mathbf{M}_t} \mathcal{O}_t + \bar{\mathcal{O}}_t \right)^\top \left(\mathcal{T}_{\mathbf{G}} \mathcal{U}_{\Delta b}(t) + \mathbf{R}_t \right) \\ & + \left(\mathcal{T}_{\mathbf{G}} \mathcal{U}_{\Delta b}(t) + \mathbf{R}_t \right)^\top \left(\mathcal{T}_{\mathbf{G}} \mathcal{T}_{\mathbf{M}_t} \mathcal{O}_t + \bar{\mathcal{O}}_t \right) \boldsymbol{\eta}_t + \left(\mathcal{T}_{\mathbf{G}} \mathcal{U}_{\Delta b}(t) + \mathbf{R}_t \right)^\top \left(\mathcal{T}_{\mathbf{G}} \mathcal{U}_{\Delta b}(t) + \mathbf{R}_t \right) \Big] \end{aligned}$$

$$\begin{aligned} \sigma_{\min}\left(\mathbb{E}\left[\phi_t \phi_t^\top\right] \right) &\geq \sigma_c^2 \min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\} \\ &\quad - 2\kappa_b \left(\kappa_{\mathcal{M}} + \kappa_{\mathcal{M}} \kappa_{\mathbf{G}} + 1 \right) \sqrt{H_{est}} ((1+\kappa_{\mathbf{G}}) \kappa_{\mathcal{M}} \kappa_u \epsilon_{\mathbf{G}}(1,\delta) \sqrt{H_{est}} + \sqrt{H_{est}} \kappa_u / 10T) \\ &\geq \sigma_c^2 \min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\} - 2\kappa_u^2 \kappa_y H_{est} (2\kappa_{\mathbf{G}} \kappa_{\mathcal{M}} \epsilon_{\mathbf{G}}(1,\delta) + 1/10T) \end{aligned}$$

Note that for $T_{burn} \geq T_{cl}$, $\epsilon_{\mathbf{G}}(1,\delta) \leq \frac{1}{2\kappa_{\mathcal{M}} \kappa_{\mathbf{G}}} \left(\frac{3\sigma_c^2 \min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}}{8\kappa_u^2 \kappa_y H_{est}} - \frac{1}{10T} \right)$ with probability at least $1-2\delta$. Thus, we get

$$\sigma_{\min}\left(\mathbb{E}\left[\phi_t \phi_t^\top\right] \right) \geq \frac{\sigma_c^2}{4} \min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}, \tag{25}$$

for all $t \geq T_{burn}$. Using Lemma 4.2, we have that for $\Upsilon_c := (\kappa_y + \kappa_u)$, $\|\phi_t\| \leq \Upsilon_c \sqrt{H_{est}}$ with probability at least $1-2\delta$. Therefore, for a chosen $\mathbf{M} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$, using Theorem E.1, we have the following with probability $1-3\delta$:

$$\lambda_{\max}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top - \mathbb{E}[\phi_i \phi_i^\top] \right) \leq 2\sqrt{2t} \Upsilon_c^2 H_{est} \sqrt{\log\left(\frac{H_{est}(m+p)}{\delta} \right)}. \tag{26}$$

In order to show that this holds for any chosen $\mathbf{M} \in \mathcal{M}(H', \kappa_{\mathcal{M}})$, we adopt a standard covering argument. We know that from Lemma 5.4 of Simchowitz et al. [2020], the Euclidean diameter of $\mathcal{M}(H', \kappa_{\mathcal{M}})$ is at most $2\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}$, *i.e.* $\|\mathbf{M}_t\|_F \leq \kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}$ for all $\mathbf{M}_t \in \mathcal{M}(H', \kappa_{\mathcal{M}})$. Thus, we can upper bound the covering number as follows,

$$\mathcal{N}(B(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}}), \|\cdot\|_F, \epsilon) \leq \left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}} + \frac{2}{\epsilon}\right)^{H'mp}.$$

The following holds for all the centers of ϵ-balls in $\|\mathbf{M}_t\|_F$, for all $t \geq T_{burn}$, with probability $1-3\delta$:

$$\lambda_{\max}\left(\sum_{i=1}^{t} \phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_c^2 H_{est}\sqrt{\log\left(\frac{H_{est}(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}} + \frac{2}{\epsilon}\right)}. \tag{27}$$

