Regret Bound of Adaptive Control in Linear Quadratic Gaussian (LQG) 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 quadratic Gaussian control systems, where the model dynamics are unknown a priori. We propose LqgOpt, a novel reinforcement learning algorithm based on the principle of optimism in the face of uncertainty, to effectively minimize the overall control cost. We employ the predictor state evolution representation of the system dynamics and propose a new approach for closed-loop system identification, estimation, and confidence bound construction. LqgOpt efficiently explores the system dynamics, estimates the model parameters up to their confidence interval, and deploys the controller of the most optimistic model for further exploration and exploitation. We provide stability guarantees for LqgOpt, and prove the regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ for adaptive control of linear quadratic Gaussian (LQG) systems, where T is the time horizon of the problem.

1 Introduction

One of the core challenges in the field of control theory and reinforcement learning is adaptive control. It is the problem of controlling dynamical systems when the dynamics of the systems are unknown to the decision-making agents. In adaptive control, agents interact with given systems in order to explore and control them while the long-term objective is to minimize the overall average associated costs. The agent has to balance between *exploration* and *exploitation*, learn the dynamics, strategize for further exploration, and exploit the estimation to minimize the overall costs. The sequential nature of agent-system interaction results in challenges in the system identifying, estimation, and control under uncertainty, and these challenges are magnified when the systems are partially observable, *i.e.* contain hidden underlying dynamics.

In the linear systems, when the underlying dynamics are fully observable, the asymptotic optimality of estimation methods has been the topic of study in the last decades [Lai et al., 1982, Lai and Wei, 1987]. Recently, novel techniques and learning algorithms have been developed to study the finite-time behavior of adaptive control algorithms and shed light on the design of optimal methods [Peña et al., 2009, Fiechter, 1997, Abbasi-Yadkori and Szepesvári, 2011]. In particular, Abbasi-Yadkori and Szepesvári [2011] proposes to use the principle of optimism in the face of uncertainty (OFU) to balance exploration and exploitation in LQR, where the state of the system is observable. OFU principle suggests to estimate the model parameters up to their confidence

interval, and then act according to the policy advised by the model in confidence set with the lowest optimal cost, known as the optimistic model.

When the underlying dynamics of linear systems are partially observable, estimating the systems' dynamics requires considering and analyzing unobservable events, resulting in a series of significant challenges in learning and controlling the partially observable systems. A line of prior works are dedicated to the problem of open-loop model estimation Oymak and Ozay [2018], Sarkar et al. [2019], Tsiamis and Pappas [2019] where the proposed methods highly rely on random excitation, uncorrelated Gaussian noise, and do not allow feedback control. Additionally, in general, computing the optimal controller requires inferring the latent state of the system, given the history of observations. When the model dynamics are not known precisely, the uncertainties in the system estimation result in inaccurate latent state estimation and inaccurate linear controller. The possibility of accumulation of these errors creates a challenging problem in adaptive control of partially observable linear systems. Therefore, we need to consider these challenges in designing an algorithm that performs desirably. In this work, we employ *regret*, a metric in quantifying the performance of learning algorithms that measures the difference between the cost encountered by an adaptive control agent and that of an optimal controller, knowing the underlying system [Lai and Robbins, 1985].

Contributions: In this work, we study the adaptive control of partially observable linear systems from both model estimation/system identification and the controller synthesis perspective. We introduce a novel estimation method for the general cases of both closed- and open-loop identification of linear dynamical systems with unobserved hidden states, even in the presence of feedback loop and correlated Gaussian noise. We provide the detailed finite time estimation analysis and construction of confidence sets.

We propose LqgOpt, an adaptive control algorithm for learning and controlling unknown partially observable linear systems with quadratic cost and Gaussian disturbances, i.e., linear quadratic Gaussian (LQG), for which optimal control exists and has a closed form [Bertsekas, 1995]. LqgOpt interacts with the system, collects samples, estimates the model parameters, and adapts accordingly. LqgOpt deploys OFU principle to balance the *exploration* vs. *exploitation* trade-off. Using the predictor form of the state-space equations of the partially observable linear systems, we define a least-squares estimation problem and obtain confidence sets on the system parameters. LqgOpt then uses these confidence sets to find the optimistic model and use the optimal controller for the chosen model for further exploration-exploitation. To analyze the finite-time regret of LqgOpt, we first provide a stability analysis for the sequence of optimistic controllers. Finally, we prove that LqgOpt achieves a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$, an improvement to the $\tilde{\mathcal{O}}(T^{2/3})$ regret upper bound in the prior work Lale et al. [2020], where T is the number of total interactions.

Independently and simultaneously to our paper, a new arxiv paper Simchowitz et al. [2020] propose an algorithm which also achieves $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound in partially observable linear systems, under different problem setup and semi-adversarial disturbances. Simchowitz et al. [2020] employ the theory of online learning, and propose to start with an initial phase of pure exploration, long enough for accurate predictive-model estimation. Then this phase is followed by committing to the learned predictive-model and deploying online learning for the policy updates. Under a set of different assumptions, e.g., on noise model, access to a set of stabilizing controllers, computation, the convexity of loss functions, the authors show that their method attains a similar order regret bound. These two works develop a principally different set of theoretical tools and analyses that require further investigation.

2 Preliminaries

We denote the Euclidean norm of a vector x as $\|x\|_2$. We denote $\rho(A)$ as the spectral radius of a matrix A, $\|A\|_F$ as its Frobenius norm and $\|A\|_2$ as its spectral norm. $\mathrm{Tr}(A)$ is its trace, $A^\top$ is the transpose, $A^\dagger$ is the Moore-Penrose inverse. 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 represents the identity matrix with the appropriate dimensions.

Consider the following discrete time linear time-invariant system $\Theta = (A, B, C)$ and with dynamics as:

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

At each time step t, the system is at (hidden) state $x_t \in \mathbb{R}^n$, the agent receives observation $y_t \in \mathbb{R}^m$ under a measurement noise $z_t \sim \mathcal{N}\left(0, \sigma_z^2 I\right)$. Then the agent applies a control input $u_t \in \mathbb{R}^p$, and receives a cost of $c_t = y_t^\top Q y_t + u_t^\top R u_t$ where Q and R are positive semidefinite and positive definite matrices, respectively. After taking u_t, the state of the system evolves to x_{t+1} for the time step $t+1$ under a process noise $w_t \sim \mathcal{N}\left(0, \sigma_w^2 I\right)$. Here the noises are i.i.d. random vectors and $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ.

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

$$\mathbf{C}(A, B, n) = [B \;\; AB \;\; A^2B \ldots A^{n-1}B]$$

has full row rank. For all $H \geq n$, $\mathbf{C}(A, B, H)$ defines the extended (A, B) controllability matrix. Similarly, a linear system $\Theta = (A, B, C)$ is A, C observable if the observability matrix,

$$\mathbf{O}(A, C, n) = [C^\top \;\; (CA)^\top \;\; (CA^2)^\top \ldots (CA^{n-1})^\top]^\top$$

has full column rank. For all $H \geq n$, $\mathbf{O}(A, C, H)$ defines the extended (A, C) observability matrix.

Suppose the underlying system is controllable and observable. Then, the agent chooses control inputs as a function of past observations and aims to minimize the expected cost,

$$J_\star(\Theta) = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} y_t^\top Q y_t + u_t^\top R u_t\right].$$

This problem is known as LQG control. The optimal solution to LQG control problem is a linear feedback control policy given as $u_t = -K\hat{x}_{t|t,\Theta}$. Here K is the optimal feedback gain matrix,

$$K = \left(R + B^\top P B\right)^{-1} B^\top P A,$$

where P is the unique positive semidefinite solution to the following discrete-time algebraic Riccati equation (DARE):

$$P = A^\top P A + C^\top Q C - A^\top P B \left(R + B^\top P B\right)^{-1} B^\top P A, \tag{2}$$

and $\hat{x}_{t|t,\Theta}$ is the minimum mean square error (MMSE) estimate of the underlying state using system parameters Θ and past observations, where $\hat{x}_{0|-1,\Theta} = 0$. At steady-state, this estimate is efficiently obtained by using the Kalman filter:

$$\hat{x}_{t|t,\Theta} = (I - LC)\,\hat{x}_{t|t-1,\Theta} + Ly_t, \tag{3}$$

$$\hat{x}_{t|t-1,\Theta} = (A\hat{x}_{t-1|t-1,\Theta} + Bu_{t-1}), \tag{4}$$

$$L = \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}, \tag{5}$$

where Σ is the unique positive semidefinite solution to the following DARE:

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

In the adaptive control, the underlying system parameters Θ are unknown, and the agent needs to learn them through interaction with the system with the aim of minimizing the cumulative costs $\sum_{t=1}^{T} c_t$ after T time steps. We measure the performance of the agent using regret, *i.e.,* the difference between the agent's cost and the optimal expected cost:

$$\text{REGRET}(T) = \sum_{t=0}^{T} (c_t - J_*(\Theta)).$$

The system characterization depicted in (1) is called state-space form of the system Θ. The same discrete time linear time-invariant system can be represented in several ways which has been considered in various works in control theory and reinforcement learning [Kailath et al., 2000, Tsiamis et al., 2019]. Note that these representations all have the same second order statistics. One of the most common form is the innovations form[1] of the system characterized as

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + Fe_t \\ y_t &= Cx_t + e_t. \end{aligned} \tag{6}$$

where $F = AL$ is the Kalman gain in the observer form and e_t is the zero mean white innovation process. In this equivalent representation of system, the state x_t can be seen as the estimate of the state in the state space representation, which is the expression stated in (4). In the steady state, $e(t) \sim \mathcal{N}\left(0, C\Sigma C^\top + \sigma_z^2 I\right)$. Using the relationship between e_t and y_t, we obtain the following characterization of the system Θ, known as the predictor form of the system,

$$\begin{aligned} x_{t+1} &= \bar{A}x_t + Bu_t + Fy_t \\ y_t &= Cx_t + e_t \end{aligned} \tag{7}$$

where $\bar{A} = A - FC$ and $F = AL$. Notice that at steady state, the predictor form allows the current output y_t to be described by the history of inputs and outputs with an i.i.d. Gaussian disturbance $e_t \sim \mathcal{N}\left(0, C\Sigma C^\top + \sigma_z^2 I\right)$. In this paper, we exploit these fundamental properties to estimate the underlying system, even with feedback control. We consider the set of stable systems.

[1]For simplicity, all of the system representations are presented for the steady-state of the system.

Assumption 2.1. *The system is order n and minimal in the sense that the system cannot be described by a state-space model of order less than n. The system is stable, i.e. $\rho(A) < 1$ and $\Phi(A) := \sup_{\tau \geq 0} \frac{\|A^\tau\|}{\rho(A)^\tau} < \infty$.*

Note that the assumption regarding $\Phi(A)$ is required for quantifying the finite time evolution of the system and it is a mild condition, *e.g.* if A is diagonalizable, $\Phi(A)$ is finite. Additionally for stable A, $\Phi(A)$ can be upper bounded by the $\mathcal{H}_\infty$-norm of the system $x_{t+1} = Ax_t + w_t$ [Mania et al., 2019].

We assume that the underlying system lives in the following set.

Assumption 2.2. *The unknown system $\Theta = (A, B, C)$ is a member of a set $\mathcal{S}$, such that,*

$$\mathcal{S} \subseteq \left\{ \Theta' = (A', B', C', F') \;\middle|\; \begin{array}{l} \rho(A') < 1, \\ (A', B') \text{ is controllable,} \\ (A', C') \text{ is observable,} \\ (A', F') \text{ is controllable.} \end{array} \right\}$$

The above assumptions are standard in system identification settings in order to ensure the possibility of accurate estimation of the system parameters [Knudsen, 2001, Oymak and Ozay, 2018, Tsiamis and Pappas, 2019, Sarkar et al., 2019, Tsiamis et al., 2019, Lale et al., 2020].

Assumption 2.3. *The set $\mathcal{S}$ consists of systems that are contractible, i.e.,*

$$\rho := \sup_{\Theta' = (A', B', C') \in \mathcal{S}} \left\| A' - B'K(\Theta') \right\| < 1,$$

where $K(\Theta')$ is the optimal feedback gain matrix of Θ', and

$$\upsilon := \sup_{\Theta' = (A', B', C') \in \mathcal{S}} \left\| A' - A'L(\Theta')C' \right\| < 1.$$

where $L(\Theta')$ is the optimal Kalman gain matrix of Θ'. There exists finite numbers D, Γ, ζ such that $D = \sup_{\Theta' \in \mathcal{S}} \|P(\Theta')\|$, $\Gamma = \sup_{\Theta' \in \mathcal{S}} \|K(\Theta')\|$ and $\zeta = \sup_{\Theta' \in \mathcal{S}} \|L(\Theta')\|$.

This assumption allows us to develop stability guarantees in the presence of sub-optimal closed-loop controllers.

3 Adaptive Control via LqgOpt

In this section, we present LqgOpt, an adaptive control algorithm for LQG control problems, and describe its compounding components. The outline of LqgOpt is given in Algorithm 1. The early stage of deploying LqgOpt involves a fixed warm-up period dedicated for pure exploration using Gaussian excitation. LqgOpt requires this exploration period to estimate the model parameters reliably enough that the controller designed based on the parameter estimation and their confidence set results in a stabilizing controller on the real system. The duration of this period depends on how stabilizable the true parameters are and how accurate the model estimations should be. We formally quantify these statements and the length of the warm-up period.

After the warm-up period, LqgOpt utilizes the model parameter estimations and their confidence sets to design a controller corresponding to an optimistic model in the confidence sets, obtained

by following the OFU principle. Due to the reliable estimation from the warm-up period, this controller and all the future designed controller stabilize the underlying true unknown model. The agent deploys the prescribed controller on the real system for exploration and exploitation. The agent collects samples throughout its interaction with the environment, and use these samples for further model estimation, confidence interval construction, and design of the controller regarding to an optimistic model. The agent repeats this process.

Since the Kalman filter converges exponentially fast to the steady-state gain in observer form, without loss of generality, we assume that $x_0 \sim \mathcal{N}(0, \Sigma)$, *i.e.*, the system starts at the steady-state. This consideration eases the presentation of the algorithm. We provide the overview of the analysis for any arbitrary and almost surely finite initialization in the Appendix G.

