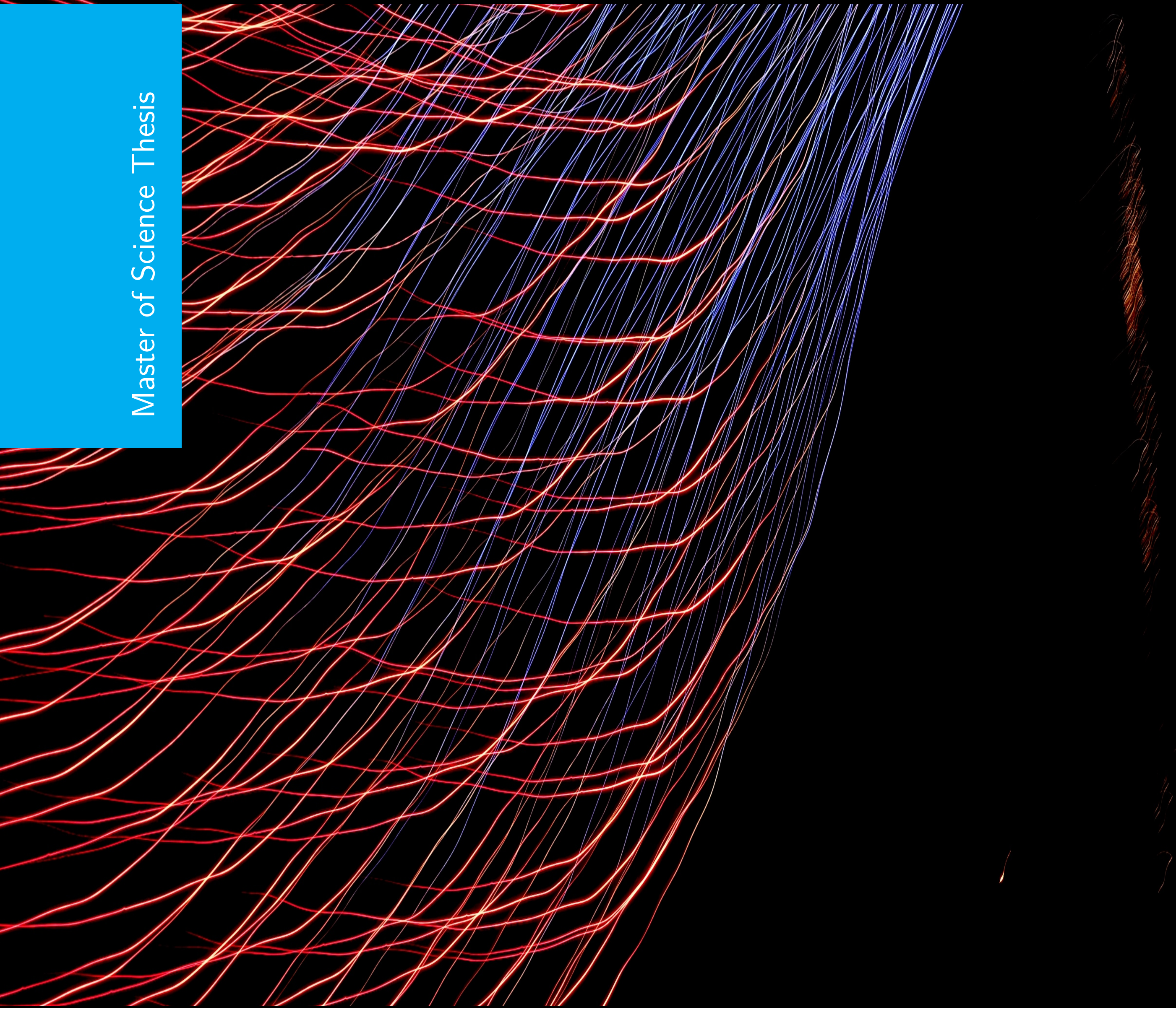
Regret Analysis of Learning-Based Linear Quadratic Gaussian Control with Additive Exploration

Archith Athrey

Master of Science Thesis



Delft University of Technology

Delft Center for Systems and Control

Regret Analysis of Learning-Based Linear Quadratic Gaussian Control with Additive Exploration

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Archith Athrey

October 15, 2023

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Delft University of Technology
Department of
Delft Center for Systems and Control (dcsc)

The undersigned hereby certify that they have read and recommend to the Faculty of Mechanical, Maritime and Materials Engineering (3mE) for acceptance a thesis entitled

Regret Analysis of Learning-Based Linear Quadratic Gaussian Control with Additive Exploration

by

Archith Athrey

in partial fulfillment of the requirements for the degree of

Master of Science Systems and Control

Dated: October 15, 2023

Supervisor(s): ____________________

Prof.dr.ir. Bart De Schutter

____

Dr.ir. Shengling Shi

Reader(s): ____________________

Dr.ir. Mohammad Khosravi

____

Dr.ir. Othmane Mazhar

Abstract

In recent years, there has been considerable interest in the Learning-Based Control (LBC) of *unknown* linear systems in the Linear Quadratic (LQ) paradigm. In the field of optimal control, the LQ control has been a benchmark for decades and is extensively used to control systems in the real world. Moreover, the insights gleaned from studying LQ control problems can be translated into a critical understanding of more complex control problems. In this setting of learning-based LQ control, the control action influences not only the control performance but also the rate at which the system is being learnt, causing a conflict between learning and control (exploration and exploitation), which is particularly challenging to address. Of particular relevance to most practical applications is the Linear Quadratic Gaussian (LQG) control problem, which addresses the control of partially observable linear dynamical systems driven by additive white Gaussian noises. The LQG control of unknown systems poses a significant challenge when compared with Linear Quadratic Regulator (LQR), where the states are measured. The primary aim of this thesis is to develop a novel LBC algorithm for *unknown* partially observable systems in the LQG setting, that is computationally efficient and can guarantee an optimal exploration-exploitation trade-off, quantified by a metric called regret. The regret quantifies the cumulative performance gap over time between the LBC policy and the ideal controller having full knowledge of the true system dynamics. The contributions in this thesis involve a novel LBC algorithm that is deployed in a two-phase structure. The first phase involves injecting Gaussian input signals to obtain an initial model of the system. The subsequent second phase deploys the proposed LBC strategy in an episodic setting, where for each episode, the model is updated, and the resulting updated LQG controller is applied with additive Gaussian signals for exploration. In addition, the thesis establishes strong theoretical guarantees on the regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$, matching the optimal regret growth in the LQR setting up to poly-logarithmic factors. This guarantee is also validated in numerical simulations. With an aim to further reduce the exploration cost, another LBC algorithm, as an extension to the above algorithm, is also developed to bridge the gap between regret minimisation and experiment design techniques in the field of system identification. Simulation results are provided in support of the proposed algorithm.

Table of Contents

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals for their invaluable support and contributions to the completion of this thesis.

First and foremost, I would like to thank Prof.dr.ir. Bart De Schutter, my primary thesis supervisor for his mentorship throughout the course of this thesis. Dr.ir. Shengling Shi, my daily thesis supervisor, deserves special recognition for his unwavering guidance, insightful feedback, expertise, and dedication to the project. His constant encouragement has been invaluable. His mentorship went beyond the academic realm, and I am truly grateful for his support. I extend my thanks to Dr.ir. Mohammad Khosravi for agreeing to be a member of the thesis defence committee. A special mention to Dr.ir. Othmane Mazhar for his thoughtful critiques and valuable suggestions that have significantly improved the quality and rigour of this work. My deepest appreciation goes to my family for their unwavering encouragement, love, and understanding. Their support has sustained and nurtured me throughout this academic journey. To my friends, thank you for being a constant source of laughter and fun.

Delft, University of Technology
October 15, 2023

Archith Athrey

To my family

Chapter 1

Introduction

In many applications, the system dynamics cannot always be determined due to the system's complexity or unmeasured disturbances. Hence in practice, knowledge about the system dynamics is generally known only with uncertainty if not completely unknown. There are several real-world applications where control of unknown or uncertain systems is required: navigation of mobile robots in unknown environments [54], optimal multi-product inventory control [42], or even analysing the anti-synchronisation behaviour of predator-prey populations [64]. These systems, which are generally nonlinear can however be sufficiently approximated by linear models near their operating points. Working with linear models is ideal because of the rich history of analytical solutions in stability analyses and optimal control strategies, and moreover, they are easy to interpret and work with. To identify such unknown *linear* systems, tools such as Prediction Error Method (PEM) [41], subspace identification [65], and maximum likelihood estimation have been developed. One could use such methods to first identify the system, and then use the estimated model to develop optimal control laws. Learning the system dynamics is of particular interest since the knowledge of system dynamics could potentially aid in deploying various other model-based methods like disturbance rejection and output feedback control, for instance. In many cases, the 'estimate-then-control' method is impractical if there are strict limitations on the resources that are involved in running the system, or if the system parameters can change over time. Further, it is not always necessary to obtain a precise understanding of the system dynamics to design a control policy to satisfy some performance criteria [67]. One of the solutions to this problem of controlling unknown systems is through adaptive control or Learning-Based Control (LBC), where the controller is updated online from the collected data, to satisfy some performance measures [44]. This type of model-based reinforcement learning is particularly appealing given the significant advancements in handling large quantities of data efficiently [47].

In recent years, several developments in the control over large state space in the field of reinforcement learning have demonstrated tremendous success in various applications like robotics [40], Atari [46], and Go [57], which have led to further developments in ensuring reliability and sample efficiency, with an emphasis on non-asymptotic guarantees [58]. Failure in such control systems could potentially lead to catastrophic consequences: loss of human life as well as economic loss [53]. While the field of reinforcement learning has skyrocketed with the development of model-free tools, the field of control theory has matured over the years in the design of robust and reliable model-based tools that can guarantee performance while adhering to the safety limitations of the system. Hence, the way forward is to combine the effectiveness of the methods developed in reinforcement learning with the strong-theoretic guarantees offered by control theory [53]. Moreover, the model-free methods of reinforcement learning are not as sample-efficient as the model-based methods [63].

In the paradigm of optimal control, Linear Quadratic (LQ) control has been a benchmark for decades and is extensively used to control complex systems [11]. Moreover, the insights gleaned from studying LQ control problems can be translated into a critical understanding of more complex control problems. Further, studying LQ control problems allows one to place reinforcement learning and control theory on equal footing [53]. The significant research that exists in the learning-based LQR control problem however dwarfs the research in the learning-based Linear Quadratic Gaussian (LQG) setting [11]. The LQG control problem which addresses the control of partially observable linear dynamical systems driven by additive white Gaussian noises, is one of the key issues in adaptive control [12]. Moreover, in most practical applications, assuming full state-measurement can be restrictive. The seemingly benign difference of not being able to measure the true states will in fact pose a significant challenge when controlling the system with unknown dynamics [36]. The errors in state estimates due to approximate models could potentially accumulate to have a significant impact on the control performance. This is precisely why LBC in the partial observability setting is a particularly challenging problem to address. The LBC strategies that do address this setting either incorporate subroutines that require non-convex optimisation [37], [36], or require restrictive assumptions on the optimal control policy [34].

Hence, this thesis aims to design an LBC algorithm in the LQG setting that is computationally efficient, and can effectively balance the *exploration-exploitation* trade-off, quantified by a metric called regret, detailed in Section 2-5. The proposed LBC algorithm is deployed in a two-phase structure. The first phase involves injecting Gaussian input signals to obtain an initial model of the system. The subsequent second phase deploys the proposed LBC strategy in an episodic setting, where for each episode, the model is updated, and the resulting updated LQG controller is applied with additive Gaussian signals for exploration. This thesis establishes a theoretical guarantee on $\tilde{\mathcal{O}}(\sqrt{T})$ upper bound on the regret growth for LQG-NAIVE (Algorithm 2), which matches the optimal rate of regret growth in the LQR setting. Further, this thesis also provides compelling simulation results for LQG-NAIVE.

The thesis is structured in the following way. In Chapter 2, the LQG control problem and the various concepts relevant to model-based reinforcement learning, are introduced. The various LBC strategies and their limitations in the current research landscape are also highlighted. Following the preliminary details, Chapter 3 provides motivation for the proposed LBC algorithms. Supporting this proposition, are finite-time stability and regret guarantees of the LQG-NAIVE algorithm (Algorithm 2), in Chapter 4. Chapter 4 also provides simulations validating the theoretical regret guarantee of LQG-NAIVE. As an extension to the thesis, LQG-IF2E is proposed (Algorithm 3), which paves the way into incorporating 'intelligence' into the LBC strategy through the Fisher Information Matrix (FIM), as detailed in Section 3-3. In this thesis, only empirical validation of the FIM-based LBC policy is provided.

Chapter 2

Background

2-1 Notations

The Euclidean norm of a vector x is denoted by $||x||$. For a matrix $X \in \mathbb{R}^{n\times m}$, $||X||$ denotes the spectral norm, $\rho(X)$ denotes the spectral radius, $||X||_{\mathrm{F}}$ denotes the Frobenius norm, $X^\top$ denotes its transpose, $X^\dagger$ denotes the Moore-Penrose inverse, and $\mathrm{Tr}(X)$ denotes the trace. The determinant of a matrix X is denoted by $\det(X)$. The j^{th} singular value of a matrix X is denoted by $\sigma_j(X)$, where $\sigma_{\max}(X) \coloneqq \sigma_1(X) \geq \sigma_2(X) \geq ... \geq \sigma_{\min}(X) \coloneqq \sigma_{\min(n,m)}(X) > 0$. Similarly, $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ have analogous meanings for the eigenvalues of X. The identity matrix with the appropriate dimension is denoted by I and similarly, 0 is a matrix or a vector of 0's with appropriate dimensions. Further, $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with a mean vector μ and a covariance matrix Σ. The expectation operator is denoted by $\mathbb{E}$, and $\mathbb{P}$ denotes the probability of an event occurring. The inequality $f \lesssim g$ denotes $f \leq Cg$ for a universal constant C, and $f \lessapprox g$ denotes informal inequality. The informal inequality is used when it is required to hide some of the terms in g. The Kronecker product is denoted by $\otimes$, `vec` denotes the vectorisation operator. Further, $\mathtt{D}_\theta$ denotes Jacobian, $\mathtt{d}_\theta$ denotes differential, and ∇_θ denotes gradient, with respect to θ.

In this thesis, $\hat{X}$ is used to denote an approximation of the true quantity X. Further, $\hat{X}_t$ is used to denote an approximation of the true quantity X, at time step t or at the t^{th} episode. The intended meaning of this notation becomes clear with the context. The 'big - O' notation ($\mathcal{O}(.)$) for two functions $f(x)$ and $g(x)$, is defined as $f(x) = \mathcal{O}(g(x))$ if $\exists C > 0$ and $\tilde{x} \in \mathbb{R}$ such that $|f(x)| \leq Cg(x)\ \forall x \geq \tilde{x}$. The 'big-omega' notation ($\Omega(.)$) for two functions $f(x)$ and $g(x)$, is defined as $f(x) = \Omega(g(x))$ if $\exists C > 0$ and $\tilde{x} \in \mathbb{R}$ such that $|f(x)| \geq Cg(x)\ \forall x \geq \tilde{x}$. The notations $\tilde{\mathcal{O}}(.)$ and $\tilde{\Omega}(.)$ ignores constants and poly-logarithmic terms.