Consider all $\mathbf{M}$ in the ϵ-balls, *i.e.* effect of epsilon perturbation in $\|\mathbf{M}\|_F$ sets, using Weyl's inequality we have with probability at lest $1-3\delta$,

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{4}\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\} - \frac{8c\kappa_b^3\kappa_{\mathbf{G}}H_{est}\epsilon\left(2\kappa_{\mathcal{M}}^2 + 3\kappa_{\mathcal{M}} + 3\right)}{\sqrt{\min\{m,p\}}}\left(1 + \frac{1}{10T}\right)\right)$$
$$- 2\sqrt{2t}\Upsilon_c^2 H_{est}\sqrt{\log\left(\frac{H_{est}(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}} + \frac{2}{\epsilon}\right)}.$$

for some enough constant c and $\epsilon \leq 1$. Let $\epsilon = \min\left\{1, \frac{\sigma_c^2\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}\sqrt{\min\{m,p\}}}{68c\kappa_b^3\kappa_{\mathbf{G}}H_{est}\left(2\kappa_{\mathcal{M}}^2+3\kappa_{\mathcal{M}}+3\right)}\right\}$. For this choice of ϵ, we get

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{8}\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}\right)$$
$$- 2\sqrt{2t}\Upsilon_c^2 H_{est}\sqrt{\log\left(\frac{H_{est}(m+p)}{\delta}\right) + H'mp\log\left(\kappa_{\mathcal{M}}\sqrt{\min\{m,p\}} + \frac{2}{\epsilon}\right)}.$$

For picking $T_{burn} \geq T_c$, we can guarantee that after T_c time steps in the first epoch we have the advertised lower bound. □

Combining Lemma C.2 with Theorem E.2 gives

$$\|\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}\| \leq \frac{\kappa_{est}}{\sigma_c\sqrt{t_i}\sqrt{\frac{\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}}{16}}},$$

for all i, with probability at least $1-4\delta$. Notice that $\sigma_\star^2 := \min\left\{\frac{\sigma_o^2\underline{\sigma}_w^2}{2}, \frac{\sigma_o^2\underline{\sigma}_z^2}{2}, \frac{\sigma_o^2\underline{\sigma}_y^2}{2}, \frac{\sigma_c^2\underline{\sigma}_w^2}{16}, \frac{\sigma_c^2\underline{\sigma}_z^2}{16}\right\} \leq \frac{\sigma_c^2\min\{\underline{\sigma}_w^2, \underline{\sigma}_z^2\}}{16}$. Thus, the statement of Theorem 4.2 holds with probability at least $1-4\delta$.

D Proofs for Regret Bound

In order to prove Theorem 4.3, we follow the proof steps of Theorem 5 of Simchowitz et al. [2020]. The main difference is that, AdaptOn updates the Markov parameter estimates in epochs throughout the adaptive control period which provides decrease in the gradient error in each epoch. These updates allow AdaptOn to remove $\mathcal{O}(\sqrt{T})$ term in the regret expression of Theorem 5. In the following, we state how the proof of Theorem 5 of Simchowitz et al. [2020] is adapted to the setting of AdaptOn.

D.1 Proof of Theorem 4.3

Recall the hypothetical "true prediction" y's, y_t^{pred} and losses, $f_t^{pred}(M)$ defined in Definition 8.1 of Simchowitz et al. [2020]. Up to truncation by H, they describe the true counterfactual output of the system for AdaptOn inputs during the adaptive control period and the corresponding counterfactual loss functions. Lemma E.2, shows that at all epoch i, at any time step $t \in [t_i, \ \ldots, \ t_{i+1}-1]$, the gradient $f_t^{pred}(M)$ is close to the gradient of the loss function of AdaptOn:

$$\left\|\nabla f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right) - \nabla f_t^{\text{pred}}\left(\mathbf{M}\right)\right\|_{\text{F}} \leq C_{\text{approx}} \epsilon_{\mathbf{G}}(i, \delta), \tag{28}$$

where $C_{\text{approx}} := \sqrt{H'} \kappa_{\mathbf{G}} \kappa_{\mathcal{M}} \kappa_b^2 \left(16\overline{\alpha}_{loss} + 24L\right)$. For a comparing controller $\mathbf{M}_{comp} \in \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$ and a restricted set $\mathcal{M}_0 = \mathcal{M}(H_0', \kappa_{\mathcal{M}}/2) \subset \mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$, where $H_0' = \lfloor \frac{H'}{2} \rfloor - H$, we have the following regret decomposition:

$$\begin{aligned}
\textsc{Regret}(T) \leq\ & \underbrace{\left(\sum_{t=1}^{T_{burn}} \ell_t\left(y_t, u_t\right)\right)}_{\text{warm-up regret}} + \underbrace{\left(\sum_{t=T_{burn}+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_{burn}+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right]\right)}_{\text{algorithm truncation error}} \\
& + \underbrace{\left(\sum_{t=T_{burn}+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] - \sum_{t=T_{burn}+1}^{T} f_t^{\text{pred}}\left(\mathbf{M}_{comp}\right)\right)}_{f^{\text{pred}} \text{ policy regret}} \\
& + \underbrace{\left(\sum_{t=T_{burn}+1}^{T} f_t^{\text{pred}}\left(\mathbf{M}_{comp}\right) - \inf_{\mathbf{M} \in \mathcal{M}_0} \sum_{t=T_{burn}+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right)\right)}_{\text{comparator approximation error}} \\
& + \underbrace{\left(\inf_{\mathbf{M} \in \mathcal{M}_0} \sum_{t=T_{burn}+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right) - \inf_{\mathbf{M} \in \mathcal{M}_0} \sum_{t=T_{burn}+1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right)\right)}_{\text{comparator truncation error}} \\
& + \underbrace{\left(\inf_{\mathbf{M} \in \mathcal{M}_0} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=0}^{T} \ell(y^{\pi_\star}, u^{\pi_\star})\right)}_{\text{policy approximation error}}
\end{aligned} \tag{29}$$

We will consider each term separately.

Warm-up Regret: From Assumption 2.1 and Lemma 4.2, we get

$$\left(\sum_{t=1}^{T_{burn}} \ell_t\left(y_t, u_t\right)\right) \leq T_{burn} L \kappa_y^2.$$

Algorithm Truncation Error: From Assumption 2.1, we get

$$\begin{aligned}
\sum_{t=T_{burn}+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_{burn}+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] &\leq \sum_{t=T_{burn}+1}^{T} \left|\ell_t\left(y_t, u_t\right) - \ell_t\left(b_t(\mathbf{G}) + \sum_{i=1}^{H} G^{[i]} u_{t-i}, u_t\right)\right| \\
&\leq \sum_{t=T_{burn}+1}^{T} L\kappa_y \left\| y_t - b_t(\mathbf{G}) + \sum_{i=1}^{H} G^{[i]} u_{t-i} \right\| \\
&\leq \sum_{t=T_{burn}+1}^{T} L\kappa_y \left\| \sum_{i=H+1} G^{[i]} u_{t-i} \right\| \\
&\leq T L \kappa_y \kappa_u \psi_{\mathbf{G}}(H+1)
\end{aligned}$$

Since $\psi_{\mathbf{G}}(H+1) \leq 1/10T$, we get $\sum_{t=T_{burn}+1}^{T} \ell_t\left(y_t, u_t\right) - \sum_{t=T_{burn}+1}^{T} F_t^{\text{pred}}\left[\mathbf{M}_{t:t-H}\right] \leq L\kappa_y\kappa_u/10$.

Comparator Truncation Error: Similar to algorithm truncation error above,

$$\begin{aligned}
\inf_{\mathbf{M}\in\mathcal{M}_0} \sum_{t=T_{burn}+1}^{T} f_t\left(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G})\right) - \inf_{\mathbf{M}\in\mathcal{M}_0} \sum_{t=T_{burn}+1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) &\leq T L \kappa_{\mathbf{G}} \kappa_{\mathcal{M}}^2 \kappa_b^2 \psi_{\mathbf{G}}(H+1) \\
&\leq L \kappa_{\mathbf{G}} \kappa_{\mathcal{M}}^2 \kappa_b^2 / 10
\end{aligned}$$