In the warm-up period LqgOpt excites the system with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $1 \leq t \leq T_w$. Considering the predictor form representation of the system given in (7), we can roll back the state evolution H time steps back as follows,

$$x_t = \sum_{k=0}^{H-1} \bar{A}^k \left(F y_{t-k-1} + B u_{t-k-1}\right) + \bar{A}^H x_{t-H}$$

From Assumption 2.2, we have that $\bar{A}$ is stable, thus the state can be estimated in principle for large enough H. Using the generated input-output sequence $\mathcal{D} = \{y_t, u_t\}_{t=1}^{T_w}$, LqgOpt constructs N subsequences of H input-output pairs, ϕ_t for $H \leq t \leq T_w$, where $T_w = H + N - 1$,

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

Using this definition, we can write the following truncated autoregressive exogenous (ARX) model for the given system Θ,

$$y_t = \mathbf{M}\phi_t + e_t + C\bar{A}^H x_{t-H} \tag{8}$$

where $\mathbf{M} \in \mathbb{R}^{m \times (m+p)H}$ defined as

$$\mathbf{M} = \left[CF,\; C\bar{A}F,\; \dots,\; C\bar{A}^{H-1}F,\; CB,\; C\bar{A}B,\; \dots,\; C\bar{A}^{H-1}B\right]. \tag{9}$$

Thus, any input-output trajectory $\{y_i, u_t\}_{t=1}^{T}$ can be represented as

$$Y_T = \Phi_T \mathbf{M}^\top + E_T + N_T \tag{10}$$

where

$$\begin{aligned}
Y_T &= [y_H,\; y_{H+1},\; \dots,\; y_T]^\top \in \mathbb{R}^{N \times m} \\
\Phi_T &= [\phi_H,\; \phi_{H+1},\; \dots,\; \phi_T]^\top \in \mathbb{R}^{N \times (m+p)H} \\
E_T &= [e_H,\; e_{H+1},\; \dots,\; e_T]^\top \in \mathbb{R}^{N \times m} \\
N_T &= \left[C\bar{A}^H x_0,\; C\bar{A}^H x_1, \dots, C\bar{A}^H x_{T-H}\right]^\top \in \mathbb{R}^{N \times m}
\end{aligned}$$

for $N = T - H + 1$.

Note that, during the warm-up period the noise terms are zero-mean including the effect of initial state since we assume that $x_0 \sim \mathcal{N}(0, \Sigma)$. After the warm-up period, LqgOpt obtains the

Algorithm 1 LqgOpt

Input: T_w, H, σ_o, σ_c, $S > 0$, $\delta > 0$, n, m, p, Q, R

—— Warm-Up ——

for $t = 0, 1, \ldots, T_w$ **do**

  Deploy $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ and store $\mathcal{D}_0 = \{y_t, u_t\}_{t=1}^{T_w}$

end for

—— Adaptive Control ——

for $i = 0, 1, \ldots$ **do**

  Calculate $\hat{\mathbf{M}}_{\mathbf{i}}$ using $\mathcal{D}_i = \{y_t, u_t\}_{t=1}^{2^i T_w}$

  Deploy SysId $(H, \hat{\mathbf{M}}_{\mathbf{i}}, n)$ for $\hat{A}_i, \hat{B}_i, \hat{C}_i, \hat{L}_i$

  Construct the confidence sets $\mathcal{C}_A(i), \mathcal{C}_B(i), \mathcal{C}_C(i)$, $\mathcal{C}_L(i)$ s.t. w.h.p. $(A, B, C, L) \in \mathcal{C}_i$, where $\mathcal{C}_i := (\mathcal{C}_A(i) \times \mathcal{C}_B(i) \times \mathcal{C}_C(i) \times \mathcal{C}_L(i))$

  Find a $\tilde{\Theta}_i = (\tilde{A}_i, \tilde{B}_i, \tilde{C}_i, \tilde{L}_i) \in \mathcal{C}_i \cap \mathcal{S}$ s.t.

$$J(\tilde{\Theta}_i) \leq \inf_{\Theta' \in \mathcal{C}_i \cap \mathcal{S}} J(\Theta') + T^{-1}$$

  for $t = 2^i T_w, \ldots 2^{i+1} T_w - 1$ **do**

    Execute the optimal controller for $\tilde{\Theta}_i$

  end for

end for

first estimate of the unknown truncated ARX model $\mathbf{M}$ by solving the following regularized least square problem,

$$\hat{\mathbf{M}}_{\mathbf{0}} = \arg\min_X \|Y_{T_w} - \Phi_{T_w} X^\top\|_F^2 + \lambda \|X\|_F^2 \tag{11}$$

where the solution

$$\hat{\mathbf{M}}_{\mathbf{0}}^\top = (\Phi_{T_w}^\top \Phi_{T_w} + \lambda I)^{-1} \Phi_{T_w}^\top Y_{T_w}.$$

Using this solution, LqgOpt deploys a system-identification algorithm and obtains the estimates of the system parameters $\hat{A}_0, \hat{B}_0, \hat{C}_0, \hat{L}_0$, with corresponding confidence sets $\mathcal{C}_A(0), \mathcal{C}_B(0), \mathcal{C}_C(0)$, $\mathcal{C}_L(0)$ in which the underlying system parameters live with high probability. With the initial confidence sets, LqgOpt starts adaptive control period using the OFU principle. It selects the optimistic model *i.e.*, the model that has the minimum average expected cost, among the plausible models and executes the optimal controller for the chosen model. As the confidence sets shrink, *i.e.*, the estimates of system parameters are *significantly* refined, LqgOpt adapts and updates its policy by deploying OFU principle on the new confidence sets.

For a linear system $\Theta = (A, B, C)$, we define truncated open-loop and closed-loop noise evolution parameters, respectively $\mathcal{G}^{ol}$ and $\mathcal{G}^{cl}$. When the controller is set to be i.i.d. Gaussian excitements, $\mathcal{G}^{ol} \in \mathbb{R}^{H(m+p) \times 2H(n+m+p)}$ encodes the open-loop evolution of the disturbances in the system, and represents the responses to these disturbances on the *batch* of observations and actions *history*. Note that the historical data is correlated even in the open-loop setting with i.i.d. Gaussian excitements. The exact definition of $\mathcal{G}^{ol}$ is provided in equation (20) of Appendix A.1. In Appendix A.1, we also show that $\mathcal{G}^{ol}$ is full row-rank, *i.e.*, $\sigma_{\min}(\mathcal{G}^{ol}) > \sigma_o > 0$, where σ_o is known to LqgOpt.

When the controller is set to be the optimal policy for the underlying system, *i.e.* closed-loop system, $\mathcal{G}^{cl} \in \mathbb{R}^{H(m+p) \times 2H(n+m)}$ represents the translation of the truncated history of process and measurement noises on the inputs, ϕ's. The exact construction of $\mathcal{G}^{cl}$ is provided in detail in equation

(24) of Appendix A.2. Briefly, it is formed by shifting a block matrix $\bar{\mathbf{G}} \in \mathbb{R}^{(m+p)\times 2H(n+m)}$ by $m+n$ in each block row where $\bar{\mathbf{G}}$ is constructed by H $(m+p)\times(n+m)$ matrices. We assume that H used in LqgOpt is large enough that $\bar{\mathbf{G}}$ is full row rank for the given system. In Appendix A.2, we show that, if $\bar{\mathbf{G}}$ is full row-rank, $\mathcal{G}^{cl}$ would be full row-rank, too. Thus, we have that for the choice of H in LqgOpt, $\sigma_{\min}(\mathcal{G}^{cl})$ is lower bounded by some positive value, *i.e.*, $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$, where LqgOpt only knows σ_c and searches for an optimistic system whose closed-loop noise evolution parameter satisfies this lower bound.

The following theorem states the main result of the paper, an end-to-end regret upper bound of the adaptive control in LQG systems.

Theorem 3.1 (Regret Upper Bound)**.** *Given a* LQG $\Theta = (A, B, C)$*, and regulating parameters Q and R, with high probability, the regret of* LqgOpt *with a warm-up duration of $T_w = poly(H, \log(T), \sigma_o, \sigma_c, \upsilon, \zeta, \Gamma, m, n, p, \rho, \Phi(A))$ is*

$$\text{REGRET}(T) = \tilde{\mathcal{O}}\left(\sqrt{T}\right) \tag{12}$$

The exact expressions that define T_w are given in Appendix with the detailed definitions.

3.1 Learning the Truncated ARX Model

First consider the effect of truncation bias term, N_t. From Assumption 2.3, we have that $\|\bar{A}\| \leq \upsilon < 1$. Thus, each term in N_t is order of υ^H. In order to get consistent estimation, for some problem dependent constant c_H, LqgOpt sets $H \geq \frac{\log(c_H T^2 \sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}$, resulting in a negligible bias term of order $1/T^2$. The following gives the self-normalized finite sample estimation error of (11).

Theorem 3.2 (Closed-Loop Identification)**.** *Let $\hat{\mathbf{M}}_\mathbf{t}$ be the solution to (11) at time t. For the given choice of H, define*

$$V_t = V + \sum_{i=H}^{t} \phi_i \phi_i^\top$$

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

$$\mathcal{C}_\mathbf{M}(t) = \{\mathbf{M}' : \text{Tr}((\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}')V_t(\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}')^\top) \leq \beta_t\},$$

for β_t defined as follows,

$$\beta_t = \left(\sqrt{m\|C\Sigma C^\top + \sigma_z^2 I\| \log\left(\frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}}\right)} + S\sqrt{\lambda} + \frac{t\sqrt{H}}{T^2}\right)^2.$$

The proof is given in Appendix B. It uses self-normalized tail inequalities to get the first two terms in the definition of β_t , and with the given choice of H, we obtain the final term in the bound. This bound can be translated to $\|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|$ in order to be utilized for the confidence set construction of the system parameters. First, we need the following lemmas that guarantee persistence of excitation during the warm-up period and adaptive control period.

Lemma 3.1 (Persistence of Excitation in Warm-Up Period)**.** *After sufficient time steps in warm-up period of* LqgOpt*, with probability at least* $1-\delta$*, we have*

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\frac{\sigma_o^2\min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2}. \tag{13}$$

Lemma 3.2 (Persistence of Excitation in Adaptive Control Period)**.** *After sufficient time steps in adaptive control period of* LqgOpt*, with probability* $1-3\delta$*, we have*

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}. \tag{14}$$

For two problem dependent parameters Υ_w and Υ_c, that uniformly bound the components of ϕ's during the warm-up and adaptive control period respectively, we have the following theorem which combines Theorem 3.2 with Lemma 3.1 and 3.2 to obtain the bound over $\|\hat{\mathbf{M}}_\mathbf{t}-\mathbf{M}\|$.

Theorem 3.3. *During the warm-up period,* $\|\phi_t\|\leq\Upsilon_w\sqrt{H}$ *with high probability. After the warm-up period of* T_w*, the initial estimation of the truncated ARX model,* $\hat{\mathbf{M}}_\mathbf{0}$*, obeys*

$$\|\hat{\mathbf{M}}_\mathbf{0}-\mathbf{M}\| \leq \frac{poly(m,H,p)}{\min\{\sigma_w,\sigma_z,\sigma_u\}\sigma_o\sqrt{T_w}}.$$

During the adaptive control, with high probability $\|\Phi_t\|\leq\Upsilon_c\sqrt{H}$*. For the adaptive control period at any time* $t\geq 2T_w$*, the least squares estimate of the truncated ARX model* $\hat{\mathbf{M}}_\mathbf{t}$ *follows*

$$\|\hat{\mathbf{M}}_\mathbf{t}-\mathbf{M}\| \leq \frac{poly(m,H,p)}{\sqrt{t\min\{\sigma_o^2\sigma_w^2,\sigma_o^2\sigma_z^2,\sigma_o^2\sigma_u^2,\frac{\sigma_c^2\sigma_w^2}{8},\frac{\sigma_c^2\sigma_z^2}{8}\}}}.$$

Note that the choice of H depends on the horizon, which is needed to be known apriori. Since the dependency of H in the horizon T is $\log(T)$, one can deploy the standard doubling trick to relax this requirement.[2]

3.2 System Identification

After estimating $\hat{\mathbf{M}}_\mathbf{t}$, LqgOpt constructs confidence sets for the unknown system parameters and uses these confidence sets to come up with the optimistic controller to exploit the information gathered. LqgOpt uses a subspace identification algorithm SysId, given in Algorithm 2 in Appendix C. SysId is similar to Ho-Kalman method [Ho and Kálmán, 1966] and estimates the system parameters from $\hat{\mathbf{M}}_\mathbf{t}$. First of all, notice that $\mathbf{M}=[\mathbf{F},\mathbf{G}]$ where

$$\mathbf{F} = \begin{bmatrix}CF,\ C\bar{A}F,\ \ldots,\ C\bar{A}^{H-1}F\end{bmatrix}\in\mathbb{R}^{m\times mH},$$
$$\mathbf{G} = \begin{bmatrix}CB,\ C\bar{A}B,\ \ldots,\ C\bar{A}^{H-1}B\end{bmatrix}\in\mathbb{R}^{m\times pH}.$$

Given the estimate for the truncated ARX model

$$\hat{\mathbf{M}}_\mathbf{t} = [\hat{\mathbf{F}}_{\mathbf{t},\mathbf{1}},\ldots,\hat{\mathbf{F}}_{\mathbf{t},\mathbf{H}},\hat{\mathbf{G}}_{\mathbf{t},\mathbf{1}},\ldots,\hat{\mathbf{G}}_{\mathbf{t},\mathbf{H}}],$$

[2]Doubling trick suggests to set the horizon to a time step, and in a repeated fashion, whenever that time step is reached, double that time step, and continue.

where $\hat{\mathbf{F}}_{\mathbf{t},\mathbf{i}}$ is the i'th $m \times m$ block of $\hat{\mathbf{F}}_{\mathbf{t}}$, and $\hat{\mathbf{G}}_{\mathbf{t},\mathbf{i}}$ is the i'th $m \times p$ block of $\hat{\mathbf{G}}_{\mathbf{t}}$ for all $1 \leq i \leq H$, SysId constructs two $d_1 \times (d_2+1)$ Hankel matrices $\mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}}$ and $\mathcal{H}_{\hat{\mathbf{G}}_{\mathbf{t}}}$ such that (i,j)th block of Hankel matrix is $\hat{\mathbf{F}}_{\mathbf{t},\mathbf{i}}$ and $\hat{\mathbf{G}}_{\mathbf{t},\mathbf{i}}$ respectively. Then, it forms the following matrix $\hat{\mathcal{H}}_t$.

$$\hat{\mathcal{H}}_t = \left[\mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}},\ \mathcal{H}_{\hat{\mathbf{G}}_{\mathbf{t}}}\right].$$

Recall that the dimension of latent state, n, is the order of the system for the observable and controllable system. For $H \geq \max\left\{2n+1, \frac{\log(c_H T^2\sqrt{m}/\sqrt{\lambda})}{\log(1/\upsilon)}\right\}$, we can pick $d_1 \geq n$ and $d_2 \geq n$ such $d_1 + d_2 + 1 = H$. This guarantees that the system identification problem is well-conditioned. Using Definition 2.1, if the input to the SysId was $\mathbf{M} = [\mathbf{F}, \mathbf{G}]$ then constructed Hankel matrix, $\mathcal{H}$ would be rank n,

$$\begin{aligned}
\mathcal{H} &= [C^\top,\ \ldots,\ (C\bar{A}^{d_1-1})^\top]^\top [F,\ \ldots,\ \bar{A}^{d_2}F,\ B,\ \ldots,\ \bar{A}^{d_2}B] \\
&= \mathbf{O}(\bar{A}, C, d_1)\ [\mathbf{C}(\bar{A}, F, d_2+1),\quad \bar{A}^{d_2}F,\quad \mathbf{C}(\bar{A}, B, d_2+1),\quad \bar{A}^{d_2}B] \\
&= \mathbf{O}(\bar{A}, C, d_1)\ [F,\quad \bar{A}\mathbf{C}(\bar{A}, F, d_2+1),\quad B,\quad \bar{A}\mathbf{C}(\bar{A}, B, d_2+1)].
\end{aligned}$$