2-2 Linear Quadratic Gaussian (LQG) control problem

In this setting, a discrete-time Linear Time Invariant (LTI) system is described by the state-space equation:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t, \;\; w_t \sim \mathcal{N}(0, \sigma_w^2 I), \\ y_t &= Cx_t + z_t, \;\; z_t \sim \mathcal{N}(0, \sigma_z^2 I), \end{aligned} \tag{2-1}$$

for $t = 0, 1, 2, 3, ...$, $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, and $C \in \mathbb{R}^{n_y \times n_x}$. At time step t, $u_t \in \mathbb{R}^{n_u}$ is the input, $x_t \in \mathbb{R}^{n_x}$ is the state, $w_t \in \mathbb{R}^{n_x}$ is the process noise, $y_t \in \mathbb{R}^{n_y}$ is the system output, and $z_t \in \mathbb{R}^{n_y}$ is the measurement noise. Let the system parameter Θ corresponding to the true system be

$$\Theta = (A, B, C, L), \tag{2-2}$$

where L is the Kalman gain as described in (2-6). To measure the performance of a controller, the cost incurred c_t at time step t is defined to be quadratically dependent on the outputs and inputs as follows:

$$c_t = y_t^\top Q y_t + u_t^\top R u_t, \tag{2-3}$$

where $Q \in \mathbb{R}^{n_y \times n_y}$ is positive semi-definite and $R \in \mathbb{R}^{n_u \times n_u}$ is positive definite. In this thesis, the infinite-horizon setting is considered wherein the goal is to design an input signal such that the long-term average expected cost is minimised. The long-term average expected cost in this setting is given by

$$J = \lim_{T \to \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} c_t \right]. \tag{2-4}$$

A linear system with parameter Θ is controllable if the controllability matrix denoted by $\mathbf{C}(A, B, n_x)$, where

$$\mathbf{C}(A, B, n_x) := \begin{bmatrix} B & AB & A^2B & . & . & . & A^{n_x-1}B \end{bmatrix},$$

has full row rank. Similarly, a linear system with parameter Θ is observable if the observability matrix denoted by $\mathbf{O}(A, C, n_x)$, where

$$\mathbf{O}(A, C, n_x) := \begin{bmatrix} C \\ CA \\ CA^2 \\ . \\ . \\ . \\ CA^{n_x-1} \end{bmatrix},$$

has full column rank.

Assumptions 2.1 The thesis assumes the following about the true system:

1. (A, B) is controllable, (A, C) is observable, and (A, F) is controllable. Here, F is the Kalman gain in the innovations form (2-10).
2. $(Q^{1/2}, A)$ is observable and $(A, \sigma_w I)$ is reachable.
3. Q is positive semi-definite and R is positive definite.

In system identification settings that guarantee the capability of accurate system parameter estimates and in optimal control settings, the aforementioned assumptions are frequently made [34], [36], [37], [50], [61]. If Θ is known, the Kalman filter can provide the state estimate $\hat{x}$ given the measured inputs and outputs up to time step $t-1$, denoted by $\hat{x}_{t|t-1,\Theta}$. If the assumptions made in Assumptions 2.1 hold, the estimated state error variance $\Sigma_{t|t-1}$ converges asymptotically to [65]:

$$\lim_{t\to\infty} \Sigma_{t|t-1} = \lim_{t\to\infty} \mathbb{E}\left[(x_t - \hat{x}_{t|t-1,\Theta})(x_t - \hat{x}_{t|t-1,\Theta})^\top\right] = \Sigma > 0,$$

where Σ is the solution to the following Discrete Algebraic Riccati Equation (DARE) [65]:

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

At steady-state, i.e., after $\Sigma_{t|t-1}$ converges (exponentially) to Σ, the state estimates can be efficiently estimated by the Kalman filter:

$$\begin{aligned} \hat{x}_{t|t,\Theta} &= (I - LC)\hat{x}_{t|t-1,\Theta} + Ly_t, \\ \hat{x}_{t+1|t,\Theta} &= A\hat{x}_{t|t,\Theta} + Bu_t, \\ L &= \Sigma C^\top\left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1}, \end{aligned} \tag{2-6}$$

where L is the Kalman gain. Then, an optimal control law of the form

$$u_t = -K\hat{x}_{t|t,\Theta}, \tag{2-7}$$

minimising J can be obtained from the separation principle with K being the optimal feedback gain matrix obtained from

$$K = (B^\top PB + R)^{-1}B^\top PA, \tag{2-8}$$

where P is the solution to the following DARE [12]:

$$P = C^\top QC + A^\top PA - A^\top PB\left(B^\top PB + R\right)^{-1} B^\top PA. \tag{2-9}$$

In (2-6), the two expressions concerning $\hat{x}_{t|t,\Theta}$ and $\hat{x}_{t+1|t,\Theta}$ can be combined to obtain the innovations form:

$$\begin{aligned} \hat{x}_{t+1|t,\Theta} &= A\left((I - LC)\hat{x}_{t|t-1,\Theta} + Ly_t\right) + Bu_t \\ &= A\hat{x}_{t|t-1,\Theta} + Bu_t + Fe_t, \\ e_t &= C\left(x_t - \hat{x}_{t|t-1,\Theta}\right) + z_t, \\ e_t &\sim \mathcal{N}(0, C\Sigma C^\top + \sigma_z^2 I), \end{aligned} \tag{2-10}$$

where F given by $F = AL$ is the Kalman gain in the innovations form. Further, the innovations form (2-10) can be expanded to obtain the one-step-ahead prediction model:

$$\begin{aligned} \hat{x}_{t+1|t,\Theta} &= (A - FC)\hat{x}_{t|t-1,\Theta} + Bu_t + Fy_t, \\ \hat{y}_{t+1|t,\Theta} &= C\hat{x}_{t+1|t,\Theta}, \end{aligned} \tag{2-11}$$

where the Kalman gain here ensures $A - FC$ is asymptotically stable. There exists a closed-form expression for the optimal long-term average expected cost when applying the optimal control law as described in (2-7) [36]:

$$J_* := \min_{u_0,u_1,...} J = \mathrm{Tr}\left(C^\top QC\bar{\Sigma}\right) + \sigma_z^2 \mathrm{Tr}\left(Q\right) + \mathrm{Tr}\left(P(\Sigma - \bar{\Sigma})\right), \tag{2-12}$$

where

$$\bar{\Sigma} = \Sigma - \Sigma C^\top \left(C\Sigma C^\top + \sigma_z^2 I\right)^{-1} C\Sigma. \tag{2-13}$$

2-3 Learning-based control

In this thesis, a variant of the LBC technique is proposed, which addresses the problem of controlling an unknown system, i.e., Θ is unknown whereas, Q and R are user-defined (known). In this setting, the acquired information about the system behaviour by the controller (or agent) is used to approximate the system parameter, where the approximation of the system parameter is denoted by $\hat{\Theta}$, thereby reducing the parameter uncertainty. As a consequence, the control law is tuned appropriately to control the true underlying system with parameter Θ, thereby achieving better control performance.

To be more precise, at time step t, the controller can access the past observations denoted by $\mathcal{I}_t$, where

$$\mathcal{I}_t = \{y_0, u_0, y_1, u_1, ..., y_{t-1}, u_{t-1}, y_t\}, \tag{2-14}$$

based on which, a control input u_t is computed. By injecting u_t into the true system described by (2-1), the state of the system transitions from x_t to x_{t+1}. Consequently, a cost c_t is incurred. The observation is then updated to $\mathcal{I}_{t+1} = \mathcal{I}_t \cup \{u_t, y_{t+1}\}$.

As mentioned before, only an approximate system parameter $\hat{\Theta}$ is known, which can lead to sub-optimal control performance. This sub-optimality in the control performance is quantified with the sub-optimality gap ($\Delta_{\hat{\Theta}}$) in the long-term average expected cost [43]:

$$\Delta_{\hat{\Theta}} = J(\hat{\Theta}) - J_*, \tag{2-15}$$

where $J(\hat{\Theta})$ is the long-term average expected cost incurred when using the control law that is optimal for a system with parameter $\hat{\Theta}$ onto the true system with parameter Θ.

2-4 Exploration-exploitation trade-off

The problem setting considered here consists of two main goals, that are to be attained simultaneously:

1. Controlling an unknown system in order to satisfy a certain performance criterion, be it maximising the performance index or minimising the cost.
2. Learning the optimal control policy for the true system by learning the true system dynamics.

This problem comes with its own challenges. To effectively determine what actions to take such that the performance index is maximised in the LBC paradigm considered here, the agent (controller) must acquire the necessary information on the relation between current observations and the corresponding performance index. In other words, the agent must be able to simultaneously learn the dynamics and plan a control policy [10]. Therefore, it is necessary to generate informative data to obtain better estimates of the system parameter. One must also consider a caveat in generating informative data, which may lead to higher costs: in the LQG setting, since the cost (2-3) is quadratically dependent on the outputs and the inputs, perturbing the system behaviour to engender informativity may cause the output and input sequences to deviate significantly from the optimal sequence, thereby incurring larger costs. Hence, there is a need to develop methods that can optimally balance exploration (actions that aid in estimating the optimal control policy by estimating Θ) and exploitation (actions that minimise the cost incurred). The essence of having a balance between exploration and exploitation is called the exploration-exploitation trade-off.

2-5 Regret minimisation

2-5-1 Definition and a brief history

One of the quantitative measures of the exploration-exploitation trade-off is the cumulative regret $R(T)$ [3], which is given by

$$R(T) = \sum_{t=0}^{T-1} (c_t - J_*). \tag{2-16}$$

The cumulative regret (2-16) quantifies the difference between the cost of an LBC policy and the optimal expected average cost J_* under the full knowledge of the true system parameter (oracle). The LBC policy converges to the optimal policy if its regret grows sub-linearly with time, i.e., $\frac{R(T)}{T} \to 0$, which is also called as the *Hannan consistency* [1]. Therefore, for any learning-based policy under consideration, it is required to guarantee at least sub-linear regret in order to pay the optimal expected average cost J_*. The smaller the regret, the faster the LBC policy is converging to the optimal policy.

Another formulation of the cumulative regret denoted by $\bar{R}(T)$ [22], is given by

$$\bar{R}(T) = \sum_{t=0}^{T-1} (c_t - c_{t,*}), \tag{2-17}$$

where $c_{t,*} = {y_{t,*}}^\top Q y_{t,*} + u_{t,*}^\top R u_{t,*}$ is the optimal instantaneous cost of the true system paid at time step t with $u_{t,*}$ being the optimal input (2-7) for the underlying true system with parameter Θ. In the current thesis, the regret definition (2-16) is used because the regret formulation in (2-17) captures only the uncertainty associated with not knowing the true system parameter but not the uncertainty associated with the stochastic behaviour induced by the measurement and process noises. This is because $\bar{R}(T)$ can become negligibly small when applying the optimal control action $u_{t,*} = -K\hat{x}_{t|t,\Theta}$.

In the literature, for the sake of brevity, the regret bounds are generally presented in a way that highlights only its dependence on the time horizon since the rate of the growth of the regret with time is considered to be one of the important factors in comparing different LBC techniques.

Regret minimisation was first studied in [32] and [33] for minimum variance controllers. The proposed methods use white-noise probing inputs when the available information is inadequate for parameter estimation thereby, guaranteeing an asymptotic rate of regret of $\mathcal{O}(\log(T))$. Whereas for LQR control problems, it is shown that even when applying the optimal control action $u_{t,*} = -Kx_t$ to the true system with parameter Θ, the distribution of $\lim_{T\to\infty} \frac{R(T)}{T^{1/2}}$ is a Gaussian random variable centred at zero [20]. This implies $R(T) = \Omega(\sqrt{T})$. This result, which is also confirmed in [58], is not trivial as it provides a lower bound for the regret of LBC policies in the LQR setting.

Whereas in the LQG setting, it was shown that given a set of convex reparameterisation of linear dynamic controllers, persistently exciting the true underlying system and *strongly* convex loss functions e.g. $Q, R > 0$, a polylogarithmic regret upper bound on $\bar{R}(T)$ can be achieved [34]. Complementing this result, [70] establishes that polylogarithmic regret is not possible if $KK^\top \not\succ 0$ or in other words, if the optimal control law does not persistently excite the true system in the LQG setting. In this case, the best regret upper bound that one can achieve is $\mathcal{O}(\sqrt{T})$.

Remark 1 The regret lower bound of $\bar{R}(T) = \Omega(\sqrt{T})$ in the LQG setting derived in [70] under the condition that $KK^\top \not\succ 0$, motivates answering a more fundamental question: which system instances are easy to control, and which are easy to learn? To answer this, the work in [70] shows that systems that are marginally stable or with large Kalman filter gain (poor observability) are fundamentally hard to (be learned to) control. Further, the work in [62] concludes that under-actuated and/or under-excited systems with weak state coupling are hard to learn. More precisely, it is shown that the controllability index directly influences the ease with which the system can be identified. Therefore, the performance of a regret minimisation algorithm critically depends on the system properties.

2-5-2 Relation between the two regret definitions

In [23], the difference between the two regret definitions ((2-16) and (2-17)) is investigated in the LQR setting, which shows

$$\limsup_{T\to\infty} \frac{R(T) - \bar{R}(T)}{T^{1/2}\log T} = \limsup_{T\to\infty} \frac{\sum_{t=0}^{T-1} c_{t,*} - TJ_*}{T^{1/2}\log T} < \infty. \tag{2-18}$$

The result (2-18) implies that the difference between the two definitions of regret given by $\sum_{t=0}^{T-1} c_{t,*} - TJ_*$, grows at the rate $\mathcal{O}(T^{1/2}\log T)$ with high probability. This is shown to hold when the moment condition $\sup_{t\geq 1} \mathbb{E}[||w(t)||^4] < \infty$ is satisfied.

Corollary 2.1 If (2-18) holds, then $R(T) = \tilde{\mathcal{O}}(T^{1/2})$ if and only if $\bar{R}(T) = \tilde{\mathcal{O}}(T^{1/2})$.

Therefore, if one were to provide a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ for either definition of regret, i.e., (2-16) or (2-17), it does not matter which definition of regret is used. Further, establishing a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ for one expression of regret implies that the other expression also scales at $\tilde{\mathcal{O}}(\sqrt{T})$. This result is presented in [23] but without proof. The proof of this corollary is presented in Appendix A-3 for the sake of completeness.

2-5-3 Regret bound under high probability vs under expectation

Most of the existing works on LQ-LBC provide regret guarantees that hold with high probability $1-\delta$, where $\delta \in (0,1)$ [11]. This probabilistic regret guarantee is derived by considering that certain event(s) holds with a probability of at least $1-\delta$ [3], [29]. Such an event could be, for instance, when a certain confidence bound on the parameter estimate is satisfied [37], or when the magnitude of the state vector remains bounded [3]. This implies that the algorithm is parameterised by δ, i.e., the regret guarantees with high probability show that

$$\mathbb{P}[R(T) \leq \text{poly}(n_x, n_u, n_y, T, 1/\delta)] \geq 1-\delta. \tag{2-19}$$

On the other hand, the algorithms that derive expected regret require the consideration of the case where the event fails to hold as well [29]. In short, expected regret [44] intends to show that

$$\mathbb{E}[R(T)] \leq \text{poly}(n_x, n_u, n_y, T). \tag{2-20}$$

However, by considering δ as a function of T, it is possible to transform an algorithm that provides probabilistic regret guarantees into the one that provides expected regret guarantees [29]. In this thesis, a probabilistic regret upper bound is provided and the extension to an expected regret guarantee is deferred to future work.

2-6 Open-loop vs closed-loop system identification

Since there is an emphasis on learning in the considered online setting, one must weigh several factors to decide the kind of input signal to adopt for system identification. Some of the relevant factors are mentioned below.

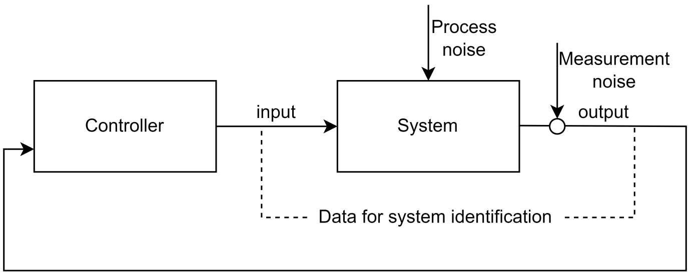


Figure 2-1: Schematic of closed-loop system identification.