Policy Approximation Error: By the assumption that $M_\star$ lives in the given convex set $\mathcal{M}\left(H', \kappa_{\mathcal{M}}\right)$ and Assumption 2.1, using Lemma 3.2, we get

$$\begin{aligned}
\inf_{\mathbf{M}\in\mathcal{M}_0} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=1}^{T} \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) &\leq \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}_\star}, u_t^{\mathbf{M}_\star}\right) - \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) \\
&\leq T L \kappa_y \left(\psi(H_0')\kappa_b + \psi(H_0')\kappa_{\mathbf{G}}\kappa_b\right) \\
&\leq 2 T L \kappa_y \kappa_{\mathbf{G}} \kappa_b \psi(H_0')
\end{aligned}$$

Since $\psi(H_0') \leq \kappa_{\mathcal{M}}/T$, we get $\inf_{\mathbf{M}\in\mathcal{M}_0} \sum_{t=1}^{T} \ell_t\left(y_t^{\mathbf{M}}, u_t^{\mathbf{M}}\right) - \sum_{t=1}^{T} \ell_t(y_t^{\pi_\star}, u_t^{\pi_\star}) \leq 2L\kappa_{\mathcal{M}}\kappa_y\kappa_{\mathbf{G}}\kappa_b$.

$\mathbf{f}^{\text{pred}}$ Policy Regret : In order to utilize Theorem E.4, we need the strong convexity, Lipschitzness and smoothness properties stated in the theorem. Lemmas E.3-E.5 provide those conditions. Combining these with (28), we obtain the following adaptation of Theorem E.4:

Lemma D.1. *For step size $\eta = \frac{12}{\alpha t}$, the following bound holds with probability $1-\delta$:*

$$\mathbf{f}^{\text{pred}} \textit{ policy regret } + \frac{\alpha}{48}\sum_{t=T_{burn}+1}^{T} \|\mathbf{M}_t - \mathbf{M}_{comp}\|_F^2$$
$$\lesssim \frac{L^2 {H'}^3 \min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) + \frac{1}{\alpha}\sum_{t=T_{burn}+1}^{T} C_{approx}^2 \epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)$$

Proof. Let $d = \min\{m,p\}$. We can upper bound the right hand side of Theorem E.4 via following proof steps of Theorem 4 of Simchowitz et al. [2020]:

$$\mathbf{f}^{\text{pred}}\mathbf{p.r.} - \left(\frac{6}{\alpha}\sum_{t=k+1}^{T}\|\boldsymbol{\epsilon}_t\|_2^2 - \frac{\alpha}{48}\sum_{t=1}^{T}\|\mathbf{M}_t - \mathbf{M}_{comp}\|_F^2\right) \lesssim \frac{L^2 {H'}^3 d\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{dL\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right)$$

$$\mathbf{f}^{\text{pred}}\mathbf{p.r.} + \frac{\alpha}{48}\sum_{t=1}^{T}\|\mathbf{M}_t - \mathbf{M}_{comp}\|_F^2 \lesssim \frac{L^2 {H'}^3 d\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{dL\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right)$$
$$+ \frac{1}{\alpha}\sum_{t=T_{burn}+1}^{T} C_{\text{approx}}^2 \epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right), \qquad (30)$$

where (30) follows from (28). ☐

Comparator Approximation Error:

Lemma D.2. *Suppose that $H' \geq 2H_0' - 1 + H$, $\psi_{\mathbf{G}}(H+1) \leq 1/10T$. Then for all $\tau > 0$,*

$$\textbf{\textit{Comp. app. err.}} \leq 4L\kappa_y\kappa_u\kappa_{\mathcal{M}}$$
$$+ \sum_{t=T_{burn}+1}^{T}\left[\tau\|\mathbf{M}_t - \mathbf{M}_{\text{comp}}\|_F^2 + 8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)\right]$$

Proof. The lemma can be proven using the proof of Proposition 8.2 of Simchowitz et al. [2020]. Combining Lemma E.3 and Lemma E.4 in Simchowitz et al. [2020],