Notice that $\mathbf{M}$ and $\mathcal{H}$ are uniquely identifiable for a given system Θ, whereas for any invertible $\mathbf{T} \in \mathbb{R}^{n\times n}$, the system resulting from

$$A' = \mathbf{T}^{-1}A\mathbf{T},\ B' = \mathbf{T}^{-1}B,\ C' = C\mathbf{T},\ L' = \mathbf{T}^{-1}L$$

gives the same $\mathbf{M}$ and $\mathcal{H}$. Similar to Ho-Kalman algorithm, SysId computes the SVD of $\hat{\mathbf{M}}_{\mathbf{t}}$ and estimates the extended observability and controllability matrices and eventually system parameters up to similarity transformation. To this end, SysId constructs $\hat{\mathcal{H}}_t^-$ by discarding (d_2+1)th and $(2d_2+2)$th block columns of $\hat{\mathcal{H}}_t$, *i.e.* if it was $\mathcal{H}$ then we have,

$$\mathcal{H}^- = \mathbf{O}(\bar{A}, C, d_1)\ [\mathbf{C}(\bar{A}, F, d_2+1),\quad \mathbf{C}(\bar{A}, B, d_2+1)].$$

The algorithm then calculates $\hat{\mathcal{N}}_t$, the best rank-n approximation of $\hat{\mathcal{H}}_t^-$, obtained by setting its all but top n singular values to zero. The estimates of $\mathbf{O}(\bar{A}, C, d_1)$, $\mathbf{C}(\bar{A}, F, d_2+1)$ and $\mathbf{C}(\bar{A}, B, d_2+1)$ are given as

$$\hat{\mathcal{N}}_t = \mathbf{U}_{\mathbf{t}}\mathbf{\Sigma}_{\mathbf{t}}{}^{1/2}\ \mathbf{\Sigma}_{\mathbf{t}}{}^{1/2}\mathbf{V}_{\mathbf{t}}{}^\top = \hat{\mathbf{O}}_{\mathbf{t}}(\bar{A}, C, d_1)\ [\hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, F, d_2+1),\quad \hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, B, d_2+1)].$$

From these estimates SysId recovers $\hat{C}_t$ as the first $m \times n$ block of $\hat{\mathbf{O}}_{\mathbf{t}}(\bar{A}, C, d_1)$, $\hat{B}_t$ as the first $n \times p$ block of $\hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, B, d_2+1)$ and $\hat{F}_t$ as the first $n \times m$ block of $\hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, F, d_2+1)$. Let $\hat{\mathcal{H}}_t^+$ be the matrix obtained by discarding 1st and (d_2+2)th block columns of $\hat{\mathcal{H}}_t$, *i.e.* if it was $\mathcal{H}$ then

$$\mathcal{H}^+ = \mathbf{O}(\bar{A}, C, d_1)\ \bar{A}\ [\hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, F, d_2+1),\quad \hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, B, d_2+1)].$$

Therefore, SysId recovers

$$\hat{\bar{A}}_t = \hat{\mathbf{O}}_{\mathbf{t}}^\dagger(\bar{A}, C, d_1)\ \hat{\mathcal{H}}_t^+\ [\hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, F, d_2+1),\quad \hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, B, d_2+1)]^\dagger.$$

Using the definition of $\bar{A} = A - FC$, the algorithm obtains $\hat{A}_t = \hat{\bar{A}}_t + \hat{F}_t\hat{C}_t$. Recall that $F = AL$. Using the Assumption 2.2, SysId finally recovers $\hat{L}_t$ as the first $n \times m$ block of $\hat{A}_t^\dagger\hat{\mathbf{O}}_{\mathbf{t}}^\dagger(\bar{A}, C, d_1)\hat{\mathcal{H}}_t^-$. The following theorem essentially translates the bound in Theorem 3.2 to individual bounds of system parameter estimates. It provides the high probability confidence sets required for deploying OFU principle for the adaptive control.

Theorem 3.4 (Confidence Set Construction)**.** *Let $\mathcal{H}$ be the concatenation of two Hankel matrices obtained from $\mathbf{M}$. Let $\bar{A}, \bar{B}, \bar{C}, \bar{L}$ be the system parameters that* SysId *provides for $\mathbf{M}$. 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, $\hat{\mathbf{M}}_\mathbf{t}$. Suppose Assumptions 2.1 and 2.2 hold, thus $\mathcal{H}$ is rank-n. After the warm-up period of T_w, for the given choice of H, there exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that, with high probability, $\bar{\Theta} = (\bar{A}, \bar{B}, \bar{C}, \bar{L}) \in (\mathcal{C}_A \times \mathcal{C}_B \times \mathcal{C}_C \times \mathcal{C}_L)$ where*

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

for

$$\beta_A(t) = c_1 \left(\frac{\sqrt{nH}(\|\mathcal{H}\| + \sigma_n(\mathcal{H}))}{\sigma_n^2(\mathcal{H})} \right) \|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|, \quad \beta_B(t) = \beta_C = \sqrt{\frac{20nH}{\sigma_n(\mathcal{H})}} \|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|, \tag{15}$$
$$\beta_L(t) = \frac{c_2\|\mathcal{H}\|}{\sqrt{\sigma_n(\mathcal{H})}}\beta_A + c_3 \frac{\sqrt{nH}(\|\mathcal{H}\| + \sigma_n(\mathcal{H}))}{\sigma_n^{3/2}(\mathcal{H})} \|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|,$$

for some problem dependent constants c_1, c_2 and c_3.

The proof is given in the Appendix C. It combines Lemma B.1 of Oymak and Ozay [2018] with careful perturbation analysis on the system parameter estimates provided by SysId.

3.3 Adaptive Control

Using the confidence sets, LqgOpt implements OFU principle. At time t, the algorithm chooses a system $\tilde{\Theta}_t = (\tilde{A}_t, \tilde{B}_t, \tilde{C}_t, \tilde{L}_t)$ from $\mathcal{C}_t \cap \mathcal{S}$ where $\mathcal{C}_t := (\mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t))$ such that

$$J(\tilde{\Theta}_t) \le \inf_{\Theta' \in \mathcal{C}_t \cap \mathcal{S}} J(\Theta') + 1/T. \tag{16}$$

The algorithm designs the optimal feedback policy $(\tilde{P}_t, \tilde{K}_t, \tilde{L}_t)$ for the chosen system $\tilde{\Theta}_t$. It uses this optimistic controller to control the underlying system Θ for twice as long as the duration of the previous control policy. This technique known as "doubling trick" in reinforcement learning and online learning prevents frequent policy updates and balances the policy changes so that the overall regret of the algorithm is affected by a constant factor only.

4 Regret Analysis of LqgOpt

Now that the confidence set constructions and the adaptive control procedure of LqgOpt are explained, it only remains to analyze the regret of LqgOpt. Lemma 4.1 of Lale et al. [2020] shows that the random exploration in the warm-up period acquires linear regret, *i.e.* $\mathcal{O}(T_w)$.

In order to analyze the regret obtained during the adaptive control period, we first need to show that system will be well-controlled during the adaptive control period. The following lemma achieves that.

Lemma 4.1. *Suppose Assumptions 2.1-2.3 hold. After the warm-up period of T_w,* LQGOPT *satisfies the following with high probability for all $T \geq t \geq T_w$,*

1. $\Theta \in (\mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t))$
2. $\|\hat{x}_{t|t,\hat{\Theta}}\| \leq \tilde{\mathcal{X}}$
3. $\|y_t\| \leq \tilde{\mathcal{Y}}$

where $\tilde{\mathcal{X}}, \tilde{\mathcal{Y}} = \mathcal{O}(\sqrt{\log(T)})$. Here, $\mathcal{O}$ hides the problem dependent constants.

The proof of the lemma with the precise expressions is given in Appendix D. This lemma is critical for the regret analysis due to the nature of the adaptive control problem in partially observable environments. The inaccuracies in the system parameter estimates affect both the optimal feedback gain synthesis and the estimation of the underlying state. If these inaccuracies are not tolerable in the adaptive control of the system, they will accumulate fast and cause explosion and unboundedness in the input and the output of the system. This would result in linear, and potentially super linear regret. The main technical challenge in the proof is to show that with T_w length warm-up period, the error between the optimistic controller's state estimation $\hat{x}_{t|t,\hat{\Theta}}$ and the true state estimation $\hat{x}_{t|t,\Theta}$ does not blow up. Lemma 4.1 shows that while the system parameter estimates are refining, the input to the system and the system's output stays bounded during the adaptive control period.

Given the verification of stability in the adaptive control period, we bound the regret of adaptive control. The regret analysis is based on the Bellman optimality equation for LQG control problem provided in Lemma 4.3 of Lale et al. [2020]. The following theorem gives the regret upper bound of the adaptive control period of LQGOPT.

Theorem 4.1 (The regret of adaptive control)**.** *Suppose Assumptions 2.1-2.3 hold. After the warm-up period of T_w, with high probability, for any time T in adaptive control period, the regret of* LQGOPT *is bounded as follows:*

$$\text{REGRET}(T) = \tilde{\mathcal{O}}\left(\sqrt{T}\right). \tag{17}$$

where $\tilde{\mathcal{O}}(\cdot)$ hides the logarithmic factors and problem dependent constants.

The proof is given in the Appendices E and F. Here we provide the main proof ideas. Since we know that the optimistic controller can attain smaller average expected cost than the optimal controller of the given system, we decompose the regret using the Bellman optimality equation for the optimistic system. For each time step t, $(\hat{x}_{t|t-1}, y_t)$ is treated as the given state of the system and the differences between the system evolutions of the true system and optimistic system are analyzed in the regret decomposition. The regret decomposition is given in Appendix E. In Appendix F, we bound each term individually. The main pieces are the facts that the confidence sets shrink with $\tilde{\mathcal{O}}(1/\sqrt{t})$ (Theorem 3.4), LQGOPT avoids frequent policy changes and the control inputs and system outputs are well-controlled (Lemma 4.1). Combining Theorem 4.1 with $\mathcal{O}(T_w)$ regret from the warm-up period gives the overall regret upper bound of LQGOPT, stated in Theorem 3.1.

5 Related Works

The problem of sequential decision making under uncertainty is one the core studies in the field of control theory and reinforcement learning. Decision making in dynamical systems, when the environment is known and regulating costs are considered, results in a reduction to the study of optimal control. Optimal controls in the general setting of partially observable linear quadratic Gaussian systems, when highly crafted sensory observations of the system are available, and a fidelity approximation of the physics of dynamical systems is provided, has a long history of applications and successes. [Åström, 2012, Bertsekas, 1995, Hassibi et al., 1999].

When there is a high uncertainty in the modeling of the system, learning algorithms are required to learn the system behavior. In such situations, the learning agent estimates the system behaviour and adapt accordingly [Ljung, 1999, Kailath et al., 2000]. For the class of fully observable systems, Lai et al. [1982], Chen and Guo [1987] study this problem in asymptotic optimality sense, mainly developed on pure exploration approaches. Along with the regret analysis, the principle of pure exploration and betting on the best, or OFU has been studied for fully observable environments [Lai and Robbins, 1985, Campi and Kumar, 1998, Bittanti et al., 2006]. Recent works, deploy the OFU principle, and study tabular fully and partially observable Markov decision processes [Jaksch et al., 2010, Azizzadenesheli et al., 2016]. In Abbasi-Yadkori and Szepesvári [2011], the authors extend the OFU principle and employ recent advances in the estimation theory [Peña et al., 2009, Abbasi-Yadkori et al., 2011] and provide the first regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ for the fully observable case. In the setting of fully observable environments, an extensive advances and development have been proposed to provide generalized methods [Faradonbeh et al., 2017, Abeille and Lazaric, 2017, 2018, Ouyang et al., 2017, Dean et al., 2018]. Simultaneously, pure exploration methods along with uncertainty equivalence methods shed lights into the design of efficient algorithms [Abbasi-Yadkori et al., 2019, Mania et al., 2019, Faradonbeh et al., 2018, Cohen et al., 2019].

The system identification in partial observable linear systems in the presence of Gaussian noise, LQGs, has recently sparked a flurry of research interests [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]. Most of the proposed methods in prior works utilize open-loop system identification methods (without a history dependent controller), using independent Gaussian excitation, which makes it easy to show the persistence of excitation and deal with the biases in the estimation using Markov parameters. However, in Lee and Lamperski [2019], the authors use the innovations form of the state-space model to deal with the biases in closed-loop system identification whereas in Tsiamis and Pappas [2019], it is shown that process and measurement noises are sufficient for persistence of excitation in the absence of a control input. Another line of novel approaches is proposed 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 [Hazan et al., 2017, Arora et al., 2018, Hazan et al., 2018].

In this work, we propose the first learning algorithm to estimate the model parameters using any arbitrary bounded sequence of samples, even with feedback controls where the future events are correlated with historical data. Along with the estimation, we provide statistically tight high probability confidence intervals over the model parameters where the true model parameters live in. A recent work by Lale et al. [2020] provides a regret bound of $\tilde{\mathcal{O}}(T^{2/3})$ for such problem. The current

work, through deploying this novel estimation procedure improves the $\tilde{\mathcal{O}}(T^{2/3})$ bound to $\tilde{\mathcal{O}}(\sqrt{T})$. Another recent work by Simchowitz et al. [2020] study a general setting in partially observable with the presence of adversarial disturbances, and given access to stabilizing controller, provide a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$. These two mentioned works and the current paper, are amongst the first to provide sublinear regret bounds for partially observable linear systems.

6 Conclusion

In this work, we study the problem of adaptive control in partially observable linear systems, also known as linear systems with imperfect observation. While the prior work relies on open-loop system identification, we propose a novel method to estimate the system parameters even in the presence of feedback loop and correlation induced by feedback controllers. We deploy the principles of the Ho-Kalman method to estimate the model parameters and construct their corresponding confidence bound. We deploy the principle of optimism in the face of uncertainty and propose LqgOpt, a reinforcement algorithm for LQGs. LqgOpt sequentially interacts with the environment for a few time steps, collect samples, and exploit the samples to estimate the model parameters up to their confidence sets. LqgOpt computes the optimal controller associated with the most optimistic model in the set of plausible models, and then deploy this controller on the systems, but this time for a bit longer. LqgOpt repeats this process. We show that following LqgOpt results in a sublinear regret of $\tilde{\mathcal{O}}(\sqrt{T})$ which is the first $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound on LQG along with Simchowitz et al. [2020].