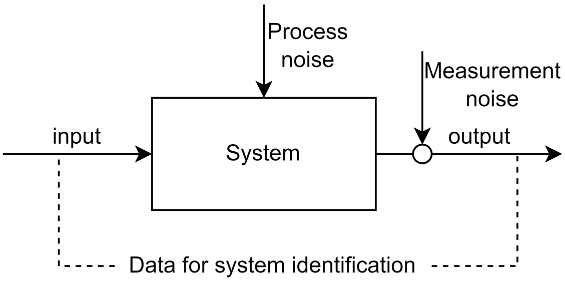


Figure 2-2: Schematic of open-loop system identification.

In closed-loop system identification as depicted in Figure 2-1, the correlation between the input and output data is present especially when the closed-loop controller inherently uses the past observations to deploy control actions [34]. Moreover, if the closed-loop controller has regulating properties then the information obtained from closed-loop data is often less. This could lead to a case where the collected data is not rich or informative enough to obtain accurate model estimates [69]. That being said, there could be cases where the closed-loop identification is preferred [39]:

- when there is a constraint on the system input or output [14], or
- if regulation is required, a stabilising controller may be required.

Moreover, it is possible to design input signals to address the potential lack of data informativity in closed-loop identification, as briefly discussed in Section 2-8.

On the other hand, in open-loop system identification as depicted in Figure 2-2, the correlation between input-output data is significantly reduced with the use of independent inputs. To elaborate, firstly roll back the output equation H time steps in the history as such:

$$y_t = CA^H x_{t-H} + \sum_{i=1}^{H} CA^{i-1} B u_{t-i} + \sum_{i=1}^{H} CA^{i-1} w_{t-i} + z_t. \tag{2-21}$$

Since A is assumed to be stable, the first term in (2-21) decays exponentially and becomes negligible with large enough H. Therefore, from (2-21) we have

$$y_t \approx \sum_{i=1}^{H} G_{[i]} u_{t-i} + \sum_{i=1}^{H} CA^{i-1} w_{t-i} + z_t, \tag{2-22}$$

where the Markov parameters, $G_{[i]} = CA^{i-1}B$, $\forall i > 0$ with $G_{[0]} = 0$, uniquely describes the system behaviour. Moreover, $\mathbf{G} = \begin{bmatrix} G_{[0]} & G_{[1]} & \dots & G_{[H]} \end{bmatrix} \in \mathbb{R}^{n_y \times n_u(H+1)}$ denotes the H-length Markov parameters matrix. From the above formulation in (2-22), a least-squares estimate of $\mathbf{G}$ can be obtained by taking $\bar{u}_t = \begin{bmatrix} u_t & \dots & u_{t-H} \end{bmatrix}$ as the regressor:

$$\hat{\mathbf{G}} = \arg\min_X \sum_{t=H}^{T-1} ||y_t - X\bar{u}_t||^2. \tag{2-23}$$

If i.i.d. Gaussian input signals independent of the noises are used as inputs then, a consistent estimate of $\hat{\mathbf{G}}$ can be attained using (2-23). From this estimate, one can obtain an estimate of (A, B, C) up to similarity transformation by using the popular Ho-Kalman algorithm [28].

As a consequence of this, it is much easier to provide guarantees on the accuracy of the parameter estimates with open-loop input signals [37], [50]. Existing works which employ such methods guarantee that the model estimation error diminishes at the rate $\mathcal{O}(1/\sqrt{T})$, after collecting observation data for T time steps [50], [56], [59], [61]. The regret minimisation techniques in the partial observability setting that utilise this method of open-loop identification, have proposed LBC strategies in the *explore*-then-*commit* framework [36], [59]. In this framework, Gaussian input signals are injected during the *explore* phase following which, the system is identified. Using the identified system, the LBC strategy is deployed during the *commit* phase. These techniques incur a regret of $\tilde{\mathcal{O}}(T^{2/3})$, which is larger when compared with the well-established $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bound [34], [37]. As a consequence of this fragmented approach, these techniques do not generalise well to the current setting where the past observations are perpetually utilised to update the LQG controller. In closed-loop system identification, the inputs will become correlated with the noise sequences and as such, consistent estimates of the Markov parameters cannot be guaranteed by solving (2-23). To address this challenge, [37] proposed an identification technique to estimate the Markov parameters both in open-loop as well as in closed-loop. This identification technique, which is detailed in section 3-1, is adopted in the present thesis.

2-7 Episodic policy update

As the agent is exploring and learning more about the true underlying system, there is a need to determine when to update the current estimate of the model to a 'better' estimate. Such a need can arise if for some algorithms, switching the policy (updating the system parameter estimate) too frequently or at every time step can degrade the control performance [1]. Moreover, by making policy updates less frequent, the computational load can be reduced [22].

The time steps within which a particular system parameter estimate (and the corresponding optimal policy) is maintained is called an episode or an epoch. In LBC, the episodic policy update is categorised based on the number of time steps in each episode, as follows:

- Varying episode length: The algorithm can increase the episode length geometrically, for instance, in a doubling fashion (cf. [58]).
- Constant episode length: The algorithm maintains a constant episode length, that is either given by the user or is estimated (cf. [48]).
- Step-wise episode: The algorithm updates the policy at every time step (cf. [29]).
- Anytime episode length: The policy can be updated at any time step (cf. [29]).

Moreover, switching control policies at the end of each episode may also provide the excitation necessary for system identification [19].

2-8 Exploration strategies to address the exploration-exploitation trade-off

There are three main exploration strategies in the LQ setting for LBC [11]:

1. **Optimism in the Face of Uncertainty (OFU):** Exploration driven by choosing optimistic system parameters from a confidence set [3],[7], [37], [38]. In this exploration strategy, exploration is engendered with the optimal control law of the optimistic system

parameter estimate, i.e., $\exists\, \tilde{\Theta}_t : J_*(\tilde{\Theta}_t) \coloneqq \inf_{\Theta' \in \mathcal{C}_t} J_*(\Theta') \leq J_*$, where $\mathcal{C}_t$ is the confidence set at time step t. Here, $J_*(\tilde{\Theta}_t)$ is the optimal long-term average expected cost of the system with parameter $\tilde{\Theta}_t$. The OFU-based LBC policy is given by

$$u_t = \begin{cases} -K(\tilde{\Theta}_t)x_t, \text{ if } C = I, \sigma_z^2 = 0 \\ -K(\tilde{\Theta}_t)\hat{x}_{t|t,\tilde{\Theta}_t}, \text{ otherwise} \end{cases},$$

where $K(\tilde{\Theta}_t)$ is the optimal feedback gain for $\tilde{\Theta}$.

2. **Forced exploration:** Certainty Equivalence Controller (CEC) with ϵ - greedy exploration [18], [21], [43], [58]. The certainty equivalence principle is based on one of the most simple techniques for regulating dynamical systems with uncertain or unknown dynamics: a model of the system is fitted by collecting its temporal history, and a control strategy is then developed by taking the fitted model as the true system [8]. Whereas, with CEC with ϵ - greedy exploration, the LBC policy is given by

$$u_t = \begin{cases} -K(\hat{\Theta}_t)x_t + \eta_t, \text{ if } C = I, \sigma_z^2 = 0 \\ -K(\hat{\Theta}_t)\hat{x}_{t|t,\hat{\Theta}_t} + \eta_t, \text{ otherwise} \end{cases},$$

where $K(\hat{\Theta}_t)$ is the optimal feedback gain for $\hat{\Theta}_t$ and η_t being an additive excitatory signal.

3. **Thompson Sampling (TS):** Here, either the true system parameter is assumed to belong to a *known* prior distribution and the optimal policy of the sampled system parameter from the posterior distribution is deployed (Bayesian setting), or with no prior assumption on the true system parameter, an optimistic system parameter is sampled from a confidence set around the system parameter estimate, and the corresponding optimistic policy is deployed (frequentist setting) [4], [6], [30], [31]. Here, the LBC policy is given by

$$u_t = \begin{cases} -K(\Theta_t)\hat{x}_{t|t,\Theta_t}, \;\; \Theta_t \sim \mathcal{D}_t, \text{ for the Bayesian setting} \\ -K(\tilde{\Theta}_t)\hat{x}_{t|t,\tilde{\Theta}_t}, \;\; \tilde{\Theta}_t \sim \mathcal{C}_t, \text{ for the frequentist setting} \end{cases},$$

where $\mathcal{D}_t$ is the posterior distribution of the system parameter at time step t, $\mathcal{C}_t$ is the confidence set around the system parameter estimate at time step t, and $K(\Theta_t)$ is the optimal feedback gain of the system Θ_t. The posterior distribution and the confidence set are constructed such that $\Theta \in \mathcal{D}_t$ or $\Theta \in \mathcal{C}_t$, with high probability. For the state feedback setting, one can replace $\hat{x}_{t|t}$ with x_t.

Many of the works employing the above-mentioned LBC policies adopt a two-phase structure: the first phase consists of an initial system identification phase where rich excitatory inputs are deployed to obtain a 'good' initial estimate of the model parameter following which, the LBC strategy is deployed online, where system identification and control take place sequentially.

2-8-1 Related works

The early works of regret minimisation emerged for Auto Regressive Model (ARX) Single-Input-Single-Output (SISO) systems, that use minimum variance controllers [32], [33]. These works show that the regret grows at a rate $\mathcal{O}(\log(T))$ asymptotically whereas for LQR, the regret is shown to grow at a rate $\mathcal{O}(\sqrt{T})$ asymptotically. In the learning-based LQR setting, the seminal work of [3] incorporating OFU, achieving a finite-time regret guarantee of $\tilde{\mathcal{O}}(\sqrt{T})$, reignited the research into regret minimisation for LQ control problems. A similar rate of regret is also shown in the TS-based approach as well as in forced exploration [21], [30], [31], [43]. In fact, $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bound is the optimal rate achievable for unknown systems in the LQR setting [58]. Such a rate can be attained by a simple naive exploration, i.e., CEC with an additive white Gaussian excitatory signal, with a variance diminishing at a rate $\mathcal{O}(1/\sqrt{t})$.

Remark 2 Although the $\sqrt{T}$ regret bound is rate optimal for the LQR control problem with unknown dynamics, poly-logarithmic regret upper bound can be achieved for a known A or B matrix, as a consequence of the extra available information [16], [29].

Both TS and naive exploration are a better substitute for OFU when considering the computational complexity of computing the actions the agent must execute. To elaborate, TS requires sampling only a single instance be it in the Bayesian setting or in the frequentist setting and then, deploying the optimal control law of the sampled instance. Further, the works in the Bayesian TS framework assume Gaussian posterior distribution, whose mean and covariance can be updated from analytical expressions [25], [49]. Naive exploration only requires the system parameter estimate to deploy the optimal control law of the estimated system parameter with an additive perturbation. On the other hand, OFU requires either solving a non-convex optimisation or solving a complex optimisation problem based on semi-definite programming, which are not as computationally efficient as the routines in TS and naive exploration [3], [7], [17].

Most of the works in the literature address the full-state measurement case, i.e., the LQR setting, and there are very few works which address the LQG case [11]. Among those, [36] and [37] use OFU, guaranteeing a $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bound but requires solving a non-convex optimisation to find optimistic system parameters. Following these two works, [34] provides guarantees of poly-logarithmic regret by deploying a disturbance feedback control law, which is a convex reparameterisation of a linear dynamic control law, under the assumption that the optimal policy persistently excites the true system. Recently, $\tilde{\mathcal{O}}(\sqrt{T})$ regret upper bound has also been established with TS in the LQG setting [30]. Forced exploration has not yet been investigated for regret minimisation in the learning-based LQG control problem.

There is a parallel line of research which focuses on designing algorithms, that deploy input signals to generate the required data informativity necessary for estimating an accurate model of the true system while accounting for various experimental constraints [13], [15], [27], [67]. This problem is called the optimal experiment design problem. Although these works in optimal experiment design take the application into account, they are not particularly suitable for regret minimisation [11]. Most of these works optimise some function of the FIM that depends on the unknown true system parameter. To circumvent this issue, an adaptive experiment design framework has been developed, where the initial estimate of the system, as well as the corresponding optimal input sequence to identify the system, are improved as more data is collected [13], [26], [27], [51], [66], [67]. The task-optimal experiment design proposed in [67] is the closest to its application for regret minimisation: in this work, the minimisation of the sub-optimality gap (2-15) is addressed. Further, the work in [67] designs an adaptive scheme to converge to a sequence of input signals that achieves a smaller sub-optimality gap when compared with injecting Gaussian input signals. However, the cost of running this algorithm is not evaluated online and therefore, it cannot be directly adopted for regret minimisation [67].

Of the three methods, OFU can be said to incorporate some form of intelligence in engendering exploration: OFU being a confidence-based method, selects control actions to explore the regions of the parameter space that has the most influence on the control performance [7]. But as mentioned previously, OFU requires solving a non-convex optimisation problem. The work in [18] bridges the gap between experiment design and regret minimisation by proposing the following LBC policy for the LQR setting:

$$\begin{aligned} u_t &= -K(\hat{\Theta}_t)x_t + \eta_t \\ \eta_t &= \sqrt{\frac{\gamma}{\lambda_{\min}\left(I_t(\hat{\Theta}_t)\right)}}\, r_0 \quad \text{for some } \gamma > 0 \text{ and } r_0 \sim \mathcal{N}(0, I), \end{aligned} \tag{2-24}$$

where $I_t(\hat{\Theta}_t)$ is the FIM at time step t evaluated on the estimated system parameter $\hat{\Theta}_t$. This type of exploration is called the Inverse Fisher Feedback Exploration (IF2E). The motivation for using the FIM is provided in Section 3-3. This work [18], guarantees an asymptotic regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ since $\lambda_{\min}\left(I_t(\hat{\Theta}_t)\right)$ is shown to grow at a rate $\mathcal{O}(\sqrt{t})$ asymptotically. The extension to the LQG setting is however lacking.

2-9 Additional assumptions and definitions

The following assumptions aid in simplifying the exposition of stability and regret analyses in the LQG setting [36], [37]:

1. The system is assumed to be open-loop stable, i.e., $\rho(A) < 1$. Define $\Phi(A) := \sup_{\tau \geq 0} \frac{||A^\tau||}{\rho(A)^\tau}$. It is assumed that $\Phi(A) < \infty$. This is a mild assumption that is necessary to quantify the finite-time evolution of the system. The stability of the open-loop plant is assumed to avoid explosive behaviour during the initial system identification phase. For details on the initial system identification phase, refer to Section 3-2.

2. The unknown system parameter Θ is assumed to be member of a set $\mathcal{S}$, such that,

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

The above two assumptions are standard in the majority of the literature on system identification and regret minimisation [37], [43], [50], [56], [59].

3. There exist real numbers ρ, ν, D, Γ, and ζ such that,

$$\begin{aligned} \rho &= \sup_{\Theta' \in \mathcal{S}} ||A' - B'K(\Theta')|| < 1, \\ \nu &= \sup_{\Theta' \in \mathcal{S}} ||A' - A'L(\Theta')C'|| < 1, \\ D &= \sup_{\Theta' \in \mathcal{S}} ||P(\Theta')||, \\ \Gamma &= \sup_{\Theta' \in \mathcal{S}} ||K(\Theta')||, \\ \zeta &= \sup_{\Theta' \in \mathcal{S}} ||L(\Theta')||. \end{aligned}$$

   The assumptions on ρ and ν are restrictive because they constrain the type of systems on which the proposed exploration strategy can be applied. The assumptions on D, Γ, and ζ are not restrictive because their existence can be ensured given that the set $\mathcal{S}$ consists of system parameters that are controllable and observable. With a similar reasoning, we have $||\mathbf{M}||_{\mathrm{F}} \leq \bar{S}$. That being said, such assumptions can aid in simplifying the stability and the regret analyses [37].

4. It is assumed that $\hat{x}_{0|-1,\hat{\Theta}} = \hat{x}_{0|-1,\Theta} = 0$. Further, the system is assumed to start at the steady state, i.e., $x_0 \sim \mathcal{N}(0, \Sigma)$. At steady state, we have $e_0 \sim \mathcal{N}(0, C\Sigma C^\top + \sigma_z^2 I)$. These assumptions are made to simplify the analysis and to streamline the exposition.