$$\sum_{t=T_{burn}+1}^{T} f_t^{\text{pred}}(\mathbf{M}_{comp}) - \inf_{\mathbf{M}\in\mathcal{M}_0}\sum_{t=T_{burn}+1}^{T} f_t(\mathbf{M}, \mathbf{G}, b_1(\mathbf{G}), \ldots, b_t(\mathbf{G}))$$
$$\leq 4L\kappa_y \sum_{t=T_{burn}+1}^{T}\epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)\kappa_{\mathcal{M}}^2\kappa_b\left(\kappa_{\mathcal{M}} + \frac{\kappa_b}{4\tau}\right) + \kappa_u\kappa_{\mathcal{M}}\psi_{\mathbf{G}}(H+1) + (H+H')\tau\|\mathbf{M}_t - \mathbf{M}_{\text{comp}}\|_F^2$$
$$\leq \sum_{t=T_{burn}+1}^{T}\left[\tau\|\mathbf{M}_t - \mathbf{M}_{\text{comp}}\|_F^2 + 8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)\right]$$
$$+ 4TL\kappa_y\kappa_u\kappa_{\mathcal{M}}\psi_{\mathbf{G}}(H+1)$$
$$\leq 4L\kappa_y\kappa_u\kappa_{\mathcal{M}} + \sum_{t=T_{burn}+1}^{T}\left[\tau\|\mathbf{M}_t - \mathbf{M}_{\text{comp}}\|_F^2 + 8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{L^2}{\tau}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)\right]$$

☐

Combining all the terms bounded above, with $\tau = \frac{\alpha}{48}$ gives

$$\begin{aligned}
&\textsc{Regret}(T) \\
&\lesssim T_{burn} L\kappa_y^2 + L\kappa_y\kappa_u/10 + L\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}^2\kappa_b^2/10 + 2L\kappa_{\mathcal{M}}\kappa_y\kappa_{\mathbf{G}}\kappa_b + 4L\kappa_y\kappa_u\kappa_{\mathcal{M}} \\
&+ \frac{L^2 H'^3 \min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) + \frac{1}{\alpha}\sum_{t=T_{burn}+1}^{T} C_{\text{approx}}^2 \epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right) \\
&+ \sum_{t=T_{burn}+1}^{T} 8\kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{48L^2}{\alpha}\right\}\epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right) \\
&\lesssim T_{burn} L\kappa_y^2 \\
&+ \frac{L^2 H'^3 \min\{m,p\}\kappa_b^4\kappa_{\mathbf{G}}^4\kappa_{\mathcal{M}}^2}{\min\{\alpha, L\kappa_b^2\kappa_{\mathbf{G}}^2\}}\left(1+\frac{\overline{\alpha}_{loss}}{\min\{m,p\}L\kappa_{\mathcal{M}}}\right)\log\left(\frac{T}{\delta}\right) \\
&+ \sum_{t=T_{burn}+1}^{T} \epsilon_{\mathbf{G}}^2\left(\left\lceil \log_2\left(\frac{t}{T_{burn}}\right)\right\rceil, \delta\right)\left\{\frac{H'\kappa_{\mathbf{G}}^2\kappa_{\mathcal{M}}^2\kappa_b^4\left(\overline{\alpha}_{loss}+L\right)^2}{\alpha} + \kappa_y^2\kappa_b^2\kappa_{\mathcal{M}}^2(H+H')\max\left\{L, \frac{48L^2}{\alpha}\right\}\right\}
\end{aligned}$$

□

E Technical Lemmas and Theorems

Theorem E.1 (Matrix Azuma [Tropp, 2012])**.** *Consider a finite adapted sequence $\{\boldsymbol{X}_k\}$ of self-adjoint matrices in dimension d, and a fixed sequence $\{\boldsymbol{A}_k\}$ of self-adjoint matrices that satisfy*

$$\mathbb{E}_{k-1}\boldsymbol{X}_k = \mathbf{0} \text{ and } \boldsymbol{A}_k^2 \succeq \boldsymbol{X}_k^2 \text{ almost surely.}$$

Compute the variance parameter

$$\sigma^2 := \left\| \sum_k \boldsymbol{A}_k^2 \right\|$$

Then, for all $t \geq 0$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_k \boldsymbol{X}_k\right) \geq t\right\} \leq d \cdot \mathrm{e}^{-t^2/8\sigma^2}$$