In future work, we plan to consider the setting where the cost function is strongly convex as in Simchowitz et al. [2020] and see if one can obtain $poly\log(T)$ regret in adaptive control of partially observable linear systems. We also aim to utilize the estimation method developed in this work and study the safety in adaptive control. Along with safety, we plan to extend this work to the problem of constraint control. While the Gaussian assumption on the noise has been long considered for partially observable linear dynamical systems, this assumption introduces limitation and model mismatch. Due to the generality of estimation analysis proposed methods in this work, in the future work, we aim to extend the current results to the case of sub-Gaussian with unknown but bounded parameters.

Acknowledgements

S. Lale is supported in part by DARPA PAI. K. Azizzadenesheli is supported in part by Raytheon and Amazon Web Service. 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, 2011.

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.

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

Karl J Åström. *Introduction to stochastic control theory*. Courier Corporation, 2012.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

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

Sergio Bittanti, Marco C Campi, et al. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

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

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, 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.

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.

BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

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.

Torben Knudsen. Consistency analysis of subspace identification methods based on a linear regression approach. *Automatica*, 37(1):81–89, 2001.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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*, 2020.

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.

Lingsheng Meng and Bing Zheng. The optimal perturbation bounds of the moore–penrose inverse under the frobenius norm. *Linear algebra and its applications*, 432(4):956–963, 2010.

Yi Ouyang, Mukul Gagrani, and Rahul Jain. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

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, 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.

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 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.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973.

Appendix

In the following, we first provide the definitions of truncated noise evolution parameters for both warm-up period and adaptive control period in Appendix A. Appendix A also contains lower bounds on the smallest singular value for $\|\Phi_t\Phi_t^\top\|$ for warm-up period and adaptive control period which are used in showing persistence of excitation and thus proving Theorem 3.3. In Appendix B, we show how the self-normalized bound is obtained for $\hat{\mathbf{M}}_\mathbf{t}$ and provide the proof of Theorem 3.2.

Appendix C gives the SysId algorithm and describes the construction of confidence sets using the outputs of SysId and provides the theoretical guarantees for them. In Appendix D, we give the proof of Lemma 4.1 and show that with the given warm-up period, the inputs and the outputs of the system stay bounded with high probability.

Appendix E provides regret decomposition for LqgOpt and states the differences arise from the policy updates in adaptive control period compared to explore and commit algorithm proposed in Lale et al. [2020]. In the Appendix F, we provide the proof of regret upper bound for the adaptive control period of LqgOpt. Finally, in Appendix G, we give the overview of the case when the initial state for the system is not coming from the steady state distribution.

Note that the warm-up period is chosen to be the following,

$$T_w \geq \max\{T_A, T_B, T_c, T_o, T_u, T_\mathbf{M}, T_N, T_\alpha, T_\beta, T_\gamma, T_\mathcal{G}\}$$

where each term satisfies different condition in order to obtain $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bound. The meanings of the terms are explained in detail throughout the Appendix.

A $H-$length Truncated Noise Evolution Parameters

In this section, we provide definitions of truncated open-loop and closed-loop noise evolution parameters, $\mathcal{G}^{ol}$ and $\mathcal{G}^{cl}$ respectively. They will play significant role in the confidence set for $\mathbf{M}$ in showing the persistence of excitation. They represent the effect of noises in the system on the outputs and the inputs. We will define $\mathcal{G}^{ol}$ and $\mathcal{G}^{cl}$ for $2H$ time steps back in time and show that last $2H$ process and measurement noises provide sufficient persistent excitation for the covariates in the estimation problem. In the following, $\bar{\phi}_t = P\phi_t$ for a permutation matrix P that gives

$$\bar{\phi}_t = \begin{bmatrix} y_{t-1}^\top & u_{t-1}^\top \dots y_{t-H}^\top & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{(m+p)H}.$$

A.1 Truncated Open-Loop Noise Evolution Parameter

Recall the state-space form of the system,

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + z_t. \end{aligned} \tag{18}$$

During the warm-up period, $t \leq T_w$, the input to the system is $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$. Let $f_t = [y_t^\top u_t^\top]^\top$. From the evolution of the system with given input we have the following:

$$f_t = \mathbf{G^o} \begin{bmatrix} w_{t-1}^\top & z_t^\top & u_t^\top & \dots & w_{t-H}^\top & z_{t-H+1}^\top & u_{t-H+1}^\top \end{bmatrix}^\top + \mathbf{r_t^o}$$

where

$$\mathbf{G^o} := \begin{bmatrix} 0_{m\times n} \; I_{m\times m} \; 0_{m\times p} & C \; 0_{m\times m} \; CB & CA \; 0_{m\times m} \; CAB & \dots & CA^{H-2} \; 0_{m\times m} \; CA^{H-2}B \\ 0_{p\times n} \; 0_{p\times m} \; I_{p\times p} & 0_{p\times n} \; 0_{p\times m} \; 0_{p\times p} & 0_{p\times n} \; 0_{p\times m} \; 0_{p\times p} & \dots & 0_{p\times n} \; 0_{p\times m} \; 0_{p\times p} \end{bmatrix} \tag{19}$$

and $\mathbf{r_t^o}$ is the residual vector that represents the effect of $[w_{i-1} \;\; z_i \;\; u_i]$ for $0 \le i < t-H$, which are independent. Notice that $\mathbf{G^o}$ is full row rank even for $H = 1$, due to first $(m+p) \times (m+n+p)$ block. Using this, we can represent $\bar{\phi}_t$ as follows

$$\bar{\phi}_t = \underbrace{\begin{bmatrix} f_{t-1} \\ \vdots \\ f_{t-H} \end{bmatrix}}_{\mathbb{R}^{(m+p)H}} + \begin{bmatrix} \mathbf{r^o_{t-1}} \\ \vdots \\ \mathbf{r^o_{t-H}} \end{bmatrix} = \mathcal{G}^{ol} \underbrace{\begin{bmatrix} w_{t-2} \\ z_{t-1} \\ u_{t-1} \\ \vdots \\ w_{t-2H-1} \\ z_{t-2H} \\ u_{t-2H} \end{bmatrix}}_{\mathbb{R}^{2(n+m+p)H}} + \begin{bmatrix} \mathbf{r^o_{t-1}} \\ \vdots \\ \mathbf{r^o_{t-H}} \end{bmatrix} \quad \text{where}$$

$$\mathcal{G}^{ol} := \begin{bmatrix} [\quad \mathbf{G^o} \quad] & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \cdots \\ 0_{(m+p)\times(m+n+p)} & [\quad \mathbf{G^o} \quad] & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \cdots & [\quad \mathbf{G^o} \quad] & 0_{(m+p)\times(m+n+p)} \\ 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & 0_{(m+p)\times(m+n+p)} & \cdots & [\quad \mathbf{G^o} \quad] \end{bmatrix}. \tag{20}$$

Define

$$T_o = \frac{32\Upsilon_w^4 \log^2\left(\frac{2H(m+p)}{\delta}\right)}{\sigma_{\min}^4(\mathcal{G}^{ol}) \min\{\sigma_w^4, \sigma_z^4, \sigma_u^4\}}.$$

We now prove Lemma 3.1, which shows that the inputs are persistently exciting uniformly during the warm-up period for $t \ge T_o$.

Lemma A.1 (Precise Statement of Lemma 3.1)**.** *If the warm-up duration $T_w \ge T_o$, then for $T_o \le t \le T_w$, with probability at least $1-\delta$ we have*

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

Proof. Let $\bar{\mathbf{0}} = 0_{(m+p)\times(m+n+p)}$. Since each block row is full row-rank, we get the following decomposition using QR decomposition for each block row:

$$\mathcal{G}^{ol} = \underbrace{\begin{bmatrix} Q^o & 0_{m+p} & 0_{m+p} & 0_{m+p} & \dots \\ 0_{m+p} & Q^o & 0_{m+p} & 0_{m+p} & \dots \\ & & \ddots & & \\ 0_{m+p} & 0_{m+p} & \dots & Q^o & 0_{m+p} \\ 0_{m+p} & 0_{m+p} & 0_{m+p} & \dots & Q^o \end{bmatrix}}_{\mathbb{R}^{(m+p)H\times(m+p)H}} \underbrace{\begin{bmatrix} R^o & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots \\ \bar{\mathbf{0}} & R^o & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots \\ & & \ddots & & \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots & R^o & \bar{\mathbf{0}} \\ \bar{\mathbf{0}} & \bar{\mathbf{0}} & \bar{\mathbf{0}} & \dots & R^o \end{bmatrix}}_{\mathbb{R}^{(m+p)H\times 2(m+n+p)H}}$$

where $R^o = \begin{bmatrix} \times & \times & \times & \times & \times & \times & \dots \\ 0 & \times & \times & \times & \times & \times & \dots \\ & \ddots & & & & & \\ 0 & 0 & 0 & \times & \times & \times & \dots \end{bmatrix} \in \mathbb{R}^{(m+p)\times H(m+n+p)}$ where the elements in the diagonal are positive numbers. Notice that the first matrix with Q^0 is full rank. Also, all the rows of second matrix are in row echelon form and second matrix is full row-rank. Thus, we can deduce that $\mathcal{G}^{ol}$ is full row-rank. Since $\mathcal{G}^{ol}$ is full row rank, we have that

$$\mathbb{E}[\bar{\phi}_t \bar{\phi}_t^\top] \succeq \mathcal{G}^{ol} \Sigma_{w,z,u} \mathcal{G}^{ol\top}$$

where $\Sigma_{w,z,u} \in \mathbb{R}^{2(n+m+p)H \times 2(n+m+p)H} = \text{diag}(\sigma_w^2, \sigma_z^2, \sigma_u^2, \dots, \sigma_w^2, \sigma_z^2, \sigma_u^2)$. This gives us

$$\sigma_{\min}(\mathbb{E}[\bar{\phi}_t \bar{\phi}_t^\top]) \geq \sigma_{\min}^2(\mathcal{G}^{ol}) \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}$$

for $t \leq T_w$. As given in (33)-(36), we have that $\|\phi_t\| \leq \Upsilon_w \sqrt{H}$ with probability at least $1 - \delta/2$. Given this holds, one can use Theorem H.1, to obtain the following which holds with probability $1 - \delta/2$:

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

Using Weyl's inequality, during the warm-up period with probability $1 - \delta$, we have

$$\sigma_{\min}\left(\sum_{i=1}^{t} \phi_i \phi_i^\top\right) \geq t\sigma_o^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\} - 2\sqrt{2t}\Upsilon_w^2 H \sqrt{\log\left(\frac{2H(m+p)}{\delta}\right)}.$$

For all $t \geq T_o := \frac{32\Upsilon_w^4 H^2 \log\left(\frac{2H(m+p)}{\delta}\right)}{\sigma_o^4 \min\{\sigma_w^4, \sigma_z^4, \sigma_u^4\}}$, we have the stated lower bound. □

A.2 Truncated Closed-Loop Noise Evolution Parameter

After the warm-up period, for $t \geq T_w$, the input to the system is $u_t = -\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}}$. Recall the following relation for state estimation updates using the optimistic parameters:

$$\begin{aligned}
\hat{x}_{t|t-1,\tilde{\Theta}} &= \tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
\hat{x}_{t|t,\tilde{\Theta}} &= \hat{x}_{t|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t\hat{x}_{t|t-1,\tilde{\Theta}}) \\
&= (\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(Cx_t + z_t - \tilde{C}_t(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}}) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(C(Ax_{t-1} - B\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} + w_{t-1}) + z_t).
\end{aligned} \tag{22}$$

Again, let $f_t = [y_t^\top u_t^\top]^\top$. Using (18) and (22), the following can be written for f_t

$$\begin{bmatrix} x_t \\ \hat{x}_{t|t,\tilde{\Theta}} \end{bmatrix} = \underbrace{\begin{bmatrix} A & -B\tilde{K}_{t-1} \\ \tilde{L}_t CA & (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1}) - \tilde{L}_t CB\tilde{K}_{t-1} \end{bmatrix}}_{\tilde{\mathbf{G}}_2^{(\mathbf{t})}} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\tilde{\Theta}} \end{bmatrix} + \underbrace{\begin{bmatrix} I & 0 \\ \tilde{L}_t C & \tilde{L}_t \end{bmatrix}}_{\tilde{\mathbf{G}}_3^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix}$$

$$f_t = \begin{bmatrix} CA & -CB\tilde{K}_{t-1} \\ -\tilde{K}_t\tilde{L}_tCA & -\tilde{K}_t(I-\tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1}-\tilde{B}_{t-1}\tilde{K}_{t-1})+\tilde{K}_t\tilde{L}_tCB\tilde{K}_{t-1} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\tilde{\Theta}} \end{bmatrix} + \begin{bmatrix} Cw_{t-1}+z_t \\ -\tilde{K}_t\tilde{L}_t(z_t+Cw_{t-1}) \end{bmatrix}$$

$$f_t = \underbrace{\begin{bmatrix} C & 0 \\ 0 & -\tilde{K}_t \end{bmatrix}}_{\tilde{\mathbf{\Gamma}}_\mathbf{t}} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\tilde{\Theta}} \end{bmatrix} + \underbrace{\begin{bmatrix} C & 0 \\ 0 & -\tilde{K}_t \end{bmatrix}}_{\tilde{\mathbf{\Gamma}}_\mathbf{t}} \underbrace{\begin{bmatrix} I & 0 \\ \tilde{L}_tC & \tilde{L}_t \end{bmatrix}}_{\mathbf{G}_\mathbf{3}^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix} + \begin{bmatrix} z_t \\ 0 \end{bmatrix}.$$

Rolling back in time for H time steps we get the following,

$$f_t = \tilde{\mathbf{\Gamma}}_\mathbf{t} \left(\sum_{i=t-H+1}^{t} \left(\prod_{j=i}^{t} \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{j})} \right) \tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{i-1})} \begin{bmatrix} w_{i-2} \\ z_{i-1} \end{bmatrix} \right) + \underbrace{\begin{bmatrix} C & I \\ -\tilde{K}_t\tilde{L}_tC & -\tilde{K}_t\tilde{L}_t \end{bmatrix}}_{\tilde{\mathbf{G}}_\mathbf{1}^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix} + \mathbf{r}_\mathbf{t}^\mathbf{c}$$

where $\mathbf{r}_\mathbf{t}^\mathbf{c}$ is the residual vector that represents the effect of $[w_{i-1} \;\; z_i]$ for $0 \leq i < t-H$, which are independent. Using this, we can represent $\bar{\phi}_t$ as follows

$$\bar{\phi}_t = \underbrace{\begin{bmatrix} f_{t-1} \\ \vdots \\ f_{t-H} \end{bmatrix}}_{\mathbb{R}^{(m+p)H}} + \begin{bmatrix} \mathbf{r}_\mathbf{t-1}^\mathbf{c} \\ \vdots \\ \mathbf{r}_\mathbf{t-H}^\mathbf{c} \end{bmatrix} = \mathcal{G}_t^{cl} \underbrace{\begin{bmatrix} w_{t-2} \\ z_{t-1} \\ \vdots \\ w_{t-2H-1} \\ z_{t-2H} \end{bmatrix}}_{\mathbb{R}^{2(n+m)H}} + \begin{bmatrix} \mathbf{r}_\mathbf{t-1}^\mathbf{c} \\ \vdots \\ \mathbf{r}_\mathbf{t-H}^\mathbf{c} \end{bmatrix}$$

where

$$\mathcal{G}_t^{cl} = \begin{bmatrix} [\quad \bar{\mathbf{G}}_\mathbf{t-1} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ 0_{(m+p)\times(m+n)} & [\quad \bar{\mathbf{G}}_\mathbf{t-2} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}}_\mathbf{t-H+1} \quad] & 0_{(m+p)\times(m+n)} \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}}_\mathbf{t-H} \quad] \end{bmatrix} \tag{23}$$

for

$$\bar{\mathbf{G}}_\mathbf{t} = \left[\tilde{\mathbf{G}}_\mathbf{1}^{(\mathbf{t})},\ \tilde{\mathbf{\Gamma}}_\mathbf{t}\tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})}\tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-1})},\ \tilde{\mathbf{\Gamma}}_\mathbf{t}\tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})}\tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t-1})}\tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-2})}, \ldots,\ \tilde{\mathbf{\Gamma}}_\mathbf{t}\tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t})}\tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t-1})} \cdots \tilde{\mathbf{G}}_\mathbf{2}^{(\mathbf{t-H+1})}\tilde{\mathbf{G}}_\mathbf{3}^{(\mathbf{t-H})} \right] \in \mathbb{R}^{(m+p)\times H(n+m)}$$