These assumptions hold throughout the thesis.

Chapter 3

Learning-Based Control Strategy

This chapter provides motivation for the proposed LBC algorithm and for the selected system identification technique. Firstly, the system identification technique chosen is based on a recent work in closed-loop subspace system identification [37], as motivated in Section 2-6. The proposed LBC algorithms are extensions to the LQG setting from the previously proposed LBC strategies in the LQR setting, namely, the naive exploration strategy [58], and the IF2E strategy [18].

3-1 System identification

As motivated in Section 2-6, this thesis adopts the closed-loop system identification technique proposed in [37]. This system identification technique can be broadly structured into two sequential phases:

1. Using the predictor form of the state-space equation as described in (2-11), estimate the Markov parameters.
2. Estimating the system parameter Θ from the estimated Markov parameter by using a variant of the subspace system identification technique.

Estimating the Markov parameters

Consider the predictor form of the state-space equation (2-11). Rolling back the evolution of the system H-time steps back, we get

$$\hat{x}_{t|t-1,\Theta} = (A - FC)^H \hat{x}_{t-H|t-H-1,\Theta} + \sum_{k=0}^{H-1} (A - FC)^k \left[Bu_{t-k-1} + Fy_{t-k-1} \right]. \tag{3-1}$$

For the sake of brevity, let $\bar{A} = (A - FC)$. Let us also define the matrices:

$$\mathbf{F} = \begin{bmatrix} CF & C\bar{A}F & \dots & C\bar{A}^{H-1}F \end{bmatrix} \in \mathbb{R}^{n_y \times n_y H},$$
$$\mathbf{G} = \begin{bmatrix} CB & C\bar{A}B & \dots & C\bar{A}^{H-1}B \end{bmatrix} \in \mathbb{R}^{n_y \times n_u H}.$$

Now for $t = H, H+1, \dots, T-1$, we have

$$\begin{aligned} y_t &= C\hat{x}_{t|t-1,\Theta} + e_t \\ &= C\bar{A}^H \hat{x}_{t-H|t-H-1,\Theta} + \sum_{k=0}^{H-1} C\bar{A}^k \left[Bu_{t-k-1} + Fy_{t-k-1}\right] + e_t \\ &= \mathbf{M}\phi_t + e_t + C\bar{A}^H \hat{x}_{t-H|t-H-1,\Theta}, \end{aligned} \tag{3-2}$$

where

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} \mathbf{F} & \mathbf{G} \end{bmatrix} \in \mathbb{R}^{n_y \times (n_y+n_u)H}, \\ \phi_t &= \begin{bmatrix} y_{t-1}^\top & \dots & y_{t-H}^\top & u_{t-1}^\top & \dots & u_{t-H}^\top \end{bmatrix}^\top \in \mathbb{R}^{(n_y+n_u)H}. \end{aligned} \tag{3-3}$$

Since $\bar{A}$ is stable, the last term in (3-2) becomes negligible for large enough H. Specifically, we need

$$H \geq \max \left\{ 2n_x + 1, \frac{\log\left(\sqrt{n_y/\lambda} c_H T^2\right)}{\log(1/\nu)} \right\}, \tag{3-4}$$

which can also be written as $H = \mathcal{O}(\log(T))$ [37]. The expression of the constant c_H can be found in (4-53).

Remark 3 Let the number of time steps in the k^{th} episode be l_k. If the duration of the episode is varied in a doubling fashion, i.e., $l_{k+1} = 2l_k$, then the requirement that $H = \mathcal{O}(\log(T))$ can be relaxed to $H = \mathcal{O}(k_{\text{fin}})$, where k_{fin} is the number of episodes [35].

Now with $\{y_t\}_{t=0}^{\tau-1}$ and $\{u_t\}_{t=0}^{\tau-1}$, define the following matrices:

$$\begin{aligned} Y_{\tau-1} &= \begin{bmatrix} y_H & y_{H+1} & \dots & y_{\tau-1} \end{bmatrix}^\top \in \mathbb{R}^{N \times n_y}, \\ \Phi_{\tau-1} &= \begin{bmatrix} \phi_H & \phi_{H+1} & \dots & \phi_{\tau-1} \end{bmatrix}^\top \in \mathbb{R}^{N \times (n_y+n_u)H}, \\ E_{\tau-1} &= \begin{bmatrix} e_H & e_{H+1} & \dots & e_{\tau-1} \end{bmatrix}^\top \in \mathbb{R}^{N \times n_y}, \\ N_{\tau-1} &= \begin{bmatrix} C\bar{A}^H \hat{x}_{0|-1,\Theta} & C\bar{A}^H \hat{x}_{1|0,\Theta} & \dots & C\bar{A}^H \hat{x}_{\tau-H-1|\tau-H-2,\Theta} \end{bmatrix}^\top \in \mathbb{R}^{N \times n_y}, \end{aligned} \tag{3-5}$$

where $N = (\tau - 1) - H + 1$. Since $N_{\tau-1}$ is negligibly small,

$$Y_{\tau-1} \approx \Phi_{\tau-1}\mathbf{M}^\top + E_{\tau-1}. \tag{3-6}$$

Therefore, from (3-6), the Markov parameters $\mathbf{M}$ can be estimated from the regularised least-squares problem as follows [37]:

$$\begin{aligned} & \hat{\mathbf{M}}^\top = \arg\min_X ||Y_{\tau-1} - \Phi_{\tau-1}X^\top||_{\text{F}}^2 + \lambda ||X||_{\text{F}}^2 \\ \implies & \hat{\mathbf{M}}^\top = (\Phi_{\tau-1}^\top \Phi_{\tau-1} + \lambda I)^{-1} \Phi_{\tau-1}^\top Y_{\tau-1}, \end{aligned} \tag{3-7}$$

where $\lambda > 0$. Therefore, from the predictor form of the system description with H satisfying (3-4), the Markov parameters can be estimated consistently.

Estimating the model parameters Θ

With the estimated Markov parameters at time step t, denoted by $\hat{\mathbf{M}}_{\mathbf{t}}$, the system parameter Θ is estimated from a variant of the Ho-Kalman algorithm [37] (refer to Algorithm 1). Recall that $\mathbf{M} = [\mathbf{F}, \ \mathbf{G}]$. At time step t, we have

$$\hat{\mathbf{M}}_{\mathbf{t}} = \begin{bmatrix} \hat{\mathbf{F}}_{\mathbf{t},\mathbf{1}} & \dots & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{H}} & \hat{\mathbf{G}}_{\mathbf{t},\mathbf{1}} & \dots & \hat{\mathbf{G}}_{\mathbf{t},\mathbf{H}} \end{bmatrix},$$

where $\hat{\mathbf{F}}_{\mathbf{t,i}}$ is the i^{th} $n_y \times n_y$ block of $\hat{\mathbf{F}}_{\mathbf{t}}$, and $\hat{\mathbf{G}}_{\mathbf{t,i}}$ is the i^{th} $n_y \times n_u$ block of $\hat{\mathbf{G}}_{\mathbf{t}}$. Define the Hankel matrix $\mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}}$ as

$$\mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}} := \begin{bmatrix} \hat{\mathbf{F}}_{\mathbf{t},\mathbf{1}} & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{2}} & \dots & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{d_2+1}} \\ \hat{\mathbf{F}}_{\mathbf{t},\mathbf{2}} & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{3}} & \dots & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{d_2+2}} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \hat{\mathbf{F}}_{\mathbf{t},\mathbf{d_1}} & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{d_1+1}} & \dots & \hat{\mathbf{F}}_{\mathbf{t},\mathbf{H}} \end{bmatrix} \in \mathbb{R}^{d_1 n_y \times (d_2+1) n_y}. \tag{3-8}$$

Analogously, $\mathcal{H}_{\hat{\mathbf{G}}_{\mathbf{t}}}$ has a similar definition as above.

Algorithm 1 SYSID [37]

```
Input: $\hat{\mathbf{M}}_{\mathbf{t}}$, $H$, $n_x$, $n_y$, $n_u$, $d_1 \geq n_x$, $d_2 \geq n_x$ such that $d_1 + d_2 + 1 = H$
Construct two Hankel matrices $\mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}} \in \mathbb{R}^{d_1 n_y \times (d_2+1) n_y}$ and $\mathcal{H}_{\hat{\mathbf{G}}_{\mathbf{t}}} \in \mathbb{R}^{d_1 n_y \times (d_2+1) n_u}$ from $\hat{\mathbf{F}}_{\mathbf{t}}$
  and $\hat{\mathbf{G}}_{\mathbf{t}}$ respectively. Let $\hat{\mathcal{H}}_t = \begin{bmatrix} \mathcal{H}_{\hat{\mathbf{F}}_{\mathbf{t}}} & \mathcal{H}_{\hat{\mathbf{G}}_{\mathbf{t}}} \end{bmatrix}$.
Obtain $\hat{\mathcal{H}}_t^-$ by discarding $(d_2+1)^{\text{th}}$ and $(2d_2+2)^{\text{th}}$ block columns of $\hat{\mathcal{H}}_t$.
Perform SVD on $\hat{\mathcal{H}}_t^-$, and then obtain $\hat{\mathcal{N}}_t$, the best $n_x$-rank approximation by setting all
  but the first $n_x$ singular values to zero.
Obtain $\mathbf{U}_{\mathbf{t}}, \mathbf{\Sigma}_{\mathbf{t}}, \mathbf{V}_{\mathbf{t}} = \text{SVD}(\hat{\mathcal{N}}_t)$.
Construct $\mathbf{O}(\hat{\hat{A}}_t, \hat{C}_t, d_1) = \mathbf{U}_{\mathbf{t}} \mathbf{\Sigma}_{\mathbf{t}}^{1/2}$.                ▷ $\hat{\hat{A}}_t = \hat{A}_t - \hat{F}_t \hat{C}_t$
Construct $[\mathbf{C}(\hat{\hat{A}}_t, \hat{F}_t, d_2+1), \mathbf{C}(\hat{\hat{A}}_t, \hat{B}_t, d_2+1)] = \mathbf{\Sigma}_{\mathbf{t}}^{1/2} \mathbf{V}_{\mathbf{t}}$.
Obtain $\hat{C}_t$ from the first $n_y$ rows of $\mathbf{O}(\hat{\hat{A}}_t, \hat{C}_t, d_1)$.
Obtain $\hat{B}_t$ from the first $n_u$ columns of $\mathbf{C}(\hat{\hat{A}}_t, \hat{B}_t, d_2+1)$.
Obtain $\hat{F}_t$ from the first $n_y$ columns of $\mathbf{C}(\hat{\hat{A}}_t, \hat{F}_t, d_2+1)$.
Obtain $\hat{\mathcal{H}}_t^+$ by discarding $1^{\text{st}}$ and $(d_2+2)^{\text{th}}$ block columns of $\hat{\mathcal{H}}_t$.
Obtain $\hat{\hat{A}}_t = \mathbf{O}^\dagger(\hat{\hat{A}}_t, \hat{C}_t, d_1) \ \hat{\mathcal{H}}_t^+ \ [\mathbf{C}(\hat{\hat{A}}_t, \hat{F}_t, d_2+1), \mathbf{C}(\hat{\hat{A}}_t, \hat{B}_t, d_2+1)]^\dagger$.
Obtain $\hat{A}_t = \hat{\hat{A}}_t + \hat{F}_t \hat{C}_t$.
Obtain $\hat{L}_t$ from the first $n_x \times n_y$ block of $\hat{A}_t^\dagger \mathbf{O}^\dagger(\hat{\hat{A}}_t, \hat{C}_t, d_1) \hat{\mathcal{H}}_t^-$.
Return: $\hat{A}_t$, $\hat{B}_t$, $\hat{C}_t$, and $\hat{L}_t$.
```

3-2 Naive exploration-based LBC algorithm

In the present thesis, the focus is on designing an LBC algorithm in the LQG setting that

1. incorporates an LBC policy having a simple structure,
2. provides an effective balance between exploration and exploitation,
3. is computationally efficient, and
4. guarantees the boundedness of the inputs and outputs.

In Section 2-8, it was mentioned that the simple structure of the CEC with ϵ-greedy exploration attains a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ in the LQR setting. The seemingly simple implementation of the CEC is however sensitive to a model mismatch: the controller can only be guaranteed to stabilise the system when the sub-optimality gap (2-15) is small [19], [31], [43]. To circumvent this issue, an initial stabilising controller can be assumed with additive exploration signals to enable sufficiently long exploration before updating the system parameter [58]. Instead of assuming an initial stabilising controller, [36] and [37], which implements an OFU scheme in the LQG setting, incorporates an initial warm-up period (injecting Gaussian input signals) to obtain an initial system parameter estimate such that the corresponding CEC stabilises the true system. Following this warm-up phase, the LBC phase is deployed where system identification and control with the proposed LBC strategy take place sequentially. This modular scheme is considered for the present thesis given the well-established finite-time guarantees on system parameter estimation error with the considered warm-up phase [37], [50], [61]. An auxiliary feature of this modular scheme is that it gives the designer the freedom to choose the type of input signals to provide during the warm-up phase, depending on the application (cf. [67]).

Considering the above argument, the present thesis incorporates a LBC strategy that is deployed in two phases:

1. **Warm-up phase:** Gaussian input signals are injected for T_{w} time steps to provide informative data in order to obtain an initial system parameter estimate such that the corresponding CEC stabilises the true system. The length of this phase depends on how accurate the initial estimate needs to be [37].
2. **LBC phase:** Naive exploration, as briefly discussed in Section 2-8, is deployed in an episodic fashion.

3-2-1 LBC phase

As mentioned previously in Section 2-8, naive exploration strategy, i.e., a CEC with an additive Gaussian input signal whose covariance diminishes at a rate $\mathcal{O}(\frac{1}{\sqrt{t}})$, is sufficient to attain a regret growth rate of $\tilde{\mathcal{O}}(\sqrt{T})$ in the LQR setting. Moreover, this LBC policy has a simple structure and is computationally efficient to deploy, thereby making it a promising candidate. Establishing a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ with this scheme in the LQG setting is however in question. In this thesis, this question is answered in the affirmative in Theorem 4.4 with the LBC policy:

$$\begin{aligned} u_t &= -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t, \\ \eta_t &= \sigma_{\eta_k} r_t, \;\; r_t \sim \mathcal{N}(0, I), \\ \sigma^2_{\eta_k} &= \frac{\gamma}{\sqrt{l_k}}, \gamma > 0, \end{aligned} \tag{3-9}$$

where k is the episode number, and l_k is the number of time steps in the k^{th} episode. Further, $\hat{K}_k$ denotes the optimal feedback gain for the system parameter $\hat{\Theta}_k$. From (3-9), it becomes evident that the covariance of the additive Gaussian exploration signal is kept constant during each episode. This setting is considered to simplify the regret analysis.

Episode length

Following the warm-up phase, the algorithm proceeds in an episodic fashion wherein, the number of time steps l_k of the k^{th} episode satisfies $l_k = 2^k T_{\text{w}}$ for $k = 0, 1, 2, 3..., k_{\text{fin}} - 1$. Since $l_{k+1} = 2l_k$, the number of episodes is approximately $\log_2(T)$. The use of such a 'doubling' episode length scheme can be motivated as follows.