Theorem E.2 (Closed-Loop Identification [Lale et al., 2020b])**.** *Let $\widehat{\mathcal{G}}_{\mathbf{yu,i}}$ be the solution to (11) at the beginning of epoch i. For $H_{est} \geq \max\{2n+1, \frac{\log(c_H T^2\sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}\}$, define*

$$V_{t_i} = V + \sum_{k=H_{est}}^{t_i} \phi_k \phi_k^\top$$

where $V = \lambda I$. Let $\|\mathcal{G}_{\mathbf{yu}}\|_F \leq S$. For $\delta \in (0,1)$, with probability at least $1-\delta$, for all i, $\mathcal{G}_{\mathbf{yu}}$ lies in the set $\mathcal{C}_{\mathcal{G}_{\mathbf{yu}}}(i)$, where

$$\mathcal{C}_{\mathcal{G}_{\mathbf{yu}}}(i) = \{\mathbf{M}' : \mathrm{Tr}((\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}')V_t(\widehat{\mathcal{G}}_{\mathbf{yu,i}} - \mathcal{G}_{\mathbf{yu}}')^\top) \leq \beta_i\},$$

for β_t defined as follows,

$$\beta_i = \left(\sqrt{m\|C\Sigma C^\top + \overline{\sigma}_z^2 I\| \log\left(\frac{\det(V_{t_i})^{1/2}}{\delta \det(V)^{1/2}}\right)} + S\sqrt{\lambda} + \frac{t_i\sqrt{H_{est}}}{T^2} \right)^2 .$$

Theorem E.3 (System Parameter Estimation Error using SysId [Lale et al., 2020b])**.** *Let $\mathcal{H}$ be the concatenation of two Hankel matrices obtained from $\mathcal{G}_{\mathbf{yu}}$. Let $\bar{A}, \bar{B}, \bar{C}, \bar{L}$ be the system parameters that* SysId *provides for $\mathcal{G}_{\mathbf{yu}}$. At time step t, let $\hat{A}_t, \hat{B}_t, \hat{C}_t, \hat{L}_t$ denote the system parameters obtained by* SysId *using the least squares estimate of the truncated ARX model, $\widehat{\mathcal{G}}_{\mathbf{yu,i}}$. Suppose Assumptions 2.2 and 2.3 hold, thus $\mathcal{H}$ is rank-n. After long enough warm-up period of T_{burn}, for the given choice of H_{est}, there exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that, with high probability, $\bar{\Theta} = (\bar{A}, \bar{B}, \bar{C}) \in (\mathcal{C}_A \times \mathcal{C}_B \times \mathcal{C}_C)$ where*

$$\begin{aligned}
\mathcal{C}_A(t) &= \left\{ A' \in \mathbb{R}^{n\times n} : \|\hat{A}_t - \mathbf{T}^\top A' \mathbf{T}\| \le \beta_A(t) \right\}, \\
\mathcal{C}_B(t) &= \left\{ B' \in \mathbb{R}^{n\times p} : \|\hat{B}_t - \mathbf{T}^\top B'\| \le \beta_B(t) \right\}, \\
\mathcal{C}_C(t) &= \left\{ C' \in \mathbb{R}^{m\times n} : \|\hat{C}_t - C' \mathbf{T}\| \le \beta_C(t) \right\},
\end{aligned}$$

for

$$\beta_A(t) = c_1 \left(\frac{\sqrt{nH}(\|\mathcal{H}\| + \sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})} \right) \|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|, \quad \beta_B(t) = \beta_C = \sqrt{\frac{20nH}{\sigma_n(\mathcal{H})}} \|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|, \tag{31}$$

for some problem dependent constant c_1.

Theorem E.4 (Theorem 8 of Simchowitz et al. [2020])**.** *Suppose that $\mathcal{K} \subset \mathbb{R}^d$ and $h \ge 1$. Let $F_t := \mathcal{K}^{h+1} \to \mathbb{R}$ be a sequence of L_c coordinatewise-Lipschitz functions with the induced unary functions $f_t(x) := F_t(x, \ldots, x)$ which are L_f-Lipschitz and β-smooth. Let $f_{t;k}(x) := \mathbb{E}\left[f_t(x) | \mathcal{F}_{t-k}\right]$ be α-strongly convex on $\mathcal{K}$ for a filtration $(\mathcal{F}_t)_{t\ge 1}$. Suppose that $z_{t+1} = \Pi_\mathcal{K}(z_t - \eta \boldsymbol{g}_t)$, where $\boldsymbol{g}_t = \nabla f_t(z_t) + \epsilon_t$ for $\|\boldsymbol{g}_t\|_2 \le L_\mathbf{g}$, and $\mathrm{Diam}(\mathcal{K}) \le D$. Let the gradient descent iterates be applied for $t \ge t_0$ for some $t_0 \le k$, with $z_0 = z_1 = \cdots = z_{t_0} \in \mathcal{K}$ for $k \ge 1$. Then with step size $\eta_t = \frac{3}{\alpha t}$, the following bound holds with probability $1-\delta$ for all comparators $z_\star \in \mathcal{K}$ simultaneously:*