By knowing the underlying system, the agent can deploy the optimal control policy. $\mathcal{G}^{cl}$ represents the translation of the process and measurement noises into $\bar{\phi}_t$ while using the optimal policy:

$$\mathcal{G}^{cl} = \begin{bmatrix} [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ 0_{(m+p)\times(m+n)} & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots \\ & & \ddots & & \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}} \quad] & 0_{(m+p)\times(m+n)} \\ 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & 0_{(m+p)\times(m+n)} & \cdots & [\quad \bar{\mathbf{G}} \quad] \end{bmatrix} \tag{24}$$

where

$$\bar{\mathbf{G}} = \begin{bmatrix}\mathbf{G_1}, & \mathbf{\Gamma G_2 G_3}, & \mathbf{\Gamma G_2}^2\mathbf{G_3}, & \ldots, & \mathbf{\Gamma G_2}^{H-1}\mathbf{G_3}\end{bmatrix} \in \mathbb{R}^{(m+p)\times H(n+m)}$$

for

$$\mathbf{G_1} = \begin{bmatrix} C & I \\ -KLC & -KL \end{bmatrix}, \mathbf{\Gamma} = \begin{bmatrix} C & 0 \\ 0 & -K \end{bmatrix}, \mathbf{G_2} = \begin{bmatrix} A & -BK \\ LCA & (I-LC)(A-BK)-LCBK \end{bmatrix}, \mathbf{G_3} = \begin{bmatrix} I & 0 \\ LC & L \end{bmatrix}.$$

Note that length of H is chosen such that $\bar{\mathbf{G}}$ is full row rank. Similar to the case with truncated open-loop noise evolution parameter, having full row rank block rows provides a full row rank $\mathcal{G}^{cl}$ via the same QR decomposition argument. Thus, the assumption on the lower bound of the smallest singular value of the $H-$length truncated closed-loop noise evolution parameter, $\sigma_{\min}(\mathcal{G}^{cl}) > \sigma_c > 0$, is valid. Due to boundedness of the set $\mathcal{S}$ that LqgOpt is searching on, let $\|\tilde{\mathcal{G}^{cl}}\|_F \leq G$ for all model in $\mathcal{S}$. Define $G_r = G + \frac{\sigma_c\sqrt{H(m+p)}}{2}$ and

$$T_c = \frac{2048\Upsilon_c^4 H^2 \left(\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{32H\Upsilon_c\sqrt{2}\eta_T + 32H\eta_T^2 + 16\max\{\sigma_w^2,\sigma_z^2\}}{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}\right)\right)}{\sigma_c^4\min\{\sigma_w^4,\sigma_z^4\}}.$$

We now prove Lemma 3.2, which shows that the inputs are persistently exciting uniformly during the adaptive control period for $t \geq T_c$.

Lemma A.2 (Precise Statement of Lemma 3.2)**.** *After T_c time steps in adaptive control period, with probability $1-3\delta$, we have the following for all $t \geq T_c$,*

$$\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}. \tag{25}$$

Proof. Define $\tilde{\mathcal{G}^{cl}}$, which is the translation parameter for the process and measurement noises into $\bar{\phi}_t$ for the system that is governed by the **optimistically chosen parameter by** LqgOpt **while using the optimal optimistic controller**. Recall that we are searching for the optimistic system model which attains the optimal LQG cost over the set of $\mathcal{C}_t \cap \mathcal{S}$ and whose closed-loop noise evolution parameter satisfies the lower bound on the smallest singular value of the $H-$length truncated closed-loop noise evolution parameter, σ_c. Therefore, LqgOpt has the guarantee that $\sigma_{\min}(\tilde{\mathcal{G}^{cl}}) \geq \sigma_c$. Let

$$T_{\mathcal{G}} = T_B\left(\frac{2H + 2H\Gamma\zeta + 2H(H-1)\Gamma\zeta}{\sigma_c}\right)^2.$$

Picking $T_w \geq T_{\mathcal{G}}$, guarantees that in adaptive control period for all $t \geq T_w$, $\|\mathcal{G}_t^{cl} - \tilde{\mathcal{G}^{cl}}\| \leq \frac{\sigma_c}{2}$. Using Weyl's inequality on singular values, we have that $\sigma_{\min}(\mathcal{G}_t^{cl}) \geq \frac{\sigma_c}{2}$. Hence, for all $t \geq T_w$, we have that

$$\mathbb{E}[\bar{\phi}_t\bar{\phi}_t^\top] \succeq \mathcal{G}_t^{cl}\Sigma_{w,z}\mathcal{G}_t^{cl\top}$$

where $\Sigma_{w,z} \in \mathbb{R}^{2(n+m)H\times 2(n+m)H} = \mathrm{diag}(\sigma_w^2,\sigma_z^2,\ldots,\sigma_w^2,\sigma_z^2)$. This gives us $\sigma_{\min}(\mathbb{E}[\bar{\phi}_t\bar{\phi}_t^\top]) \geq \frac{\sigma_c^2}{4}\min\{\sigma_w^2,\sigma_z^2\}$ for $t \geq T_w$. As given in (37)-(39), we have that $\|\phi_t\| \leq \Upsilon_c\sqrt{H}$ with probability at least $1-2\delta$.

Given this holds, **for a given optimistic model**, one can use Theorem H.1 as in the truncated open-loop noise evolution parameter, to obtain the following which holds with probability $1-\delta$:

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

Notice that this bound holds only for a single model. However, we need to show that for any random model within the confidence set, the lower bound holds. Thus, we need a standard covering argument. Using the perturbation result that holds for all $t \geq T_w$, we have $\|\mathcal{G}_t^{cl}\|_F \leq G_r$. We have the following upper bound on the covering number:

$$\mathcal{N}(B(G_r), \|\cdot\|_F, \epsilon) \leq \left(G_r + \frac{2}{\epsilon}\right)^{(m+p)(n+m)H^2}.$$

Thus, the following holds for all the centers of ϵ-balls in $\|\mathcal{G}_t^{cl}\|_F$, for all $t \geq T_w$, with probability $1-\delta$:

$$\lambda_{\max}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right) \leq 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{2}{\epsilon}\right)}. \tag{27}$$

Let $\eta_T = \sigma_w\sqrt{2n\log\left(\frac{2nT}{\delta}\right)} + \sigma_z\sqrt{2m\log\left(\frac{2mT}{\delta}\right)}$. Considering all the systems in the ϵ-balls, during the adaptive control period with probability $1-3\delta$, we have

$$\begin{aligned}\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq\ & t\left(\frac{\sigma_c^2}{4}\min\{\sigma_w^2,\sigma_z^2\} - 2\epsilon\left(H\Upsilon_c\sqrt{2}\eta_T + H\eta_T^2 + \max\{\sigma_w^2/2, \sigma_z^2/2\}\right)\right)\\ & - 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{2}{\epsilon}\right)}.\end{aligned}$$

Let $\epsilon = \frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16\left(H\Upsilon_c\sqrt{2}\eta_T + H\eta_T^2 + \max\{\sigma_w^2/2,\sigma_z^2/2\}\right)}$. This gives the following bound

$$\begin{aligned}&\sigma_{\min}\left(\sum_{i=1}^{t}\phi_i\phi_i^\top\right) \geq t\left(\frac{\sigma_c^2}{8}\min\{\sigma_w^2,\sigma_z^2\}\right)\\ &- 2\sqrt{2t}\Upsilon_c^2 H\sqrt{\log\left(\frac{H(m+p)}{\delta}\right) + H^2(m+p)(m+n)\log\left(G_r + \frac{32H\Upsilon_c\sqrt{2}\eta_T + 32H\eta_T^2 + 16\max\{\sigma_w^2,\sigma_z^2\}}{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}\right)}.\end{aligned}$$

For all $t \geq T_c$, we have the stated lower bound. □

B System Identification

Recall that for a single input-output trajectory $\{y_t, u_t\}_{t=1}^T$, using the ARX model, we can write the following for the given system,

$$Y_t = \Phi_t \mathbf{M}^\top + \underbrace{E_t + N_t}_{\text{Noise}} \qquad \text{where} \tag{28}$$

$$\begin{aligned}
\mathbf{M} &= \big[CF,\ C\bar{A}F,\ \ldots,\ C\bar{A}^{H-1}F,\ CB,\ C\bar{A}B,\ \ldots,\ C\bar{A}^{H-1}B\big] \in \mathbb{R}^{m\times(m+p)H} \\
Y_t &= [y_H,\ y_{H+1},\ \ldots,\ y_t]^\top \in \mathbb{R}^{(t-H)\times m} \\
\Phi_t &= [\phi_H,\ \phi_{H+1},\ \ldots,\ \phi_t]^\top \in \mathbb{R}^{(t-H)\times(m+p)H} \\
E_t &= [e_H,\ e_{H+1},\ \ldots,\ e_t]^\top \in \mathbb{R}^{(t-H)\times m} \\
N_t &= \big[C\bar{A}^H x_0,\ C\bar{A}^H x_1, \ldots, C\bar{A}^H x_{t-H}\big]^\top \in \mathbb{R}^{(t-H)\times m}
\end{aligned}$$

$\hat{\mathbf{M}}_\mathbf{t}$ is the solution to $\min_X \|Y_t - \Phi_t X^\top\|_F^2 + \lambda\|X\|_F^2$. Hence, we get $\hat{\mathbf{M}}_\mathbf{t}^\top = (\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top Y_t$.

Proof of Theorem 3.2

$$\begin{aligned}
\hat{\mathbf{M}}_\mathbf{t} &= \big[(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top(\Phi_t\mathbf{M}^\top + E_t + N_t)\big]^\top \\
&= \big[(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top\,(E_t + N_t) + (\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top\Phi_t\mathbf{M}^\top \\
&\qquad + \lambda(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top - \lambda(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top\big]^\top \\
&= \big[(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t + (\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t + \mathbf{M}^\top - \lambda(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top\big]^\top
\end{aligned}$$

Using $\hat{\mathbf{M}}_\mathbf{t}$, we get

$$\begin{aligned}
&|\operatorname{Tr}(X(\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})^\top)| \qquad (29)\\
&= |\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t) + \operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t) - \lambda\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)| \\
&\le |\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t)| + |\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t)| + \lambda|\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)| \\
&\le \sqrt{\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\operatorname{Tr}(E_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t)} \qquad (30)\\
&\quad + \sqrt{\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\operatorname{Tr}(N_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t)} \\
&\quad + \lambda\sqrt{\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\operatorname{Tr}(\mathbf{M}(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)} \\
&= \sqrt{\operatorname{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)}\ \times \\
&\Big[\sqrt{\operatorname{Tr}(E_t^\top\Phi_t(\Phi_t^\top\Phi_t+\lambda I)^{-1}\Phi_t^\top E_t)} + \sqrt{\operatorname{Tr}(N_t^\top\Phi_t(\Phi_t^\top\Phi_t+\lambda I)^{-1}\Phi_t^\top N_t)} + \lambda\sqrt{\operatorname{Tr}(\mathbf{M}(\Phi_t^\top\Phi_t+\lambda I)^{-1}\mathbf{M}^\top)}\Big]
\end{aligned}$$

where (30) follows from $|\operatorname{Tr}(ABC^\top)| \le \sqrt{\operatorname{Tr}(ABA^\top)\operatorname{Tr}(CBC^\top)}$ for positive definite B due to Cauchy Schwarz (weighted inner-product). For $X = (\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})(\Phi_t^\top\Phi_t + \lambda I)$, we get

$$\sqrt{\operatorname{Tr}((\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})V_t(\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})^\top)} \le \sqrt{\operatorname{Tr}(E_t^\top\Phi_t V_t^{-1}\Phi_t^\top E_t)} + \sqrt{\operatorname{Tr}(N_t^\top\Phi_t V_t^{-1}\Phi_t^\top N_t)} + \sqrt{\lambda}\|\mathbf{M}\|_F \tag{31}$$

The first term on the right hand side of (31) can be bounded using Theorem H.2 since e_t is $\|C\Sigma C^\top + \sigma_z^2 I\|$-sub-Gaussian vector. Therefore,

$$\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} \leq \sqrt{m\|C\Sigma C^\top + \sigma_z^2 I\| \log\left(\frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}}\right)} \tag{32}$$

For the second term,

$$\begin{aligned}\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \leq \frac{1}{\sqrt{\lambda}}\|N_t^\top \Phi_t\|_F &\leq \sqrt{\frac{m}{\lambda}}\left\|\sum_{i=H}^{t} \phi_i (C\bar{A}^H x_{i-H})^\top\right\| \\ &\leq t\sqrt{\frac{m}{\lambda}} \max_{i\leq t}\left\|\phi_i (C\bar{A}^H x_{i-H})^\top\right\| \\ &\leq t\sqrt{\frac{m}{\lambda}}\|C\|\upsilon^H \max_{i\leq t}\|\phi_i\|\|x_{i-H}\|\end{aligned}$$