The bulk of the algorithm's regret can be expressed as the sum of the sub-optimality in the control law and the cost associated with the additive exploration signal, as detailed in 4-75:

$$\begin{aligned} R(T) &\lessapprox \sum_{k=0}^{k_{\text{fin}}-1} l_k (J(\hat{\Theta}_k) - J_*) + l_k \sigma^2_{\eta_k} n_u \\ &\lessapprox \sum_{k=0}^{k_{\text{fin}}-1} l_k c_\Theta \left(||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||_{\text{F}} \right)^2 + l_k \sigma^2_{\eta_k} n_u \\ &\lessapprox \sum_{k=0}^{k_{\text{fin}}-1} l_k c_\Theta \frac{1}{l_k} + l_k \frac{1}{\sqrt{l_k}} n_u \\ &\lessapprox k_{\text{fin}} c_\Theta + \sqrt{T} n_u \\ &\approx \log_2(T) c_\Theta + \sqrt{T} n_u. \end{aligned} \tag{3-10}$$

The third inequality is a consequence of Theorem 4.3. From the above exposition, it becomes evident that in order to ensure that the sub-optimality in the control law scales only with $\log_2(T)$, the doubling scheme of the episode length must be adopted. For the detailed treatment of the regret upper bound, please refer to the proof of Theorem 4.4 in Chapter 4.

3-2-2 Algorithm with naive exploration

The LBC algorithm with naive exploration in the LQG setting is given below. The algorithm consists of both the warm-up phase and the LBC phase. Specifically, it shows when the system parameter is updated as well as the corresponding input signals of the warm-up phase and the LBC phase.

Algorithm 2 LQG-NAIVE

```
Initialise $Q, R, \gamma > 0$, $H$, $T_\mathrm{w}$, $n_x$, $n_y$, $n_u$, $\sigma_u^2$, $k_\mathrm{fin}$
procedure WARM-UP                                        ▷ An initial SYS ID phase
    for $t = 0, 1, ..., T_\mathrm{w} - 1$ do
        Inject $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$
    end for
    Store $\{y_t, u_t\}_{t=0}^{T_\mathrm{w}-1}$
end procedure
procedure LEARNING-BASED CONTROL                         ▷ LBC phase
    for $k = 0, 1, ..., k_\mathrm{fin} - 1$ episodes do
        Calculate $\hat{\mathbf{M}}_k$ using $\{y_t, u_t\}_{t=0}^{2^k T_\mathrm{w}-1}$
        Perform SYSID to obtain $\hat{A}_k, \hat{B}_k, \hat{C}_k, \hat{L}_k$
        Determine $\hat{K}_k$ from (2-8)
        Let $l_k = 2^k T_\mathrm{w}$
        for $t = 2^k T_\mathrm{w}, ..., 2^{k+1} T_\mathrm{w} - 1$ do
            Inject $u_t = -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t, \ \eta_t \sim \mathcal{N}(0, \frac{\gamma}{\sqrt{l_k}} I)$
        end for
    end for
end procedure
```

The finite-time guarantees that are critical for establishing a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ are presented in Chapter 4. Further, the simulation results validating the theoretical guarantees are also presented in Chapter 4.

3-3 FIM-based LBC algorithm

It was mentioned in Section 2-8 that experiment design incorporates the intended application into account when designing the input signals. This is achieved by optimising over some function of the FIM (cf. [13], [27], [67]). As promising as this scheme is, it is not suitable for regret minimisation: the cost of deploying these algorithms is not evaluated online. Naive exploration, on the other hand, has a simple and intuitive structure that can be easily deployed for regret minimisation problems but lacks any form of 'intelligence' in designing the exploration signal. That is, the exploration signal which is a Gaussian signal with a diminishing covariance, does not take the application or any feedback from the system into account. There is hence a need to combine the principles of experiment design with the LBC strategies for regret minimisation to 'intelligently' design the exploration signal.

Recently, there have been such efforts in the LQR setting [18], [24], as shown in (2-24). As an extension to this thesis, this LBC strategy which is based on the FIM, is extended to the LQG setting from the LQR setting. The use of FIM is motivated in the following.

3-3-1 FIM

Since learning the system parameters is intimately tied to LBC, understanding the role of the FIM in LBC becomes imperative in designing effective exploration strategies for regret minimisation [27], [52], [70]. The FIM quantifies the amount of information contained in a random variable. For a sequence of random variables $\{V_t\}_{t=0}^{T-1}$, let the joint probability density function be $f_T(\mathbf{v}|\theta) = \prod_{t=0}^{T-1} f(v_i|\theta)$. The FIM $I_T(\theta)$ after T time steps is given by

$$I_T(\theta) = \text{Var}\left(\frac{\partial}{\partial\theta'} l_T(\mathbf{V}|\theta')\right)\Bigg|_{\theta'=\theta}, \tag{3-11}$$

where $l_T(\mathbf{V}|\theta')$ is the log-likelihood of the joint probability density function. Alternatively, the FIM can be defined as follows.

Definition 3.1 [69] For a family of parameterised probability densities $\{p_\theta, \theta \in \bar{\Theta}\}$, $\bar{\Theta} \in \mathbb{R}^d$, FIM $\bar{I}_p(\theta) \in \mathbb{R}^{d\times d}$ is given by

$$\bar{I}_p(\theta) = \int \nabla_\theta \log p_\theta(x) \left(\nabla_\theta \log p_\theta(x)\right)^\top p_\theta(x)dx, \tag{3-12}$$

whenever the integral exists.

Since the FIM quantifies the amount of information contained in the random variable about some parameter θ, one can construct the FIM on the output signals to quantify the amount of the information these signals have on the true system parameter Θ.

The FIM has profound significance in the field of system identification and control. Experiment design methods revolve around the FIM, which was briefly discussed in section 2-8. Many works have used the FIM for control albeit not for exploration: in [5] and [49], FIM is used to decide when the controller must be updated to a new one. Likewise, in [29], the FIM is used to decide when to switch to a known stabilising controller. Finally, the recent work of [18] uses the FIM to explicitly design the exploration signal, the motivation for which is based on the works of [69] and [70], which emphasise the significance of the FIM in regret minimisation in the LQ setting, where it is shown that the optimal policy renders the FIM singular and that the smallest eigenvalue of the FIM is upper bounded by the regret of the corresponding policy. This intuition is used to influence the rate of growth of the regret through the smallest eigenvalue of the FIM, as shown in (2-24) for the LQR setting. In the present thesis, the structure of the LBC policy in (2-24) is extended to the LQG setting by constructing the FIM on the output measurements.

Lemma 3.1 For a partially observable system as defined in (2-1) with the output expressed in an ARX form as shown in (3-6), the FIM under any policy π is given by

$$I_{H,T-1}(\mathbf{M}) = \sum_{t=H}^{T-1} \mathbb{E}_\Theta \left[\phi_t \phi_t^\top \otimes \Sigma_e^{-1}\right], \tag{3-13}$$

where $\Sigma_e = C\Sigma C^\top + \sigma_z^2 I$.

The proof of Lemma 3.1, which can be found in Section 3-4, is an extension to the LQG setting from the earlier result in [70] for the LQR setting.

From (3-13), it becomes evident that the covariates $\phi_t\phi_t^\top$ that directly influence the estimation accuracy of the Markov parameters, as shown in the least-squares formulation in (3-7), appears in the FIM formulation. Hence, the minimum eigenvalue of the FIM can be seen as a metric to measure both the informativity of the input-output signals in terms of system identification as well as a measure of how much information the output measurements have about the true system parameter Θ.

The literature corresponding to the LQR setting indicates that the minimum eigenvalue of the FIM grows at a rate $\mathcal{O}(\sqrt{T})$, which is optimal when for instance, naive exploration is used [24], [29]. Moreover, it has been shown in the LQR setting that, when

$$\begin{aligned} u_t &= -\hat{K}_t x_t + \eta_t, \\ \eta_t &= \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \forall t = 0, 1, 2, ..., T-1, \end{aligned} \tag{3-14}$$

the incurred regret $R(T) = \tilde{\mathcal{O}}(\sqrt{T})$ [43], [58]. It must be noted that the naive exploration strategy satisfies (3-14). Using the result from the optimal growth of $\lambda_{\min}(\text{FIM})$ and the result corresponding to the incurred regret from naive exploration, it is possible to directly incorporate the FIM in designing the additive exploration signal. The work in [18] leveraged this relation to design an LBC as described in (2-24) to guarantee a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$, albeit only asymptotically. This is largely due to the behaviour of $\lambda_{\min}(\text{FIM})$: the work in [18] guarantees that $\lambda_{\min}(\text{FIM})$ can only grow at a rate $\mathcal{O}(\sqrt{T})$ asymptotically.

Therefore, in this thesis, the FIM-based LBC strategy is validated only with empirical simulations, which can be found in Section 4-2. Establishing finite-time guarantees on the regret with this LBC strategy is an interesting direction to pursue for future work.

There is however a caveat in using the FIM: the FIM must be evaluated at the unknown true system parameter Θ, as seen in (3-13). To circumvent this issue, in [18] the FIM is instead evaluated on $\hat{\Theta}_t$, and as the system parameter estimates improve, we have $\hat{\Theta}_t \to \Theta$ in the LQR setting and therefore, the 'estimated' FIM that is used will tend to the true FIM. But, such a guarantee cannot be provided in LQG setting because $\hat{\Theta}_t$ can only converge to some similarity transformation of Θ. That being said, what is of consequence is the behaviour of $\lambda_{\min}(\text{FIM})$ when evaluated at the estimated model parameter since it directly influences the exploration signal. Since the eigenvalues of a matrix are preserved under similarity transformation, one can evaluate the FIM with $\hat{\Theta}_t$.

3-3-2 LBC policy and algorithm

Much like LQG-NAIVE (Algorithm 2), the FIM-based LBC algorithm proceeds in two phases:

1. **Warm-up phase:** Gaussian input signals are injected for T_w time steps to provide informative data in order to obtain an initial system parameter estimate such that the corresponding CEC stabilises the true system. The length of this phase depends on how accurate the initial estimate needs to be [37].
2. **LBC phase:** CEC with additive FIM-based exploration signal, is deployed in an episodic fashion.

The estimated FIM is given by the following expression:

$$I_{H,T-1}(\hat{\mathbf{M}}_\mathbf{t}) = \sum_{t=H}^{T-1} \phi_t \phi_t^\top \otimes \hat{\Sigma}_{e,t}^{-1}, \tag{3-15}$$

where

$$\hat{\Sigma}_{e,t} = \frac{1}{t+1} \sum_{i=0}^{t} \left(y_i - \hat{y}_{i|i-1,\hat{\Theta}_{i-1}}\right)\left(y_i - \hat{y}_{i|i-1,\hat{\Theta}_{i-1}}\right)^\top,$$

where $\hat{y}_{i|i-1,\hat{\Theta}_{i-1}}$ is as defined in (2-11) but for the estimated system parameter $\hat{\Theta}_{i-1}$. The FIM cannot be constructed for the first H-time steps since the ϕ_t vector is defined only after the first H-time steps. This is acceptable because the algorithm, which is based on a similar structure to LQG-NAIVE, also utilises all the previously collected data for system identification. Therefore, during the warm-up phase, sufficient data is collected, i.e., $T_\mathrm{w} \geq H$, to construct the FIM that is to be used in the LBC phase. To ensure that the FIM is not poorly scaled, the exploration strategy in (3-9) is used until $\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_\mathbf{t})\right)$ is greater than some tolerance value. After achieving this minimum scaling, the FIM-based exploration strategy is deployed.

That is, if $\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_\mathbf{t})\right) < c_\mathrm{tol}$,

$$\begin{aligned} u_t &= -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t, \\ \eta_t &= \sigma_{\eta_k} r_t, \;\; r_t \sim \mathcal{N}(0, I), \\ \sigma_{\eta_k}^2 &= \frac{\gamma}{\sqrt{l_k}}, \gamma > 0, \end{aligned}$$

else,

$$\begin{aligned} u_t &= -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t, \\ \eta_t &= \sigma_{\eta_t} r_t, \;\; r_t \sim \mathcal{N}(0, I), \\ \sigma_{\eta_t}^2 &= \frac{\alpha}{\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_\mathbf{t})\right)}, \alpha > 0, \end{aligned} \tag{3-16}$$

where c_{tol} is the tolerance set by the designer. The proposed LBC strategy does not require any optimisation procedure to be performed thereby avoiding any computational costs with respect to computing the exploration signal. The FIM-based LBC algorithm in the LQG setting is given below. The main difference between LQG-NAIVE and LQG-IF2E lies in the exploration signal. LQG-IF2E uses the exploration signal as described in (3-16). The structure of the algorithm, on the other hand, is identical to LQG-NAIVE.

Algorithm 3 LQG-IF2E

```
Initialise $Q, R, \gamma > 0, \alpha > 0,\ H,\ T_{\text{w}},\ n_x,\ n_y,\ n_u,\ \sigma_u^2,\ k_{\text{fin}},\ c_{\text{tol}}$
procedure Warm-up                                          ▷ An initial SYS ID phase
    for $t = 0, 1, ..., T_{\text{w}} - 1$ do
        Inject $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$
    end for
    Store $\{y_t, u_t\}_{t=0}^{T_{\text{w}}-1}$
end procedure
procedure Learning-based control                           ▷ LBC phase
    for $k = 0, 1, ..., k_{\text{fin}} - 1$ episodes do
        Calculate $\hat{\mathbf{M}}_k$ using $\{y_t, u_t\}_{t=0}^{2^k T_{\text{w}}-1}$
        Perform SYSID to obtain $\hat{A}_k, \hat{B}_k, \hat{C}_k, \hat{L}_k$
        Determine $\hat{K}_k$ from (2-8)
        for $t = 2^k T_{\text{w}}, ..., 2^{k+1} T_{\text{w}} - 1$ do
            Compute $I_{H,t}(\hat{\mathbf{M}}_{\mathbf{t}})$ from (3-15)
            if $\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_{\mathbf{t}})\right) < c_{\text{tol}}$ then
                Inject $u_t = -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t,\ \ \eta_t \sim \mathcal{N}(0, \frac{\gamma}{\sqrt{l_k}} I)$
            else
                Inject $u_t = -\hat{K}_k \hat{x}_{t|t,\hat{\Theta}_k} + \eta_t,\ \ \eta_t \sim \mathcal{N}\left(0, \frac{\alpha}{\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_{\mathbf{t}})\right)} I\right)$
            end if
        end for
    end for
end procedure
```

3-4 Proofs

3-4-1 Proof of Lemma 3.1

This proof is an extension from the state measurement case addressed in [69] to the partial observability case. From Lemma B.6, we have for the sequences $\{y_t\}_{t=0}^{T-1}$ and $\{u_t\}_{t=0}^{T-1}$

$$I_{H,T-1}(\mathbf{M}) = \bar{I}_{p\left(\{y_t\}_{t=H}^{T-1}, \{\phi_t\}_{t=H}^{T-1}\right)} = \sum_{t=H}^{T-1} \mathbb{E}\left[\mathcal{L}_t(\mathbf{M})\right],$$

where $p\left(\{y_t\}_{t=H}^{T-1}, \{\phi_t\}_{t=H}^{T-1}\right)$ is a multivariate density function for the sequences $\{y_t\}_{t=H}^{T-1}$ and $\{\phi_t\}_{t=H}^{T-1}$. Further,

$$\mathcal{L}_t(\mathbf{M}) = \int \nabla_{\mathbf{M}} \log p\left(\bar{y} - \mathbf{M}\phi_t - e_t\right)\left(\nabla_{\mathbf{M}} \log p\left(\bar{y} - \mathbf{M}\phi_t - e_t\right)\right)^\top \cdot p\left(\bar{y} - \mathbf{M}\phi_t - e_t\right) d\bar{y},$$

where $\bar{y}$ is a dummy variable for integration. From Lemma B.7, with $y_t = \mathbf{M}\phi_t + e_t$ we have

$$\begin{aligned}\mathcal{L}_t(\mathbf{M}) &= \mathbb{E}_\Theta\left[\left(\left(\mathtt{D}_{\mathbf{M}}\mathbf{M}\right)\phi_t\right)^\top \Sigma_e^{-1}\left(\mathtt{D}_{\mathbf{M}}\mathbf{M}\right)\phi_t\right] \\ &= \mathbb{E}_\Theta\left[\left(\nabla_{\mathbf{M}}\mathtt{vec}(\mathbf{M})\right)^\top\left(\phi_t\phi_t^\top \otimes \Sigma_e^{-1}\right)\nabla_{\mathbf{M}}\mathtt{vec}(\mathbf{M})\right] \\ &= \mathbb{E}_\Theta\left[\phi_t\phi_t^\top \otimes \Sigma_e^{-1}\right].\end{aligned}$$

Then

$$I_{H,T-1}(\mathbf{M}) = \sum_{t=H}^{T-1} \mathbb{E}_\Theta\left[\phi_t\phi_t^\top \otimes \Sigma_e^{-1}\right].$$

Chapter 4

Finite-Time Guarantees and Numerical Analysis

This chapter provides a finite-time regret guarantee when deploying LQG-NAIVE. Several auxiliary results are also provided, which provide support to establish the regret guarantee. Simulation results are provided, which validate the theoretical regret guarantee. As previously mentioned, this chapter also provides numerical simulations for LQG-IF2E.