$$\begin{aligned}
&\sum_{t=k+1}^{T} f_t(z_t) - f_t(z_\star) - \left(\frac{6}{\alpha} \sum_{t=k+1}^{T} \|\boldsymbol{\epsilon}_t\|_2^2 - \frac{\alpha}{12} \sum_{t=1}^{T} \|z_t - z_\star\|_2^2 \right) \\
&\qquad \lesssim \alpha k D^2 + \frac{\left(k L_\mathrm{f} + h^2 L_\mathrm{c}\right) L_\mathrm{g} + k d L_\mathrm{f}^2 + k\beta L_\mathrm{g}}{\alpha} \log(T) + \frac{k L_\mathrm{f}^2}{\alpha} \log\left(\frac{1 + \log\left(e + \alpha D^2\right)}{\delta} \right)
\end{aligned}$$

Lemma E.1 (Norm of a subgaussian vector [Abbasi-Yadkori and Szepesvári, 2011])**.** *Let $v \in \mathbb{R}^d$ be a entry-wise R-subgaussian random variable. Then with probability $1-\delta$, $\|v\| \le R\sqrt{2d\log(2d/\delta)}$.*

Lemma E.2 (Lemma 8.1 of Simchowitz et al. [2020])**.** *For any $\mathbf{M} \in \mathcal{M}$, let $f_t^{pred}(\mathbf{M})$ denote the unary counterfactual loss function induced by true truncated counterfactuals (Definition 8.1* *of Simchowitz et al. [2020]). During the i'th epoch of adaptive control period, at any time step* $t \in [t_i, \ \ldots, \ t_{i+1}-1]$*, for all* i*, we have that*

$$\left\|\nabla f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right) - \nabla f_t^{pred}\left(\mathbf{M}\right)\right\|_{\mathrm{F}} \leq C_{approx}\, \epsilon_{\mathbf{G}}(i, \delta),$$

where $C_{approx} \coloneqq \sqrt{H'}\kappa_{\mathbf{G}}\kappa_{\mathcal{M}}\kappa_b^2\left(16\overline{\alpha}_{loss} + 24L\right)$.

Lemma E.3 (Lemma 8.2 of Simchowitz et al. [2020])**.** *For any* $\mathbf{M} \in \mathcal{M}$*,* $f_t^{pred}\left(\mathbf{M}\right)$ *is* β*-smooth, where* $\beta = 16H'\kappa_b^2\kappa_{\mathbf{G}}^2\overline{\alpha}_{loss}$.

Lemma E.4 (Lemma 8.3 of Simchowitz et al. [2020])**.** *For any* $\mathbf{M} \in \mathcal{M}$*, given* $\epsilon_{\mathbf{G}}(i, \delta) \leq \frac{1}{4\kappa_b\kappa_{\mathcal{M}}\kappa_{\mathbf{G}}}\sqrt{\frac{\alpha}{H'\underline{\alpha}_{loss}}}$*, conditional unary counterfactual loss function induced by true counterfactuals are* $\alpha/4$ *strongly convex.*

Lemma E.5 (Lemma 8.4 of Simchowitz et al. [2020])**.** *Let* $L_f = 4L\sqrt{H'}\kappa_b^2\kappa_{\mathbf{G}}^2\kappa_{\mathcal{M}}$*. For any* $\mathbf{M} \in \mathcal{M}$ *and for* $T_{burn} \geq T_{\max}$*,* $f_t^{pred}(\mathbf{M})$ *is* $4L_f$*-Lipschitz,* $f_t^{pred}[\mathbf{M}_{t:t-H}]$ *is* $4L_f$ *coordinate Lipschitz. Moreover,* $\max_{\mathbf{M}\in\mathcal{M}}\left\|\nabla f_t\left(\mathbf{M}, \widehat{\mathbf{G}}_i, b_1(\widehat{\mathbf{G}}_i), \ldots, b_t(\widehat{\mathbf{G}}_i)\right)\right\|_2 \leq 4L_f$.