During warm-up period, from Lemma D.1 of Lale et al. [2020], we have that for all $1 \leq t \leq T_w$, with probability $1-\delta/2$,

$$\|x_t\| \leq X_w := \frac{(\sigma_w + \sigma_u\|B\|)\Phi(A)\rho(A)}{\sqrt{1-\rho(A)^2}}\sqrt{2n\log(12nT_w/\delta)}, \tag{33}$$

$$\|z_t\| \leq Z := \sigma_z\sqrt{2m\log(12mT_w/\delta)}, \tag{34}$$

$$\|u_t\| \leq U_w := \sigma_u\sqrt{2p\log(12pT_w/\delta)}, \tag{35}$$

$$\|y_t\| \leq \|C\|X_w + Z. \tag{36}$$

Thus, during the warm-up phase, we have $\max_{i\leq t\leq T_w}\|\phi_i\|\|x_{i-H}\| \leq \Upsilon_w X_w\sqrt{H}$, where $\Upsilon_w = \|C\|X_w + Z + U_w$. During the adaptive control phase, from Lemma 4.1, we have that for all $t \geq T_w$, with probability $1-2\delta$,

$$\|x_t\| \leq X_{ac} := \|\Sigma\|^{1/2}\sqrt{2n\log(2nT/\delta)} + \bar{\Delta} + \tilde{\mathcal{X}}, \tag{37}$$

$$\|u_t\| \leq \Gamma\tilde{\mathcal{X}}, \tag{38}$$

$$\|y_t\| \leq \tilde{\mathcal{Y}}. \tag{39}$$

Thus, after the warm-up phase, we have $\max_{T_w\leq t\leq T}\|\phi_i\|\|x_{i-H}\| \leq \Upsilon_c X_{ac}\sqrt{H}$, where $\Upsilon_c = \tilde{\mathcal{Y}} + \Gamma\tilde{\mathcal{X}}$. Therefore for all t,

$$\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \leq t\sqrt{\frac{mH}{\lambda}}\|C\|\upsilon^H \max\{\Upsilon_c X_{ac}, \Upsilon_w X_w\}$$

Picking $H = \frac{2\log(T)+\log(\max\{\Upsilon_c X_{ac}, \Upsilon_w X_w\})+0.5\log(m/\lambda)+\log(\|C\|)}{\log(1/\upsilon)}$ gives

$$\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \leq \frac{t}{T^2}\sqrt{H} \tag{40}$$

Combining (32) and (40) gives the statement of Theorem 3.2. □

Proof of Theorem 3.3: For $\|\mathbf{M}\|_F \leq S$, we have

$$\sigma_{\min}(V_t)\|\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M}\|_F^2 \leq \mathrm{Tr}((\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M})V_t(\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M})^\top)$$
$$\leq \left(\sqrt{m\|C\Sigma C^\top+\sigma_z^2 I\|\log\left(\frac{\det(V_t)^{1/2}}{\delta\det(V)^{1/2}}\right)}+S\sqrt{\lambda}+\frac{t\sqrt{H}}{T^2}\right)^2$$

During the warm-up period, for $t \geq T_o$, using Lemma A.1, we get

$$\|\hat{\mathbf{M}}_{\mathbf{0}}-\mathbf{M}\|_F \leq \frac{\sqrt{m\|C\Sigma C^\top+\sigma_z^2 I\|\left(\log(\frac{1}{\delta})+\frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)H+t\Upsilon_w^2}{\lambda(m+p)H}\right)\right)}+S\sqrt{\lambda}+\frac{t\sqrt{H}}{T^2}}{\sqrt{t\frac{\sigma_o^2\min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2}}} \leq \frac{R_{\text{warm}}}{\sqrt{t}}.$$

where $R_{\text{warm}} = \frac{\sqrt{2m\|C\Sigma C^\top+\sigma_z^2 I\|\left(\log(1/\delta)+\frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)H+T_w\Upsilon_w^2}{\lambda(m+p)H}\right)\right)}+S\sqrt{2\lambda}+\frac{\sqrt{2H}}{T}}{\sigma_o\min\{\sigma_w,\sigma_z,\sigma_u\}}$. Let $T_{\mathbf{M}} = R_{\text{warm}}^2$. For $T_w \geq T_{\mathbf{M}}$, we will have $\|\hat{\mathbf{M}}_{\mathbf{0}}-\mathbf{M}\|_F \leq 1$.

During the adaptive control period, for $t \geq T_c + T_w$, using Lemma A.2, we get

$$\|\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M}\|_F \leq \frac{\sqrt{m\|C\Sigma C^\top+\sigma_z^2 I\|\left(\log(1/\delta)+\frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)H+t\max\{\Upsilon_w^2,\Upsilon_c^2\}}{\lambda(m+p)H}\right)\right)}+S\sqrt{\lambda}+\frac{t\sqrt{H}}{T^2}}{\sqrt{T_w\frac{\sigma_o^2\min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2}+(t-T_w)\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}}}$$
$$\leq \frac{\sqrt{m\|C\Sigma C^\top+\sigma_z^2 I\|\left(\log(1/\delta)+\frac{H(m+p)}{2}\log\left(\frac{\lambda(m+p)H+T\max\{\Upsilon_w^2,\Upsilon_c^2\}}{\lambda(m+p)H}\right)\right)}+S\sqrt{\lambda}+\frac{\sqrt{H}}{T}}{\sqrt{t}\sqrt{\min\left\{\frac{\sigma_o^2\min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2},\frac{\sigma_c^2\min\{\sigma_w^2,\sigma_z^2\}}{16}\right\}}}$$

□

C Confidence Set Construction for System Parameters

After estimating $\hat{\mathbf{M}}_{\mathbf{t}}$, we construct confidence sets for the unknown system parameters and use these confidence sets to come up with the optimistic controller to exploit the information gathered. LqgOpt uses SysId, a method similar to Ho-Kalman method [Ho and Kálmán, 1966], to estimate the system parameters from $\hat{\mathbf{M}}_{\mathbf{t}}$. The outline of the algorithm is given in the main text and in Algorithm 2. Note that the system is order n and minimal in the sense that the system cannot be described by a state-space model of order less than n. Thus, without loss of generality, $\sigma_n(A) > 0$. The results in this section follow similar steps with Oymak and Ozay [2018] with similar changes mentioned in Lale et al. [2020]. The following lemma is from Oymak and Ozay [2018], it will be used in proving confidence bounds and we provide it for completeness.

Lemma C.1 ([Oymak and Ozay, 2018])**.** *$\mathcal{H}$, $\hat{\mathcal{H}}_t$ and $\mathcal{N}$,$\hat{\mathcal{N}}_t$ satisfies the following perturbation bounds,*

$$\max\left\{\left\|\mathcal{H}^+-\hat{\mathcal{H}}_t^+\right\|,\left\|\mathcal{H}^--\hat{\mathcal{H}}_t^-\right\|\right\} \leq \|\mathcal{H}-\hat{\mathcal{H}}_t\| \leq \sqrt{\min\{d_1,d_2+1\}}\|\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M}\|$$
$$\|\mathcal{N}-\hat{\mathcal{N}}_t\| \leq 2\left\|\mathcal{H}^--\hat{\mathcal{H}}_t^-\right\| \leq 2\sqrt{\min\{d_1,d_2\}}\|\hat{\mathbf{M}}_{\mathbf{t}}-\mathbf{M}\|$$

Algorithm 2 SysId

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

The following lemma is a slight modification of Lemma B.1 in [Oymak and Ozay, 2018].

Lemma C.2 ([Oymak and Ozay, 2018])**.** *Suppose $\sigma_{\min}(\mathcal{N}) \geq 2\|\mathcal{N} - \hat{\mathcal{N}}\|$ where $\sigma_{\min}(\mathcal{N})$ is the smallest nonzero singular value (i.e. nth largest singular value) of N. Let rank n matrices $\mathcal{N}, \hat{\mathcal{N}}$ have singular value decompositions $\mathbf{U\Sigma V}^\top$ and $\hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$ There exists an $n \times n$ unitary matrix $\mathbf{T}$ so that*

$$\left\|\mathbf{U\Sigma}^{1/2} - \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}^{1/2}\mathbf{T}\right\|_F^2 + \left\|\mathbf{V\Sigma}^{1/2} - \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{1/2}\mathbf{T}\right\|_F^2 \leq \frac{5n\|\mathcal{N} - \hat{\mathcal{N}}\|^2}{\sigma_n(\mathcal{N}) - \|\mathcal{N} - \hat{\mathcal{N}}\|}$$

The following is the proof of Theorem 3.4.

Proof of Theorem 3.4: For brevity, we have the following notation $\mathbf{O} = \mathbf{O}(\bar{A}, C, d_1)$, $\mathbf{C_F} = \mathbf{C}(\bar{A}, F, d_2+1)$, $\mathbf{C_B} = \mathbf{C}(\bar{A}, B, d_2+1)$, $\hat{\mathbf{O}}_{\mathbf{t}} = \hat{\mathbf{O}}_{\mathbf{t}}(\bar{A}, C, d_1)$, $\hat{\mathbf{C}}_{\mathbf{F_t}} = \hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, F, d_2+1)$, $\hat{\mathbf{C}}_{\mathbf{B_t}} = \hat{\mathbf{C}}_{\mathbf{t}}(\bar{A}, B, d_2+1)$. Let $T_N = T_{\mathbf{M}}\frac{8H}{\sigma_n^2(\mathcal{H})}$. Directly applying Lemma C.2 with the condition that for given $T_w \geq T_N$, $\sigma_{\min}(\mathcal{N}) \geq 2\|\mathcal{N} - \hat{\mathcal{N}}\|$, we can guarantee that there exists a unitary transform $\mathbf{T}$ such that

$$\left\|\hat{\mathbf{O}}_{\mathbf{t}} - \mathbf{OT}\right\|_F^2 + \left\|[\hat{\mathbf{C}}_{\mathbf{F_t}}\ \ \hat{\mathbf{C}}_{\mathbf{B_t}}] - \mathbf{T}^\top[\mathbf{C_F}\ \ \mathbf{C_B}]\right\|_F^2 \leq \frac{10n\|\mathcal{N} - \hat{\mathcal{N}}_t\|^2}{\sigma_n(\mathcal{N})} \tag{41}$$

Since $\hat{C}_t - \bar{C}\mathbf{T}$ is a submatrix of $\hat{\mathbf{O}}_{\mathbf{t}} - \mathbf{OT}$, $\hat{B}_t - \mathbf{T}^\top\bar{B}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{B_t}} - \mathbf{T}^\top\mathbf{C_B}$ and $\hat{F}_t - \mathbf{T}^\top\bar{F}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{F_t}} - \mathbf{T}^\top\mathbf{C_F}$, we get the same bounds for them stated in (41). Using Lemma C.1, with the choice of $d_1, d_2 \geq \frac{H}{2}$, we have

$$\|\mathcal{N} - \hat{\mathcal{N}}_t\| \leq \sqrt{2H}\|\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}\|.$$

This provides the advertised bounds in the theorem:

$$\|\hat{B}_t - \mathbf{T}^\top\bar{B}\|, \|\hat{C}_t - \bar{C}\mathbf{T}\|, \|\hat{F}_t - \mathbf{T}^\top\bar{F}\| \leq \frac{\sqrt{20nH}\|\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}\|}{\sqrt{\sigma_n(\mathcal{N})}}$$

Let $T_B = T_{\mathbf{M}} \frac{20nH}{\sigma_n(\mathcal{H})}$. Notice that for $T_w \geq T_B$, we have all the terms above to be bounded by 1. In order to determine the closeness of $\hat{A}_t$ and $\bar{A}$ we first consider the closeness of $\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}$, where $\bar{\bar{A}}$ is the output obtained by Ho-Kalman for $\bar{A}$ when the input is $\mathbf{M}$. Let $X = \mathbf{OT}$ and $Y = \mathbf{T}^\top [\mathbf{C_F} \;\; \mathbf{C_B}]$. Thus, we have

$$\begin{aligned}
\|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}\|_F &= \|\hat{\mathbf{O}}_\mathbf{t}^\dagger \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - X^\dagger \mathcal{H}^+ Y^\dagger\|_F \\
&\leq \left\| \left(\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger \right) \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F + \left\| X^\dagger \left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+ \right) [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F \\
&\quad + \left\| X^\dagger \mathcal{H}^+ \left([\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right) \right\|_F
\end{aligned}$$

For the first term we have the following perturbation bound [Meng and Zheng, 2010, Wedin, 1973],

$$\|\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger\|_F \leq \|\hat{\mathbf{O}}_\mathbf{t} - X\|_F \max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_\mathbf{t}^\dagger\|^2\} \leq \|\mathcal{N} - \hat{\mathcal{N}}_t\| \sqrt{\frac{10n}{\sigma_n(\mathcal{N})}} \max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_\mathbf{t}^\dagger\|^2\}$$

Since we have $\sigma_n(\mathcal{N}) \geq 2\|\mathcal{N} - \hat{\mathcal{N}}\|$, we have $\|\hat{\mathcal{N}}\| \leq 2\|\mathcal{N}\|$ and $2\sigma_n(\hat{\mathcal{N}}) \geq \sigma_n(\mathcal{N})$. Thus,

$$\max\{\|X^\dagger\|^2, \|\hat{\mathbf{O}}_\mathbf{t}^\dagger\|^2\} = \max\left\{ \frac{1}{\sigma_n(\mathcal{N})}, \;\; \frac{1}{\sigma_n(\hat{\mathcal{N}})} \right\} \leq \frac{2}{\sigma_n(\mathcal{N})} \tag{42}$$

Combining these and following the same steps for $\|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger\|_F$, we get

$$\left\| \hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger \right\|_F, \;\; \left\| [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right\|_F \leq \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\| \sqrt{\frac{40n}{\sigma_n^3(\mathcal{N})}} \tag{43}$$

The following individual bounds obtained by using (42), (43) and triangle inequality:

$$\begin{aligned}
\left\| \left(\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger \right) \hat{\mathcal{H}}_t^+ [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F &\leq \|\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger\|_F \|\hat{\mathcal{H}}_t^+\| \|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger\| \\
&\leq \frac{4\sqrt{5n} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{\sigma_n^2(\mathcal{N})} \left(\|\mathcal{H}^+\| + \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\| \right) \\
\left\| X^\dagger \left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+ \right) [\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger \right\|_F &\leq \frac{2\sqrt{n} \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\|}{\sigma_n(\mathcal{N})} \\
\left\| X^\dagger \mathcal{H}^+ \left([\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger \right) \right\|_F &\leq \|X^\dagger\| \|\mathcal{H}^+\| \|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^\dagger - Y^\dagger\| \\
&\leq \frac{2\sqrt{10n} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{\sigma_n^2(\mathcal{N})} \|\mathcal{H}^+\|
\end{aligned}$$

Combining these we get

$$\begin{aligned}
\|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}} \mathbf{T}\|_F &\leq \frac{31\sqrt{n} \|\mathcal{H}^+\| \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{2\sigma_n^2(\mathcal{N})} + \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\| \left(\frac{4\sqrt{5n} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\|}{\sigma_n^2(\mathcal{N})} + \frac{2\sqrt{n}}{\sigma_n(\mathcal{N})} \right) \\
&\leq \frac{31\sqrt{n} \|\mathcal{H}^+\|}{2\sigma_n^2(\mathcal{N})} \left\| \mathcal{N} - \hat{\mathcal{N}}_t \right\| + \frac{13\sqrt{n}}{2\sigma_n(\mathcal{N})} \|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\|
\end{aligned}$$