4-1 Finite-time guarantees

To provide a finite-time regret guarantee when deploying LQG-NAIVE, several auxiliary results are required. The two main ingredients involve guaranteeing the system parameter estimation error to be monotonically decreasing, and the input-output signals of the system to remain bounded during the LBC phase (Lemma 4.1). The first ingredient requires showing that the Markov parameters estimation error is monotonically decreasing (Theorem 4.3), which requires showing that a 'persistence of excitation' condition is satisfied (Lemma 4.2).

Now, since the model parameter estimation error can indeed be shown to decrease monotonically, the corresponding estimation error bound during the LBC period can be upper-bounded by the estimation error bound after the warm-up period. This fact is extensively utilised to simplify the analyses. The following provides a list of the results involved in providing a finite-time regret guarantee:

1. Bounds on the input and output signals after the warm-up phase (Lemma A.1 [36]).
2. Persistence of excitation during the warm-up phase (Lemma A.2 [35]).
3. Bounds on the Markov parameter estimation error after the warm-up phase (Theorem 4.1 [35]).
4. Bounds on the system parameter estimation error (Theorem 4.2 [35]).

5. Bounds on the input and output signals during the LBC phase (Lemma 4.1).

6. Persistence of excitation during the LBC phase (Lemma 4.2).

7. Bounds on the Markov parameter estimation error during the LBC phase (Theorem 4.3).

To facilitate an easier understanding of how the various finite-time guarantees depend on each other, a flowchart is provided below.

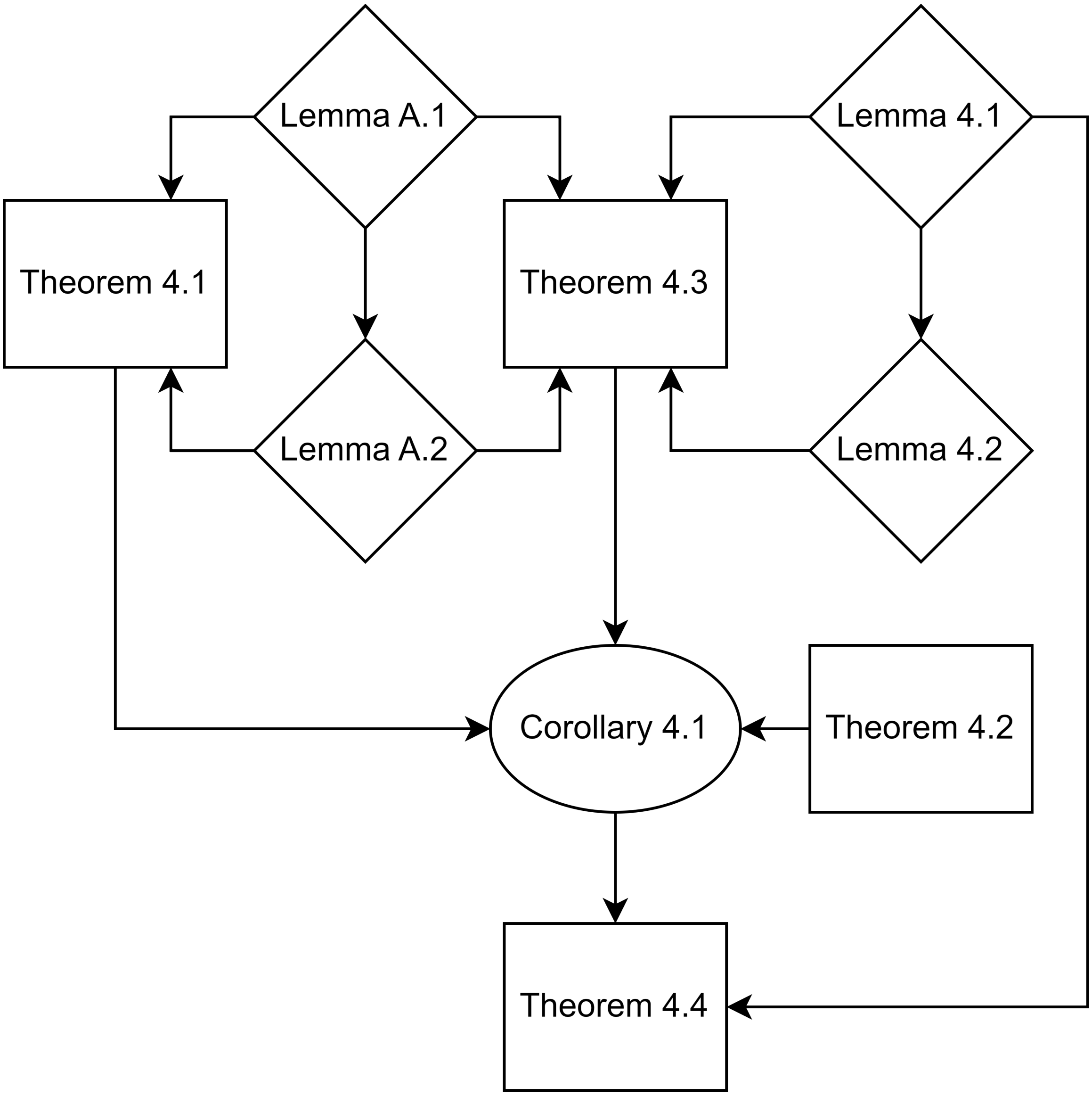


Figure 4-1: A flowchart showing the dependencies involved in providing finite-time guarantees.

For the ease of comprehension, some relevant terms are recalled here

1. From Section 2-9, we have

$$\begin{aligned}\Phi(A) &:= \sup_{\tau\geq 0}\frac{||A^{\tau}||}{\rho(A)^{\tau}},\\ \rho &= \sup_{\Theta'\in\mathcal{S}} ||A' - B'K(\Theta')|| < 1,\\ \nu &= \sup_{\Theta'\in\mathcal{S}} ||A' - A'L(\Theta')C'|| < 1,\\ \Gamma &= \sup_{\Theta'\in\mathcal{S}} ||K(\Theta')||,\\ \zeta &= \sup_{\Theta'\in\mathcal{S}} ||L(\Theta')||.\end{aligned}$$

2. From (3-4), we have

$$H \geq \max\left\{2n_x + 1, \frac{\log\left(\sqrt{n_y/\lambda}c_H T^2\right)}{\log(1/\nu)}\right\},$$

where the exact expression for c_H can be found in (4-53).

3. From (A-3), we have

$$\Upsilon_{\text{w}} = ||C||X_{\text{w}} + Z + U_{\text{w}},$$

where X_{w}, Z, and U_{w} are as defined in (A-2).

4. From (4-46), we have

$$\Upsilon_{\text{ac}} = Y_{\text{ac}} + U_{\text{ac}},$$

where U_{ac} and Y_{ac} are as defined in (4-35) and (4-37) respectively.

5. From (3-7), we have that $\lambda > 0$, which is a regularising parameter in the least-squares formulation.

4-1-1 Warm-up phase

Firstly, we state a result from the literature that provides guarantees on Markov parameters estimation error after the warm-up phase.

Theorem 4.1 [35] The initial estimate of the truncated ARX model, $\hat{\mathbf{M}}_{\mathbf{T}_{\text{w}}}$, obeys the following bound with a probability of at least $1 - 2\delta$ for $\delta \in (0, 1/2)$, after the warm-up period of T_{w} time steps:

$$||\hat{\mathbf{M}}_{\mathbf{T}_{\text{w}}} - \mathbf{M}|| \leq \frac{\text{poly}(n_y, H, n_u)}{\min\{\sigma_w, \sigma_z, \sigma_u\}\sigma_{\text{o}}\sqrt{T_{\text{w}} - H}}, \tag{4-1}$$

for some $\sigma_{\text{o}} > 0$. Specifically, if $T_{\text{w}} \geq T_{\mathbf{M}}$ then

$$||\hat{\mathbf{M}}_{\mathbf{T}_{\text{w}}} - \mathbf{M}|| \leq 1, \tag{4-2}$$

with $T_{\mathbf{M}} = R^2_{\text{warm}}$, where

$$R_{\text{warm}} = \frac{\sqrt{2n_y||C\Sigma C^\top + \sigma_z^2 I|| \left(\log(1/\delta) + \frac{H(n_u+n_y)}{2}\log\left(\frac{\lambda(n_u+n_y)H+\Upsilon_{\text{w}}^2 T_{\text{w}}}{\lambda(n_u+n_y)H}\right)\right)} + \frac{\sqrt{2H}}{T} + \sqrt{2\lambda}\bar{S}}{\sigma_{\text{o}}\min\{\sigma_w, \sigma_z, \sigma_u\}}, \tag{4-3}$$

where $||\mathbf{M}||_{\text{F}} \leq \bar{S}$.

This theorem depends on two supporting results, namely, Lemma A.1 and Lemma A.2. Essentially, Theorem 4.1 implies that after sufficient time steps in the warm-up phase, the estimate of the Markov parameters is quite close to that of the true Markov parameters.

4-1-2 LBC phase

The next step would be to provide guarantees on the system parameter estimation error after the warm-up phase, i.e., an upper bound on $||\hat{\Theta}_{T_{\text{w}}} - \Theta||$. Theorem 4.2 provides such a guarantee by generalising it for any t. This is to avoid redundancy in the results as the following theorem shows that $||\hat{\Theta}_t - \Theta|| = \mathcal{O}(||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||)$.

Theorem 4.2 [35] Let $\mathcal{H} = \begin{bmatrix}\mathcal{H}_{\mathbf{F}} & \mathcal{H}_{\mathbf{G}}\end{bmatrix}$ be the concatenation of two Hankel matrices obtained from $\mathbf{M}$. The notations $\mathcal{H}_{\mathbf{F}}$ and $\mathcal{H}_{\mathbf{G}}$ have analogous expressions to the definition in (3-8) but with the true system parameter. Let $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{L}$ be the similarity transformed system parameters obtained from $\mathbf{M}$ by using Algorithm 1. At time step t, let $\hat{A}_t, \hat{B}_t, \hat{C}_t, \hat{L}_t$ be the estimated system parameters obtained from $\hat{\mathbf{M}}_{\mathbf{t}}$ via Algorithm 1. Then, for a given choice of H satisfying (3-4), there exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n_x \times n_x}$ such that, $\tilde{\Theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{L}) \in (\mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t))$, where

$$\begin{aligned}
\mathcal{C}_A(t) &= \left\{\, A' \in \mathbb{R}^{n_x \times n_x} : ||\hat{A}_t - \mathbf{T}^\top A'\mathbf{T}|| \leq \beta_A(t)\right\}, \\
\mathcal{C}_B(t) &= \left\{\, B' \in \mathbb{R}^{n_x \times n_u} : ||\hat{B}_t - \mathbf{T}^\top B'|| \leq \beta_B(t)\right\}, \\
\mathcal{C}_C(t) &= \left\{\, C' \in \mathbb{R}^{n_y \times n_x} : ||\hat{C}_t - C'\mathbf{T}|| \leq \beta_C(t)\right\}, \\
\mathcal{C}_L(t) &= \left\{\, L' \in \mathbb{R}^{n_u \times n_y} : ||\hat{L}_t - \mathbf{T}^\top L'|| \leq \beta_L(t)\right\},
\end{aligned} \tag{4-4}$$

where

$$\begin{aligned}
\beta_A(t) &= c_A\left(\frac{\sqrt{n_x H}(||\mathcal{H}|| + \sigma_{n_x}(\mathcal{H}))}{\sigma_{n_x}^2(\mathcal{H})}\right)||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||, \\
\beta_B(t) &= \beta_C(t) = \sqrt{\frac{20 n_x H}{\sigma_{n_x}(\mathcal{H})}}||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||, \\
\beta_L(t) &= \frac{c_{L,1}||\mathcal{H}||}{\sqrt{\sigma_{n_x}(\mathcal{H})}}\beta_A(t) + c_{L,2}\frac{\sqrt{n_x H}(||\mathcal{H}|| + \sigma_{n_x}(\mathcal{H}))}{\sigma_{n_x}^{3/2}(\mathcal{H})}||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||,
\end{aligned} \tag{4-5}$$

for some problem-dependent constants $c_A, c_{L,1}$ and $c_{L,2}$.

Problem-dependent constants imply constants that depend on the true system parameter Θ. The proof is provided in Appendix A-2 for the sake of completeness. During the LBC phase, it is imperative to guarantee that the input and output signals remain bounded to ensure the safe operation of the closed-loop system. Such a guarantee can be provided with LQG-NAIVE, as shown in the following lemma. Both Lemma 4.1 and Lemma 4.2 are extensions of the results in [35]. The extension here requires accounting for the additive Gaussian excitation signals.

Lemma 4.1 After a warm-up period of T_w time steps, LQG-NAIVE satisfies the following with a probability of at least $1-\delta$ for $\delta \in (0,1)$: for all $t \in [T_\mathrm{w}, T-1]$,

1. $||\hat{x}_{t|t,\hat{\Theta}_t}|| \leq \bar{\chi}$,
2. $||\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}|| \leq X_{\mathrm{est,ac}}$,
3. $||y_t|| \leq Y_{\mathrm{ac}}$,
4. $||u_t|| \leq U_{\mathrm{ac}}$,

where $\bar{\chi}$, $X_{\mathrm{est,ac}}$, U_{ac}, and Y_{ac} are as defined in (4-33), (4-34), (4-35), and (4-37) respectively.

The proof of this lemma can be found in Section 4-3-1. It is important to show that the system parameter estimation error is monotonically decreasing in the LBC phase. A critical piece to ensure that lies in guaranteeing the persistence of excitation, which ensures the estimation accuracy of the Markov parameters. Essentially, the persistence of the excitation ensures that the cumulative sum of the covariates $\left(\sum_{i=T_\mathrm{w}}^{t-1} \phi_i \phi_i^\top\right)$ is positive definite. The significance of the positive definiteness of the covariates becomes evident in the least-squares formulation (3-7).

Lemma 4.2 After T_{ac} time steps in the LBC period and for some $\sigma_\mathrm{c} > 0$, with probability of at least $1-\delta$ for $\delta \in (0,1)$, we have the following for all $t \geq T_{\mathrm{ac}} + T_\mathrm{w}$,

$$\sigma_{\min}\left(\sum_{i=T_\mathrm{w}}^{t-1} \phi_i \phi_i^\top\right) \geq (t - T_\mathrm{w}) \frac{\sigma_\mathrm{c}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta_{t-1}}^2\}}{8}, \tag{4-6}$$

where

$$T_{\mathrm{ac}} = \frac{512 \Upsilon_{\mathrm{ac}}^4 H^2 \log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_\mathrm{c}^4 \min\{\sigma_w^4, \sigma_z^4, \sigma_{\eta_{t-1}}^4\}}. \tag{4-7}$$

A critical component of this proof is the 'truncated closed-loop noise evolution parameter' denoted by $\mathcal{G}_t^{\mathrm{cl}}$, which captures how the noise sequences influence the trajectory of the system. The proof of this lemma is deferred to Section 4-3-2.