Now consider $\hat{A}_t = \hat{\bar{A}}_t + \hat{F}_t\hat{C}_t$. Using Lemma C.1,

$$
\begin{aligned}
&\|\hat{A}_t - \mathbf{T}^\top \bar{A}\mathbf{T}\|_F \\
&= \|\hat{\bar{A}}_t + \hat{F}_t\hat{C}_t - \mathbf{T}^\top \bar{\bar{A}}\mathbf{T} - \mathbf{T}^\top \bar{F}\bar{C}\mathbf{T}\|_F \\
&\leq \|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}}\mathbf{T}\|_F + \|(\hat{F}_t - \mathbf{T}^\top \bar{F})\hat{C}_t\|_F + \|\mathbf{T}^\top \bar{F}(\hat{C}_t - \bar{C}\mathbf{T})\|_F \\
&\leq \|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\bar{A}}\mathbf{T}\|_F + \|(\hat{F}_t - \mathbf{T}^\top \bar{F})\|_F\|\hat{C}_t - \bar{C}\mathbf{T}\|_F + \|(\hat{F}_t - \mathbf{T}^\top \bar{F})\|_F\|\bar{C}\| + \|\bar{F}\|\|(\hat{C}_t - \bar{C}\mathbf{T})\|_F \\
&\leq \frac{31\sqrt{n}\|\mathcal{H}^+\|}{2\sigma_n^2(\mathcal{N})}\left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| + \frac{13\sqrt{n}}{2\sigma_n(\mathcal{N})}\|\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\| + \frac{10n\|\mathcal{N} - \hat{\mathcal{N}}_t\|^2}{\sigma_n(\mathcal{N})} + (\|\bar{F}\| + \|\bar{C}\|)\|\mathcal{N} - \hat{\mathcal{N}}_t\|\sqrt{\frac{10n}{\sigma_n(\mathcal{N})}} \\
&\leq \frac{31\sqrt{2nH}\|\mathcal{H}\|}{2\sigma_n^2(\mathcal{N})}\|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\| + \frac{13\sqrt{nH}}{2\sqrt{2}\sigma_n(\mathcal{N})}\|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\| + \frac{20nH\|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|^2}{\sigma_n(\mathcal{N})} \\
&\quad + (\|\bar{F}\| + \|\bar{C}\|)\|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\|\sqrt{\frac{20nH}{\sigma_n(\mathcal{N})}}
\end{aligned}
$$

Define T_A such that

$$
T_A = T_\mathbf{M}\left(\frac{\frac{62\sqrt{2nH}\|\mathcal{H}\|}{2\sigma_n^2(\mathcal{N})} + \frac{26\sqrt{nH}}{2\sqrt{2}\sigma_n(\mathcal{N})} + (\|\bar{F}\| + \|\bar{C}\|)\sqrt{\frac{80nH}{\sigma_n(\mathcal{N})}} + \sqrt{\frac{40nH\sigma_n(\bar{A})}{\sigma_n(\mathcal{N})}}}{\sigma_n(\bar{A})}\right)^2. \tag{44}
$$

Notice that for $T_w \geq T_A$, we have $\|\hat{A}_t - \mathbf{T}^\top \bar{A}\mathbf{T}\| \leq \sigma_n(\bar{A})/2$. Since $T_w \geq T_A$, from Weyl's inequality we have $\sigma_n(\hat{A}_t) \geq \sigma_n(\bar{A})/2$. Recalling that $X = \mathbf{O}(\bar{A}, C, d_1)\mathbf{T}$, under Assumption 2.2 we consider $\hat{L}_t$:

$$
\begin{aligned}
&\|\hat{L}_t - \mathbf{T}^\top \bar{L}\|_F \\
&= \|\hat{A}_t^\dagger \hat{\mathbf{O}}_\mathbf{t}^\dagger \hat{\mathcal{H}}_t^- - \mathbf{T}^\top \bar{A}^\dagger \mathbf{O}^\dagger \mathcal{H}^-\|_F \\
&\leq \|(\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T})\hat{\mathbf{O}}_\mathbf{t}^\dagger \hat{\mathcal{H}}_t^-\|_F + \|\mathbf{T}^\top \bar{A}^\dagger \mathbf{T}(\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger)\hat{\mathcal{H}}_t^-\|_F + \|\mathbf{T}^\top \bar{A}^\dagger \mathbf{T}X^\dagger(\hat{\mathcal{H}}_t^- - \mathcal{H}^-)\|_F \\
&\leq \|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F\|\hat{\mathbf{O}}_\mathbf{t}^\dagger\|\|\hat{\mathcal{H}}_t^-\| + \|\hat{\mathbf{O}}_\mathbf{t}^\dagger - X^\dagger\|_F\|\bar{A}^\dagger\|\|\hat{\mathcal{H}}_t^-\| + \sqrt{n}\|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\|\|\bar{A}^\dagger\|\|X^\dagger\| \\
&\leq \left(\|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F\sqrt{\frac{2}{\sigma_n(\mathcal{N})}} + \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\|\sqrt{\frac{40n}{\sigma_n^3(\mathcal{N})}}\|\bar{A}^\dagger\|\right)\left(\|\mathcal{H}^-\| + \|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\|\right) \\
&\quad + \sqrt{n}\|\bar{A}^\dagger\|\frac{1}{\sqrt{\sigma_n(\mathcal{N})}}\|\hat{\mathcal{H}}_t^- - \mathcal{H}^-\|
\end{aligned}
$$

Again using the perturbation bounds of the Moore–Penrose inverse under the Frobenius norm [Meng and Zheng, 2010], we have $\|\hat{A}_t^\dagger - \mathbf{T}^\top \bar{A}^\dagger \mathbf{T}\|_F \leq \frac{2}{\sigma_n^2(\bar{A})}\|\hat{A}_t - \mathbf{T}^\top \bar{A}\mathbf{T}\|$. Notice that the similarity transformation that transfers A to $\bar{A}$ is bounded since $S = \left([C^\top\ (C\bar{A})^\top \ldots (C\bar{A}^{d_1-1})^\top]^\top\right)^\dagger \mathbf{O}(\bar{A}, C, d_1)$. Combining all and using Lemma C.1, we obtain the confidence set for $\hat{L}_t$ given in Theorem 3.4.

□

D Boundedness of The Output and State Estimation, Proof of Lemma 4.1

The proof of Lemma 4.1 follows similar arguments with the proof of Lemma 4.2 of [Lale et al., 2020]. The main difference is that LqgOpt, the system estimations are refined during the adaptive control period, thus the control policy is refined. Also, since the behavior of a system and its similarity transformation is the same, without loss of generality we assume that similarity transformation $\mathbf{T} = I$.

Proof of Lemma 4.1:

Assume that $\Theta \in (\mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t))$ for all $t \geq T_w$, which is holds with probability $1-\delta$. We can write the decomposition for $\hat{x}_{t|t,\tilde{\Theta}}$ as follows,

$$\begin{aligned}
\hat{x}_{t|t,\tilde{\Theta}} &= \hat{x}_{t|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t\hat{x}_{t|t-1,\tilde{\Theta}}) \\
&= \tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t(y_t - \tilde{C}_t(\tilde{A}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}} - \tilde{B}_{t-1}\tilde{K}_{t-1}\hat{x}_{t-1|t-1,\tilde{\Theta}})) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + \tilde{L}_t y_t \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t\left(Cx_t - C\hat{x}_{t|t-1,\tilde{\Theta}} + C\hat{x}_{t|t-1,\tilde{\Theta}} + z_t\right) \\
&= (I - \tilde{L}_t\tilde{C}_t)(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t\left(Cx_t - C\hat{x}_{t|t-1,\tilde{\Theta}} + C(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + z_t\right) \\
&= \left(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)\right)\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta} + \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t \\
&= \left(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)\right)\hat{x}_{t-1|t-1,\tilde{\Theta}} \\
&\qquad + \tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t. \qquad (45)
\end{aligned}$$

Thus, the dynamics of $\hat{x}_{t|t,\tilde{\Theta}}$ is governed by

$$\mathbf{N}_t = \tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1} - \tilde{L}_t\left(\tilde{C}_t\tilde{A}_{t-1} - \tilde{C}_t\tilde{B}_{t-1}\tilde{K}_{t-1} - C\tilde{A}_{t-1} + C\tilde{B}_{t-1}\tilde{K}_{t-1}\right)$$

and it is driven by the process of $\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + \tilde{L}_t z_t$. Let $T_u = T_B\left(\frac{2\zeta\rho}{1-\rho}\right)^2$. With the Assumption 2.3, and for $T_w \geq T_u$, we have that $\|\tilde{C}_t - C\| \leq \frac{1-\rho}{2\zeta\rho}$ which gives $\|\mathbf{N}_t\| \leq \frac{1+\rho}{2} < 1$ for all $t \geq T_w$. Similar to the proof of Lemma 4.2 in [Lale et al., 2020], we have that $\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t z_t$ is $\zeta(\|C\|\|\Sigma\|^{1/2} + \sigma_z)$-sub-Gaussian, thus it's ℓ_2-norm can be bounded using Lemma H.1:

$$\|\tilde{L}_t C(x_t - \hat{x}_{t|t-1,\Theta}) + \tilde{L}_t z_t\| \leq \zeta(\|C\|\|\Sigma\|^{1/2} + \sigma_z)\sqrt{2n\log(2nT/\delta)}$$

for all $t \geq T_w$ with probability at least $1-\delta$. A special care is needed for $\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}$. Denote $\Delta_t = \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}$. Consider the decomposition given in equation (51) in Lale et al. [2020]. In this setting, since at each time step after the warm-up, the estimation errors are monotonically

decreasing, therefore we can upper bound the norm of each term in the decomposition by the norm of the term at the time of end of warm-up. Let

$$T_\alpha = T_B \left(\frac{\Gamma\left(1+\zeta(1+\|C\|)\right)}{\sigma - \upsilon} \right)^2, \qquad T_\gamma = T_A \frac{\sigma_n^2(\bar{A})}{4} \left(\frac{1+\Gamma(1+\zeta\|B\|)}{\sigma-\rho} \right)^2,$$
$$T_\beta = T_A \frac{\sigma_n^2(\bar{A})}{4} \left(\frac{\Gamma\|B\|(1+\zeta+\zeta\|C\|)(\Phi(A)\zeta+(1+\Gamma)(1+\zeta))}{(1-\sigma)^2} \right)^2. \tag{46}$$

Thus, using the arguments in Lale et al. [2020], we can show that after a warm-up period of $T_w \geq \max\{T_\alpha, T_\gamma\}$, we have that for all $t \geq T_w$, $\max\{\|(A+(\tilde{A}_t - A - \tilde{B}_t\tilde{K}_t + B\tilde{K}_t))(I-\tilde{L}_t\tilde{C}_t)\|, \|A - B\tilde{K}_t + B\tilde{K}_t\tilde{L}_t(\tilde{C}_t - C)\|\} \leq \sigma < 1$. Using the inductive argument given in [Lale et al., 2020], we can show that for all $t \geq T_w \geq T_\beta$, $\|\Delta_t\| \leq \bar{\Delta}$ with probability $1-\delta$. Notice that the definition of $\bar{\Delta}$ still includes the same terms given in equation (54) of Lale et al. [2020] but $\beta_A, \beta_B, \beta_C$ is replaced with $\beta_A(T_w), \beta_B(T_w), \beta_C(T_w)$ and ΔL is replaced by $2\beta_L(T_w)$ due to new estimation method, *i.e.*,

$$\bar{\Delta} = 10 \left(\frac{\bar{\kappa}}{1-\sigma} + \frac{\bar{\beta}\bar{\xi}}{(1-\sigma)^2} \right) \left(\|C\|\|\Sigma\|^{1/2} + \sigma_z \right) \sqrt{2m\log(2mT/\delta)}$$

for $\bar{\kappa} = 2\Phi(A)\beta_L(T_w) + 2\zeta(\beta_A(T_w) + \Gamma\beta_B(T_w))$, $\bar{\beta} = 2\zeta\beta_C(T_w)(\Phi(A) + 2(\beta_A(T_w) + \Gamma\beta_B(T_w))) + 2(\beta_A(T_w) + \Gamma\beta_B(T_w))$ and $\bar{\xi} = \zeta(\rho + 2(\beta_A(T_w) + \Gamma\beta_B(T_w))) + 2\|B\|\Gamma\beta_L(T_w)$. Thus, we get

$$\|\hat{x}_{t|t,\tilde{\Theta}}\| = \left\| \sum_{i=1}^{t} \mathbf{N}^{t-i} \left(\tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i \right) \right\| \tag{47}$$
$$\leq \max_{1\leq i\leq t} \left\| \tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i \right\| \left(\sum_{i=1}^{t} \|\mathbf{M}\|^{t-i} \right) \tag{48}$$
$$\leq \frac{2}{1-\rho} \max_{1\leq i\leq t} \left\| \tilde{L}_i C(x_{i-1} - \hat{x}_{i|i-1,\Theta}) + \tilde{L}_i C(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\tilde{\Theta}}) + \tilde{L}_i z_i \right\| \tag{49}$$
$$\leq \tilde{\mathcal{X}} := \frac{2\zeta \left(\|C\|\bar{\Delta} + \left(\|C\|\|\Sigma\|^{1/2} + \sigma_z \right) \sqrt{2n\log(2nT/\delta)} \right)}{1-\rho}. \tag{50}$$

with probability $1-2\delta$. For y_t, we have the following decomposition,

$$\begin{aligned} y_t &= C\hat{x}_{t|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \\ &= C\hat{x}_{t|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\Theta}) + C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \\ &= C(\tilde{A}_{t-1} - \tilde{B}_{t-1}\tilde{K}_{t-1})\hat{x}_{t-1|t-1,\tilde{\Theta}} + C(x_t - \hat{x}_{t|t-1,\Theta}) + C(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\tilde{\Theta}}) + z_t \end{aligned}$$

Using similar analysis with $\hat{x}_{t|t,\tilde{\Theta}}$, we get the following bound for y_t for all $t \geq T_w$:

$$\|y_t\| \leq \rho\|C\|\tilde{\mathcal{X}} + (\|C\|\|\Sigma\|^{1/2} + \sigma_z)\sqrt{2m\log(2mT/\delta)} + \|C\|\bar{\Delta}$$

with probability $1-2\delta$. Thus, all three statements of Lemma 4.1 hold with probability at least $1-3\delta$. □

E Regret Decomposition

Recall the following lemma from [Lale et al., 2020] on the Bellman optimality equation for LQG:

Lemma E.1 (Bellman Optimality Equation for LQG [Lale et al., 2020])**.** *Given state estimation $\hat{x}_{t|t-1} \in \mathbb{R}^n$ and an observation $y_t \in \mathbb{R}^m$ pair at time t, Bellman optimality equation of average cost per stage control of* LQG *system $\Theta = (A, B, C)$ with regulating parameters Q and R is*

$$J_*(\Theta) + \hat{x}_{t|t}^\top \left(P - C^\top Q C\right) \hat{x}_{t|t} + y_t^\top Q y_t = \min_u \Big\{ y_t^\top Q y_t + u^\top R u + \mathbb{E}\Big[\hat{x}_{t+1|t+1}^{u\top} \left(P - C^\top Q C\right) \hat{x}_{t+1|t+1}^u + y_{t+1}^{u\top} Q y_{t+1}^u\Big]\Big\} \tag{51}$$

where P is the unique solution to DARE of Θ, $\hat{x}_{t|t} = (I - LC)\hat{x}_{t|t-1} + Ly_t$, $y_{t+1}^u = C(Ax_t + Bu + w_t) + z_{t+1}$, and $\hat{x}_{t+1|t+1}^u = (I - LC)\left(A\hat{x}_{t|t} + Bu\right) + Ly_{t+1}^u$. The equality is achieved by the optimal controller of Θ.

Using Lemma E.1 for the optimistic system at time t, we derive the instantaneous regret decomposition at time t with the following expressions:

$$\hat{x}_{t|t,\tilde{\Theta}_t} = \left(I - \tilde{L}_t \tilde{C}_t\right) \hat{x}_{t|t-1} + \tilde{L}_t y_t \tag{52}$$

$$y_{t+1,\tilde{\Theta}_t} = \tilde{C}_t \left(\tilde{A}_t - \tilde{B}_t \tilde{K}_t\right) \hat{x}_{t|t,\tilde{\Theta}_t} + \tilde{C}_t \tilde{A}_t \left(x_t - \hat{x}_{t|t,\tilde{\Theta}_t}\right) + \tilde{C}_t w_t + z_{t+1} \tag{53}$$

$$\hat{x}_{t+1|t+1,\tilde{\Theta}_t} = \left(\tilde{A}_t - \tilde{B}_t \tilde{K}_t\right) \hat{x}_{t|t,\tilde{\Theta}} + \tilde{L}_t \tilde{C}_t \tilde{A}_t \left(x_t - \hat{x}_{t|t,\tilde{\Theta}}\right) + \tilde{L}_t \tilde{C}_t w_t + \tilde{L}_t z_{t+1} \tag{54}$$

$$y_{t+1,\Theta} = CA\hat{x}_{t|t,\tilde{\Theta}} - CB\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}} + Cw_t + CA(x_t - \hat{x}_{t|t,\tilde{\Theta}}) + z_{t+1} \tag{55}$$

$$\hat{x}_{t+1|t+1,\Theta} = (I - LC)(A\hat{x}_{t|t,\Theta} - B\tilde{K}_t \hat{x}_{t|t,\tilde{\Theta}}) + Ly_{t+1,\Theta} \tag{56}$$

$$= (I - LC)(A - B\tilde{K}_t)\hat{x}_{t|t,\tilde{\Theta}} + (I - LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}}) + Ly_{t+1,\Theta} \tag{57}$$

$$= (I - LC)(A - B\tilde{K}_t)\hat{x}_{t|t,\tilde{\Theta}} + LC(A - B\tilde{K})\hat{x}_{t|t,\tilde{\Theta}} + LCw_t + LCA(x_t - \hat{x}_{t|t,\tilde{\Theta}}) + (I - LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}}) + Lz_{t+1} \tag{58}$$

$$= (A - B\tilde{K}_t)\hat{x}_{t|t,\tilde{\Theta}} + LCw_t + LCA(x_t - \hat{x}_{t|t,\tilde{\Theta}_t}) + (I - LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\tilde{\Theta}_t}) + Lz_{t+1}. \tag{59}$$

Note that these expressions are the time varying counterparts for the same expressions in Lale et al. [2020]. Thus, the regret decomposition is similar to the regret decomposition derived in Lale et al. [2020], but with some changes. Since we are updating the optimistic choices during the adaptive control each regret term is written using the expressions given (52)-(59). This brings the only significant change in term R_1 in the regret decomposition of Lale et al. [2020]. In order to analyze the effect of policy changes and obtain a similar analysis for R_1, we obtain these two terms:

$$R_1 = \sum_{t=1}^{T} \Big\{ \hat{x}_{t|t,\hat{\Theta}}^\top \left(\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t\right) \hat{x}_{t|t,\hat{\Theta}} - \mathbb{E}\left[\hat{x}_{t+1|t+1,\Theta}^\top \left(\tilde{P}_{t+1} - \tilde{C}_{t+1}^\top Q \tilde{C}_{t+1}\right) \hat{x}_{t+1|t+1,\Theta} \middle| \hat{x}_{t|t-1}, y_t, u_t\right]\Big\}$$

$$R_{\text{update}} = \sum_{t=1}^{T} \mathbb{E}\left[\hat{x}_{t+1|t+1,\Theta}^\top \left((\tilde{P}_t - \tilde{C}_t^\top Q \tilde{C}_t) - (\tilde{P}_{t+1} - \tilde{C}_{t+1}^\top Q \tilde{C}_{t+1})\right) \hat{x}_{t+1|t+1,\Theta} \middle| \hat{x}_{t|t-1}, y_t, u_t\right]$$

Therefore, due to Lemma 4.1, the overall regret decomposition can be represented as

$$\begin{aligned}\sum_{t=1}^{T}\left(y_t^\top Q y_t + u_t^\top R u_t\right) &= \sum_{t=1}^{T} J_*(\hat{\Theta}) + R_1 + R_2 - R_3 - R_4 - R_5 - R_6 - R_7 - R_8 - R_9 - R_{10} - R_{11} - R_{\text{update}} \\ &\leq T J_*(\Theta) + R_1 + R_2 - R_3 - R_4 - R_5 - R_6 - R_7 - R_8 - R_9 - R_{10} - R_{11} - R_{\text{update}}\end{aligned} \tag{60}$$

for

$$\begin{aligned}
R_2 &= \sum_{t=1}^{T}\left\{y_t^\top Q y_t - \mathbb{E}\left[y_{t+1,\Theta}^\top Q y_{t+1,\Theta}\middle|\hat{x}_{t|t-1}, y_t, u_t\right]\right\}, \\
R_3 &= \sum_{t=1}^{T}\left\{\hat{x}_{t|t,\hat{\Theta}}^\top(\tilde{A}_t - \tilde{B}_t\tilde{K}_t)^\top \tilde{C}_t^\top Q\tilde{C}_t(\tilde{A}_t - \tilde{B}_t\tilde{K}_t)\hat{x}_{t|t,\hat{\Theta}} - \hat{x}_{t|t,\hat{\Theta}}^\top(A - B\tilde{K}_t)^\top C^\top QC(A - B\tilde{K}_t)\hat{x}_{t|t,\hat{\Theta}}\right\}, \\
R_4 &= \sum_{t=1}^{T}\left\{\hat{x}_{t|t,\hat{\Theta}}^\top(\tilde{A}_t - \tilde{B}_t\tilde{K}_t)^\top(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t)(\tilde{A}_t - \tilde{B}_t\tilde{K}_t)\hat{x}_{t|t,\hat{\Theta}} - \hat{x}_{t|t,\hat{\Theta}}^\top(A - B\tilde{K}_t)^\top(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t)(A - B\tilde{K}_t)\hat{x}_{t|t,\hat{\Theta}}\right\}, \\
R_5 &= -\sum_{t=1}^{T}\left\{2\hat{x}_{t|t,\hat{\Theta}}^\top(A - B\tilde{K}_t)^\top(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t)(I - LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}})\right\}, \\
R_6 &= -\sum_{t=1}^{T}\left\{(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}})^\top A^\top(I - LC)^\top(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t)(I - LC)A(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}})\right\}, \\
R_7 &= \sum_{t=1}^{T}\left\{\mathbb{E}\left[w_t^\top \tilde{C}_t^\top Q\tilde{C}_t w_t\right] - \mathbb{E}\left[w_t^\top C^\top QCw_t\right]\right\}, \\
R_8 &= \sum_{t=1}^{T}\left\{\mathbb{E}\left[w_t^\top \tilde{C}_t^\top \tilde{L}_t^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)\tilde{L}_t\tilde{C}_t w_t\right] - \mathbb{E}\left[w_t^\top C^\top L^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)LCw_t\right]\right\}, \\
R_9 &= \sum_{t=1}^{T}\left\{\mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)^\top \tilde{A}_t^\top \tilde{C}_t^\top Q\tilde{C}_t\tilde{A}_t\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)\middle|\hat{x}_{t|t-1}, y_t\right]\right. \\
&\qquad \left. - \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)^\top A^\top C^\top QCA\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)\middle|\hat{x}_{t|t-1}, y_t\right]\right\}, \\
R_{10} &= \sum_{t=1}^{T}\left\{\mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)^\top \tilde{A}_t^\top \tilde{C}_t^\top \tilde{L}_t^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)\tilde{L}_t\tilde{C}_t\tilde{A}_t\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)\middle|\hat{x}_{t|t-1}, y_t\right]\right. \\
&\qquad \left. - \mathbb{E}\left[\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)^\top A^\top C^\top L^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)LCA\left(x_t - \hat{x}_{t|t,\hat{\Theta}}\right)\middle|\hat{x}_{t|t-1}, y_t\right]\right\}, \\
R_{11} &= \sum_{t=1}^{T}\left\{2\mathbb{E}\left[z_{t+1}^\top L^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)(\tilde{L}_t - L)z_{t+1}\right] + \mathbb{E}\left[z_{t+1}^\top(\tilde{L}_t - L)^\top\left(\tilde{P}_t - \tilde{C}_t^\top Q\tilde{C}_t\right)(\tilde{L}_t - L)z_{t+1}\right]\right\},
\end{aligned}$$

where (60) follows due to optimistic choice of system parameters. This gives us the following regret decomposition for the adaptive control period of LqgOpt:

$$\text{REGRET}(T) \leq R_1 + R_2 - R_3 - R_4 - R_5 - R_6 - R_7 - R_8 - R_9 - R_{10} - R_{11} - R_{\text{update}}. \tag{61}$$

F Regret Analysis, Proof of Theorem 4.1

Notice that $R_1 - R_{11}$ given above have the same properties of $R_1 - R_{11}$ of Lale et al. [2020]. The only difference is that during the adaptive control period of LqgOpt, the agent updates its estimate of the underlying system using the doubling trick mentioned in the main text and in Algorithm 1. Therefore, with probability at least $1 - 5\delta$, the regret of each term has the following structure,

$$R_i = \tilde{\mathcal{O}}\left(\frac{T_w}{\sqrt{T_w}} + \frac{2T_w}{\sqrt{2T_w}} + \frac{4T_w}{\sqrt{4T_w}} + \dots\right)$$

for $i = 1, 3 \dots, 11$. Since $R_2 = \tilde{\mathcal{O}}(\sqrt{T - T_w})$ and using Lemma H.2, we get that $R_i = \tilde{\mathcal{O}}\left(\sqrt{T}\right)$ for $i = 1, \dots, 11$ with probability at least $1 - 5\delta$. Notice that there are $\log(T)$ policy changes, *i.e.* there are $\log(T)$ terms in the summation of R_{update}. Each term is bounded by $2\left(D + \|Q\|\left(\|C\| + \Delta C\right)^2\right)\tilde{\mathcal{X}}^2$. Thus, we have $|R_2| \leq 2\left(D + \|Q\|\left(\|C\| + \Delta C\right)^2\right)\tilde{\mathcal{X}}^2 \log(T) = \mathcal{O}(\log(T))$. Combining all, we conclude that during the adaptive control period of LqgOpt $\text{REGRET}(T) = \tilde{\mathcal{O}}\left(\sqrt{T}\right)$.

□

G System Identification with Non-Steady State Initial Point

$$\begin{aligned} x_{t+1} &= \bar{A}_t x_t + B u_t + A L_t y_t \\ y_t &= C x_t + e_t. \end{aligned} \tag{62}$$

where $\bar{A}_t = A - AL_tC$. If the system is at steady state, *i.e.* $L_t = L = \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}$. Since the system is stable, the dynamics of the system approaches exponentially fast to the steady state dynamics. Therefore, starting at $x_0 = 0$ and with a long enough burning period such that $\|F_t - F\| = O\left(\frac{1}{\text{poly}(T)}\right)$, starting from arbitrary point will provide additional bias term in the estimation which decays over time:

$$\begin{aligned} y_H &= \mathbf{M}\phi_H + e_H + (\mathbf{M_H} - \mathbf{M})\phi_H \\ y_{H+1} &= \mathbf{M}\phi_{H+1} + e_{H+1} + (\mathbf{M_{H+1}} - \mathbf{M})\phi_{H+1} + C\left(\prod_{i=1}^{H} \bar{A}_{H+1-i}\right) x_1 \\ &\vdots \\ y_t &= \mathbf{M}\phi_t + e_t + (\mathbf{M_t} - \mathbf{M})\phi_t + C\left(\prod_{i=1}^{H} \bar{A}_{t-i}\right) x_{t-H} \end{aligned}$$

where

$$\mathbf{M} = \left[CF,\ C\bar{A}F,\ \dots,\ C\bar{A}^{H-1}F,\ CB,\ C\bar{A}B,\ \dots,\ C\bar{A}^{H-1}B\right] \in \mathbb{R}^{m \times (m+p)H}$$

$$\mathbf{M_t} = \left[CF_{t-1},\ C\bar{A}_{t-1}F_{t-2},\ \dots,\ C\left(\prod_{i=1}^{H-1} \bar{A}_{t-i}\right) F_{t-H},\ CB,\ C\bar{A}_{t-1}B,\ \dots,\ C\left(\prod_{i=1}^{H-1} \bar{A}_{t-i}\right) B\right]$$

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

Note that for any t, $\mathbf{M_t} - \mathbf{M}$ represents the model mismatch from the steady-state model parameters and the parameters of the evolving system. The noise terms are zero-mean including the effect of initial state since we assume that $x_0 = 0$. The model mismatch combined with the upper bound on ϕ_t can be used to define the additional bias in the estimation. Notice that this bias will decrease over time since the system approaches exponentially fast to the steady state dynamics. We leave the exact analysis to future work.

H Technical Lemmas and Theorems

Theorem H.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 H.2 (Self-normalized bound for vector-valued martingales [Abbasi-Yadkori et al., 2011]). *Let $(\mathcal{F}_t; k \geq 0)$ be a filtration, $(m_k; k \geq 0)$ be an $\mathbb{R}^d$-valued stochastic process adapted to $(\mathcal{F}_k)$, $(\eta_k; k \geq 1)$ be a real-valued martingale difference process adapted to $(\mathcal{F}_k)$. Assume that η_k is conditionally sub-Gaussian with constant R. Consider the martingale*

$$S_t = \sum_{k=1}^{t} \eta_k m_{k-1}$$

and the matrix-valued processes

$$V_t = \sum_{k=1}^{t} m_{k-1} m_{k-1}^\top, \quad \overline{V}_t = V + V_t, \quad t \geq 0$$

Then for any $0 < \delta < 1$, with probability $1 - \delta$

$$\forall t \geq 0, \quad \|S_t\|_{V_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det\left(\overline{V}_t\right)^{1/2} \det(V)^{-1/2}}{\delta} \right)$$

Lemma H.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 $1 - \delta$, $\|v\| \leq R\sqrt{2d\log(2d/\delta)}$.*

Lemma H.2 (Doubling Trick [Jaksch et al., 2010]). *For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} \coloneqq \max\left\{1, \sum_{i=1}^{k-1} z_i\right\}$*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2}+1)\sqrt{Z_n}$$