Remark 4 An integral piece for proving Lemma 4.2 requires a perturbation bound on the 'truncated closed-loop noise evolution parameter' of the form, $||\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}}||$, where $\mathcal{G}^{\mathrm{cl}}$ is the 'truncated closed-loop noise evolution parameter' with the true system parameter Θ. This perturbation bound requires representing the bound on $||\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}}||$ as a function of the system parameter estimation error, and a detailed derivation of the perturbation bound is not provided in this thesis due to lack of time and hence it will be consolidated in the future work. Therefore, in this thesis, it is assumed that $||\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}}||$ can indeed be represented as a function of the system parameter estimation error. This assumption is not far-fetched when considering the form of $\mathcal{G}_t^{\mathrm{cl}}$, as shown in (4-41). Further, the proof of Lemma 3.2 in [35] on which Lemma 4.2 is based, also fails to provide sufficient details for deriving the above-mentioned perturbation bound.

With Lemma 4.1 and Lemma 4.2, it is possible now to provide a bound on the estimation error of the Markov parameters during the LBC phase.

Theorem 4.3 For any time step $t \geq \max\{T_{\mathrm{ac}} + T_{\mathrm{w}}, 2T_{\mathrm{w}} - 1\}$, the estimate of the truncated ARX model, $\hat{\mathbf{M}}_{\mathbf{t}}$, obeys the following bound with a probability of at least $1-2\delta$ for $\delta \in (0, 1/2)$:

$$||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}|| \leq \frac{\mathrm{poly}(n_y, H, n_u)}{\sqrt{(t - H + 1)\min\{\sigma_{\mathrm{o}}^2\sigma_w^2, \sigma_{\mathrm{o}}^2\sigma_z^2, \sigma_{\mathrm{o}}^2\sigma_u^2, \frac{\sigma_{\mathrm{c}}^2\sigma_w^2}{8}, \frac{\sigma_{\mathrm{c}}^2\sigma_z^2}{8}, \frac{\sigma_{\mathrm{c}}^2\sigma_{\eta_t}^2}{8}\}}}. \tag{4-8}$$

The proof of this theorem can be found in Section 4-3-3. The structure of the proof follows similarly to the proof of Theorem 3.2 in [35]. The difference lies in the terms that account for the additive excitation signal. Theorem 4.1 along with the result in Theorem 4.3 implies that the estimation error of Markov parameters is monotonically decreasing approximately at a rate $\mathcal{O}(\frac{1}{\sqrt{t}})$. This fact is leveraged in the following corollary to provide a bound for the model parameter estimation error at time step t with respect to the bound on the model parameter estimation error after the warm-up phase.

Corollary 4.1 [35] After the warm-up period of T_{w} time steps, for a given choice of H satisfying (3-4), for a unitary matrix $\mathbf{T} \in \mathbb{R}^{n_x \times n_x}$, with a probability of at least $1 - 2\delta$ for $\delta \in (0, 1/2)$, we have

$$\begin{aligned}
\mathcal{C}_A(T_{\mathrm{w}}) &= \left\{ A' \in \mathbb{R}^{n_x \times n_x} : ||\hat{A}_t - \mathbf{T}^\top A'\mathbf{T}|| \leq \beta_A(T_{\mathrm{w}}) \right\}, \\
\mathcal{C}_B(T_{\mathrm{w}}) &= \left\{ B' \in \mathbb{R}^{n_x \times n_u} : ||\hat{B}_t - \mathbf{T}^\top B'|| \leq \beta_B(T_{\mathrm{w}}) \right\}, \\
\mathcal{C}_C(T_{\mathrm{w}}) &= \left\{ C' \in \mathbb{R}^{n_y \times n_x} : ||\hat{C}_t - C'\mathbf{T}|| \leq \beta_C(T_{\mathrm{w}}) \right\}, \\
\mathcal{C}_L(T_{\mathrm{w}}) &= \left\{ L' \in \mathbb{R}^{n_u \times n_y} : ||\hat{L}_t - \mathbf{T}^\top L'|| \leq \beta_L(T_{\mathrm{w}}) \right\},
\end{aligned} \tag{4-9}$$

where

$$\begin{aligned}
\beta_A(T_{\mathrm{w}}) &= \frac{\sigma_{n_x}(A)}{2} \text{ if } T_{\mathrm{w}} \geq T_A, \\
\beta_B(T_{\mathrm{w}}) &= \beta_C(T_{\mathrm{w}}) = 1 \text{ if } T_{\mathrm{w}} \geq T_B, \\
\beta_L(T_{\mathrm{w}}) &= \frac{c_{L,1}||\mathcal{H}||}{\sqrt{\sigma_{n_x}(\mathcal{H})}}\beta_A(T_{\mathrm{w}}) + c_{L,2}\frac{\sqrt{n_x H}(||\mathcal{H}|| + \sigma_{n_x}(\mathcal{H}))}{\sigma_{n_x}^{3/2}(\mathcal{H})} \text{ if } T_{\mathrm{w}} \geq T_A.
\end{aligned} \tag{4-10}$$

This corollary is a consequence of Theorem 4.2 and Theorem 4.3, and is from the literature as cited above. The proof of this corollary along with the definitions of T_A and T_B are presented for the sake of completeness in Appendix A-2, along with the proof of Theorem 4.2.

With the boundedness of the input and output signals and with the guarantee of diminishing model parameter estimation error, we are almost ready to address the regret bound. The final piece in establishing the regret upper bound requires bounding the sub-optimality gap $\Delta_{\hat{\Theta}}$ as defined in (2-15). This inherently requires a way to represent the (sub)optimal long-term average expected cost incurred during the LBC phase denoted by $J(\hat{\Theta}_t)$. The following exposition allows us to do just that by representing the (sub)optimal long-term average expected cost as a function of the solution to a Lyapunov equation.

For a system as defined in (2-1) satisfying the assumptions in Assumptions 2.1, and for an estimated system parameter $\hat{\Theta} \in \mathcal{S}$ with the set $\mathcal{S}$ as defined in Section 2-9, define an alternative formulation of the LQG cost function as follows:

$$J_s(\hat{\Theta}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}} \end{bmatrix}^\top \underbrace{\begin{bmatrix} Q_c & 0 \\ 0 & \hat{K}^\top R\hat{K} \end{bmatrix}}_{\bar{\mathbf{W}}} \begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}} \end{bmatrix}\right] \quad \text{s.t.}$$

$$\begin{aligned} & x_{t+1} = Ax_t + Bu_t + w_t, \;\; w_t \sim \mathcal{N}(0, \sigma_w^2 I), \\ & y_t = Cx_t + z_t, \;\; z_t \sim \mathcal{N}(0, \sigma_z^2 I), \\ & \hat{x}_{t|t,\hat{\Theta}} = (I - \hat{L}\hat{C})\hat{x}_{t|t-1,\hat{\Theta}} + \hat{L}y_t, \\ & \hat{x}_{t+1|t,\hat{\Theta}} = \hat{A}\hat{x}_{t|t,\hat{\Theta}} + \hat{B}u_t, \\ & u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}}, \end{aligned} \tag{4-11}$$

where $Q_c = C^\top QC$, $\hat{K}$ stabilises the true system, and $\hat{A} - \hat{F}\hat{C}$ is asymptotically stable. Further, consider the following closed-loop state-space equation with extended states:

$$\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}} \end{bmatrix} = \underbrace{\begin{bmatrix} A & -B\hat{K} \\ \hat{L}CA & \left(I - \hat{L}\hat{C}\right)\left(\hat{A} - \hat{B}\hat{K}\right) - \hat{L}CB\hat{K} \end{bmatrix}}_{\hat{\mathbf{G}}_1} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}} \end{bmatrix} + \underbrace{\begin{bmatrix} I & 0 \\ \hat{L}C & \hat{L} \end{bmatrix}}_{\hat{\mathbf{G}}_2} \begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix},$$

and let $S = \text{dlyap}(\hat{\mathbf{G}}_1, \bar{\mathbf{W}}) \geq 0$ be the solution of the discrete Lyapunov equation, as defined in Definition B.1. Then, we have

$$J_s(\hat{\Theta}) = \text{Tr}\left(\hat{\mathbf{G}}_2^\top S \hat{\mathbf{G}}_2 \begin{bmatrix} \sigma_w^2 I & 0 \\ 0 & \sigma_z^2 I \end{bmatrix}\right). \tag{4-12}$$

The proof of (4-12) is presented for the sake of completeness in Appendix A-3. Now, it is a well-known fact that $J_s(\hat{\Theta}) = J(\hat{\Theta}) - \text{Tr}(Q\sigma_z^2 I)$ [43]. This property can aid in quantifying the sub-optimality gap $\Delta_{\hat{\Theta}}$ as shown in the proof of Theorem 3 in [43]. Now, we are ready to state the regret upper bound.

Theorem 4.4 (The regret of the LBC phase) After the warm-up period of T_w time steps, with a probability of at least $1 - \delta$ for $\delta \in (0, 1)$, we have for any T in the LBC phase, the regret of LQG-NAIVE is bounded as follows:

$$R(T) \lesssim \sum_{k=0}^{k_{\text{fin}}-1} l_k \left(J_s(\hat{\Theta}_k) - J_*\right) + l_k n_y \sigma_z^2 \text{Tr}\,(Q) + l_k \sigma_{\eta_k}^2 \text{poly}\,(\beta_A(T_\text{w}), \beta_B(T_\text{w}), \beta_L(T_\text{w}), ||S||)$$
$$+ \sqrt{l_k}\text{poly}\,(\beta_A(T_\text{w}), \beta_B(T_\text{w}), \beta_L(T_\text{w}), ||S||, X_\text{ac}, \bar{\chi}, ||Q||, ||R||, \Gamma)\,, \tag{4-13}$$

where l_k is the number of time steps in the k^th episode and k_fin is the number of episodes. The above bound can be refined to obtain:

$$R(T) = \tilde{\mathcal{O}}(\sqrt{T}), \tag{4-14}$$

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

Problem-dependent constants imply constants that depend on the true system parameter Θ. The proof of this theorem is deferred to Section 4-3-4. Here, an intuition is provided on how the regret bound is derived. From Algorithm 2, it becomes evident that LQG-NAIVE operates in an episodic fashion. Hence, the regret is also analysed episode-wise. Firstly, an upper bound on the cumulative cost incurred by LQG-NAIVE in any episode is obtained. From this, we can obtain an upper bound on the regret for any episode. This episode-wise regret bound is then summed over the number of episodes to obtain the final regret upper bound incurred by LQG-NAIVE during the LBC phase as shown in (4-13). From (4-13), we see that the sub-optimality gap and the exploration cost (3^rd term) have significant contributions towards the magnitude of the regret since they are linearly dependent on the number of time steps in the k^th episode. To provide a bound on the sub-optimality gap, we use an earlier result in [43], which essentially reduces the contribution from the sub-optimality gap to $\mathcal{O}(\log_2(T))$. On the other hand, the exploration cost essentially reduces to $\mathcal{O}(\sqrt{T})$ since $\sigma_{\eta_k}^2 = \mathcal{O}(1/\sqrt{l_k})$. We can further see that the system parameter estimation error along with the established bounds on the state and its estimate, also influences the regret upper bound. As was mentioned before, the system parameter estimation errors are monotonically decreasing, and as a consequence, it is possible to upper bound the system parameter estimation error at any time step t with the corresponding bound after the warm-up. This is evident in (4-13). Finally, it must be noted that the regret incurred during the warm-up phase is $\tilde{\mathcal{O}}(T_\text{w})$ [36]. This result along with the result in Theorem 4.4 gives the overall regret incurred by LQG-NAIVE.

4-2 Simulation results

4-2-1 Simulation setting

Here, we validate the performance of LQG-NAIVE and LQG-IF2E through empirical simulations. For the simulation, the web server control problem is considered, which is linearised around its operating point [1], [9]. For more details about this control problem, refer to Section 3.4 in [9]. This problem, which is formulated for the full-state measurement case, is modified to the partial observability case, i.e., the inclusion of the C matrix and the measurement noise. The system under consideration is given by

$$\begin{aligned} x_{t+1} &= \begin{bmatrix} 0.54 & -0.11 \\ -0.026 & 0.63 \end{bmatrix} x_t + \begin{bmatrix} -85 & 4.4 \\ -2.5 & 2.8 \end{bmatrix} u_t + w_t, \;\; w_t \sim \mathcal{N}(0, 0.01I), \\ y_t &= \begin{bmatrix} 0.2 & 0.3 \\ 0.3 & 0.2 \end{bmatrix} x_t + z_t, \;\; z_t \sim \mathcal{N}(0, 0.01I). \end{aligned}$$

The cost matrices for the control problem are given by [9]:

$$Q = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} \frac{1}{50^2} & 0 \\ 0 & \frac{1}{10^6} \end{bmatrix}.$$

The optimal long-term average expected cost calculated from (2-12) is 0.0707046. The length of the warm-up phase is set to $T_\mathrm{w} = 25$. During the warm-up phase, Gaussian excitatory signals are injected, where $u_t \sim \mathcal{N}(0, 0.1I)$. For the LBC phase, the number of episodes is taken to be 11. The hyper-parameters for the LBC policies (3-9) and (3-16) are $\gamma = \frac{\sqrt{T_\mathrm{w}}}{10}$ and $\alpha = 1$ respectively. To ensure proper scaling of the minimum eigenvalue of the FIM, $c_\mathrm{tol} = 1$. Finally, to obtain consistent parameter estimates, the length of the input-output data history that is used to construct the ϕ vector is set at $H = 12$. The algorithms LQG-NAIVE and LQG-IF2E are run 100 times to report the mean and the standard deviation of the observations.

4-2-2 Comparing the simulation results of LQG-NAIVE and LQG-IF2E

Figure 4-2 captures how the regret growth varies over the 100 simulations. The bold red line signifies the mean regret of LQG-NAIVE whereas, the bold blue line signifies the mean regret of LQG-IF2E. LQG-NAIVE incurs a long-term average cost of 0.074426 and LQG-IF2E incurs a long-term average cost of 0.074203, averaged over the 100 simulations. The LQG-IF2E algorithm switches to the FIM-based exploration strategy at the 35th time step, on average. This means that with a delay of approximately one episode, the algorithm is able to deploy the FIM-based exploration strategy. From Figure 4-2, it becomes evident that LQG-NAIVE and LQG-IF2E have similar behaviour of the regret growth, this is primarily due to the hyper-parameter tuning. An intuitive way to understand this similarity in regret growth is to plot the evolution of the minimum eigenvalue of the FIM.

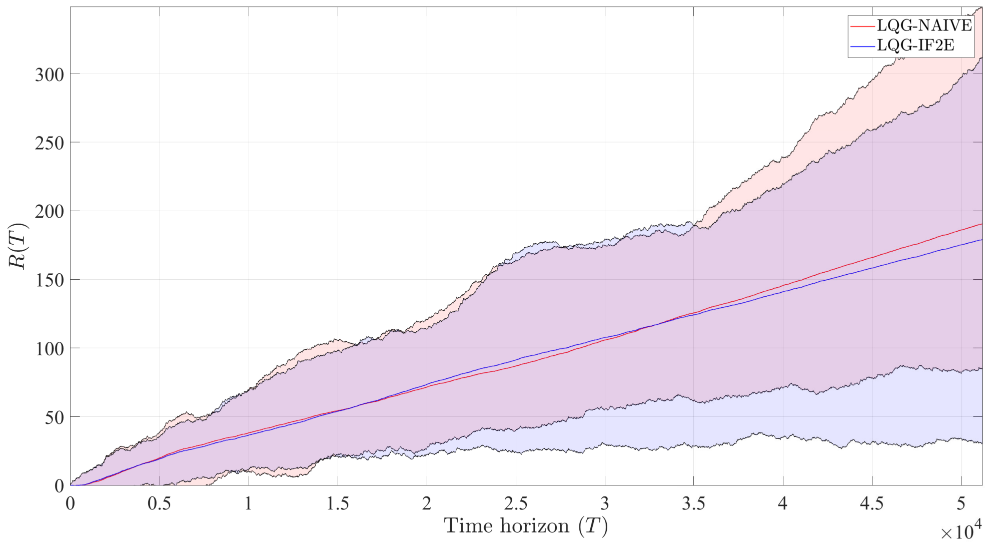


Figure 4-2: Regret growth of LQG-NAIVE and LQG-IF2E.

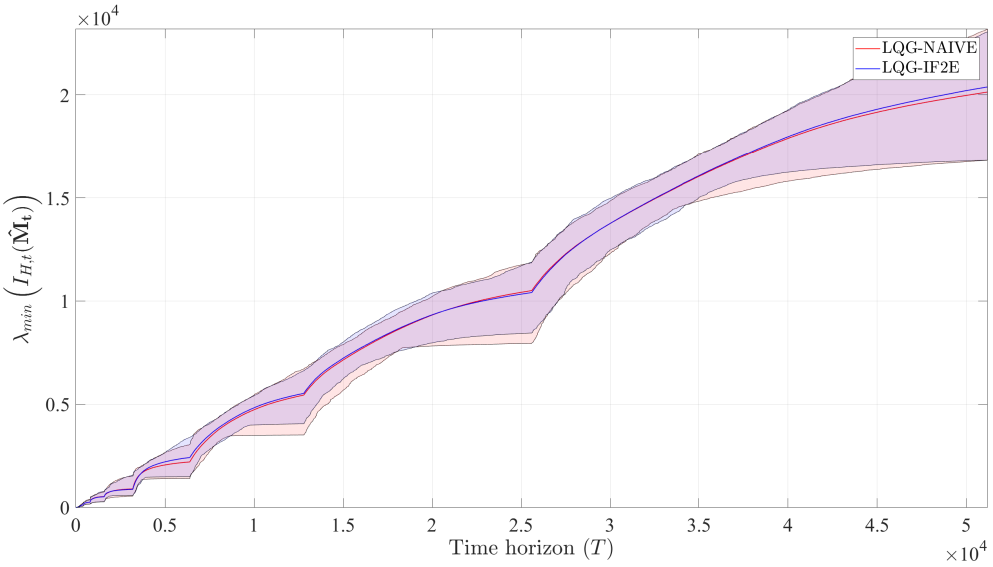


Figure 4-3: Growth of the minimum eigenvalue of the FIM estimated on $\hat{\mathbf{M}}_\mathbf{t}$.

Figure 4-3 captures how the minimum eigenvalue of the FIM varies over the 100 simulations. The bold blue line signifies the mean growth of $\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_\mathbf{t})\right)$ of LQG-IF2E whereas, the bold red line signifies the mean growth of $\lambda_{\min}\left(I_{H,t}(\hat{\mathbf{M}}_\mathbf{t})\right)$ of LQG-NAIVE. From the figure, it becomes evident that the behaviour of the FIM is also similar between the two algorithms. Since the FIM captures the informativity of the data, one can expect two algorithms to have similar regret growth if their corresponding FIMs constructed on the input-output data, also show a similar growth. The 'bumps' that are observed in Figure 4-3 closely correspond to the time steps where the system parameter estimate $\hat{\Theta}_t$ was updated. It is straightforward to notice that the length of the 'bumps' corresponds approximately to the length of the episodes.

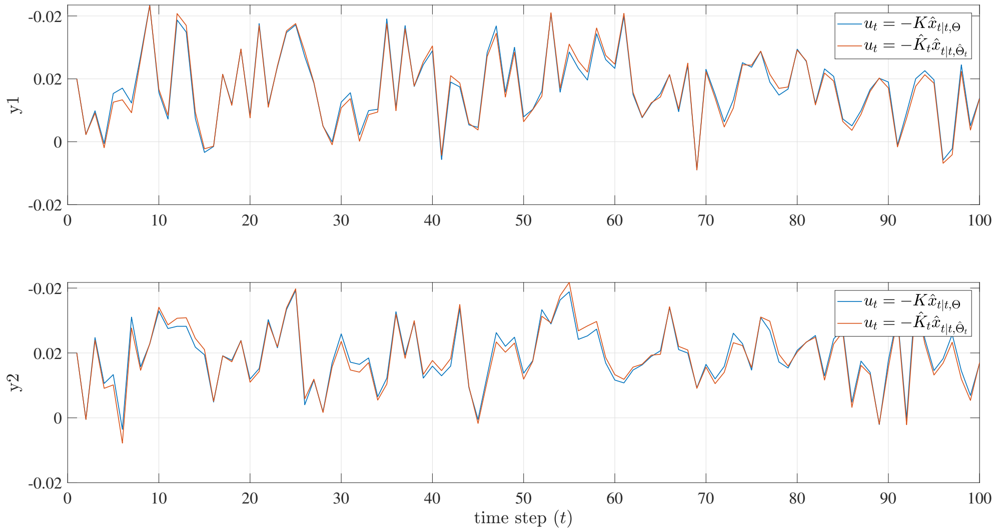


Figure 4-4: Comparing the CEC of the last updated model with that of the optimal policy: LQG-NAIVE.

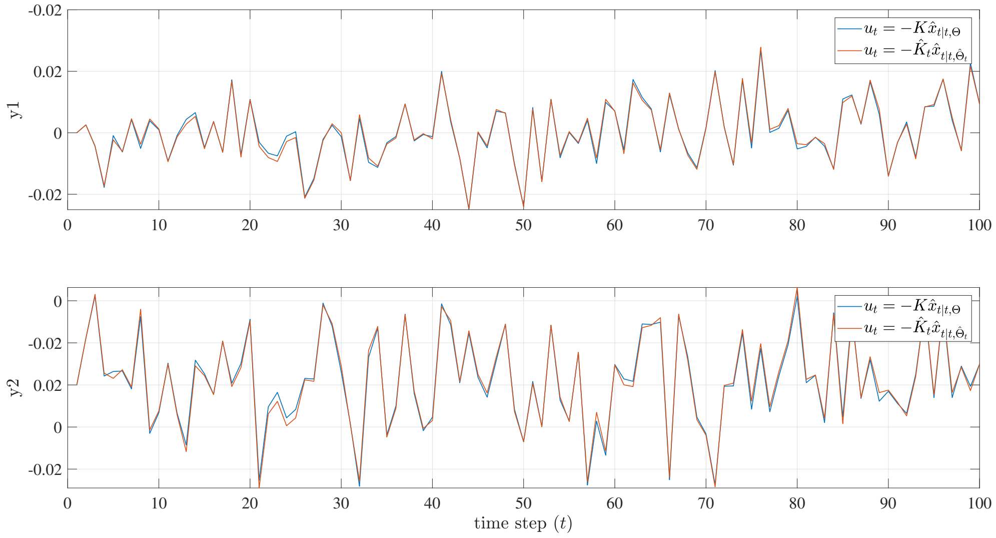


Figure 4-5: Comparing the CEC of the last updated model with that of the optimal policy: LQG-IF2E.

To validate the claim that the LBC policy is converging to that of the optimal policy, we can compare the CEC controller, i.e., $u_t = -\hat{K}_{11}\hat{x}_{t|t,\hat{\Theta}_{11}}$ that is updated at the start of the last (11^{th}) episode, with that of the optimal controller. For LQG-NAIVE, we have Figure 4-4, which compares the *average* behaviour of the CEC controller of the 'latest' model with that of the optimal controller for 100 time steps. Here, the average is taken over the 100 simulations. From the figure, it becomes evident that the estimated control policy has converged quite close to that of the optimal policy. Comparing the output data, we have a Root Mean Squared Error (RMSE) of 14×10^{-4} and 16×10^{-4}, and a Variance Accounted For (VAF) of 96.56% and 94.52%, for output channel 1 and 2 respectively. Similarly, for LQG-IF2E, we have Figure 4-5, which shows that the output behaviour of the two controllers is quite close. Comparing the output data, we have an RMSE of 7.87×10^{-4} and 9.36×10^{-4}, and a VAF of 99.72% and 98.42%, for output channel 1 and 2 respectively.

4-3 Proofs

4-3-1 Proof of Lemma 4.1

The present analysis abstracts away the episodic behaviour of the algorithm, that is, $\hat{\Theta}_{t-1}$ could either be the system parameter estimate being used in the current episode at time step $t-1$ or it could be the system parameter estimated in the previous episode if the time step $t-1$ falls in the previous episode. The analysis is independent of when the system parameter is being updated.

Since the behaviour of a system with parameter Θ and its similarity transformation is the same, without loss of generality, we assume that the similarity transformation matrix $\mathbf{T} = I$. This proof is an extension of an earlier result in [35]. The cited parts in this proof can be found in the proof of Lemma 4.1 in [35].

Based on (2-6), consider the following decomposition of $\hat{x}_{t|t,\hat{\Theta}_t}$

$$
\begin{aligned}
\hat{x}_{t|t,\hat{\Theta}_t} &= \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + \hat{L}_t(y_t - \hat{C}_t\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}) \\
&= \hat{A}_{t-1}\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} - \hat{B}_{t-1}\hat{K}_{t-1}\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} \\
&\quad + \hat{B}_{t-1}\eta_{t-1} + \hat{L}_t\Bigg(y_t - \hat{C}_t\left(\hat{A}_{t-1}\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} - \hat{B}_{t-1}\hat{K}_{t-1}\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1}\right)\Bigg) \\
&= \left(I - \hat{L}_t\hat{C}_t\right)\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \left(I - \hat{L}_t\hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1} \\
&\quad + \hat{L}_t\left(Cx_t + z_t - C\hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + C\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right) \\
&= \left(I - \hat{L}_t\hat{C}_t\right)\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \left(I - \hat{L}_t\hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1} \\
&\quad + \hat{L}_t\left(Cx_t - C\hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + C\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + C\hat{B}_{t-1}\eta_{t-1} + z_t\right) \\
&= \Bigg(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1} - \hat{L}_t\left(\hat{C}_t\hat{A}_{t-1} - \hat{C}_t\hat{B}_{t-1}\hat{K}_{t-1} - C\hat{A}_{t-1} + C\hat{B}_{t-1}\hat{K}_{t-1}\right)\Bigg)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} \\
&\quad + \left(I - \hat{L}_t\hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1} + \hat{L}_tC\left(x_t - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\Theta}\right) \\
&\quad + \hat{L}_tC\hat{B}_{t-1}\eta_{t-1} + \hat{L}_tz_t
\end{aligned}
$$

$$= \underbrace{\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1} - \hat{L}_t\left(\hat{C}_t\hat{A}_{t-1} - \hat{C}_t\hat{B}_{t-1}\hat{K}_{t-1} - C\hat{A}_{t-1} + C\hat{B}_{t-1}\hat{K}_{t-1}\right)\right)}_{\text{Can be thought of as the dynamics}} \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}}$$
$$+ \underbrace{\hat{L}_tC\left(x_t - \hat{x}_{t|t-1,\Theta}\right) + \hat{L}_tC\left(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right) + \hat{B}_{t-1}\eta_{t-1} + \hat{L}_tz_t + \hat{L}_t\left(C - \hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1}}_{\text{Can be thought of as a process noise}}. \tag{4-15}$$

We will bound $||\hat{x}_{t|t,\hat{\Theta}_t}||$ by bounding each of the above terms in the decomposition. Define the following event:

$$\mathcal{E}_{\mathbf{M}} := \left\{||\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}|| \leq 1\right\}, \tag{4-16}$$

which is assumed to hold when $t \geq T_\text{w} \geq T_\mathbf{M}$, where $T_\mathbf{M}$ is as defined in Theorem 4.1. It will be shown later in Theorem 4.3 that this event indeed holds with high probability. Assuming this event holds, we have

1. $\Theta \in \mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t)$ for all $t \geq T_\text{w}$.
2. $||\hat{C}_t - C||, ||\hat{B}_t - B||, ||\hat{F}_t - F|| \leq \beta_B(T_\text{w}) = 1$ when $T_\text{w} \geq T_B$.
3. $\left|\left|\hat{A}_t - A\right|\right| \leq \beta_A(T_\text{w}) = \sigma_{n_x}(A)/2$ when $T_\text{w} \geq T_A$.
4. $\left|\left|\hat{L}_t - L\right|\right| \leq \beta_L(T_\text{w})$,

where the similarity transformation matrix $\mathbf{T} = I$ without loss of generality. We will use the above bounds extensively in the current proof. The definitions of T_A and T_B can be found in Appendix A-2.

Bounding the norm of the 'dynamics' term [35]

Let

$$N_t = \left(I - \hat{L}_t\left(\hat{C}_t - C\right)\right)\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right).$$

Let $T_u = T_B\left(\frac{2\zeta\rho}{1-\rho}\right)^2$. This implies $\left|\left|\hat{C}_t - C\right|\right| \leq \frac{1-\rho}{2\zeta\rho}$ for $t \geq T_\text{w} \geq T_u$. Then under event $\mathcal{E}_\mathbf{M}$, we have

$$||N_t|| \leq \frac{1+\rho}{2} < 1. \tag{4-17}$$

Recalling from Section 2-9, we have

$$\begin{aligned}
\Phi(A) &:= \text{sup}_{\tau\geq 0}\frac{||A^\tau||}{\rho(A)^\tau},\\
\rho &= \sup_{\Theta'\in\mathcal{S}} ||A' - B'K(\Theta')|| < 1,\\
\nu &= \sup_{\Theta'\in\mathcal{S}} ||A' - A'L(\Theta')C'|| < 1,\\
\Gamma &= \sup_{\Theta'\in\mathcal{S}} ||K(\Theta')||,\\
\zeta &= \sup_{\Theta'\in\mathcal{S}} ||L(\Theta')||.
\end{aligned}$$

Bounding the norm of the 'process noise' term

The term $\hat{L}_t C\left(x_t - \hat{x}_{t|t-1,\Theta}\right) + \hat{L}_t z_t$ is a $\zeta\left(||C||||\Sigma||^{1/2} + \sigma_z\right)$ - sub-Gaussian random variable. Therefore, from Lemma B.1, we get [36]:

$$\left|\left|\hat{L}_t C\left(x_t - \hat{x}_{t|t-1,\Theta}\right) + \hat{L}_t z_t\right|\right| \leq \zeta\left(||C||||\Sigma||^{1/2} + \sigma_z\right)\sqrt{2n_x \log\left(\frac{2n_x T}{\delta}\right)}, \tag{4-18}$$

for all $t \geq T_\text{w}$ with probability of at least $1 - \delta/T$. Re-parameterise $\delta/T \to \delta$. Now, under $\mathcal{E}_\mathbf{M}$, we have

$$\begin{aligned}
&\left|\left|\hat{B}_{t-1}\eta_{t-1} + \hat{L}_t\left(C - \hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1}\right|\right| \\
&\leq \left(||B|| + \left|\left|\hat{B}_{t-1} - B\right|\right|\right)||\eta_{t-1}||\left(1 + \left|\left|\hat{L}_t(\hat{C}_t - C)\right|\right|\right) \\
&\leq \left(||B|| + 1\right)||\eta_{t-1}||\left(1 + \zeta\right).
\end{aligned}$$

Recall from (3-9) that $\eta_t \sim \sigma_{\eta_t}\mathcal{N}(0, I)$, where $\sigma_{\eta_t}^2 = \frac{\gamma}{\sqrt{l_k}}$ with $\gamma > 0$. Therefore, from Lemma B.1, we have

$$||\eta_t|| \leq \sigma_{\eta_t}\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)}, \tag{4-19}$$

which holds with a probability of at least $1 - \delta/T$ for all $t \geq T_\text{w}$. Re-parameterise $\delta/T \to \delta$. This implies

$$\begin{aligned}
&\left|\left|\hat{B}_{t-1}\eta_{t-1} + \hat{L}_t\left(C - \hat{C}_t\right)\hat{B}_{t-1}\eta_{t-1}\right|\right| \\
&\leq \sigma_{\eta_t}\left(||B|| + 1\right)\left(1 + \zeta\right)\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)} \\
&\leq \sqrt{\gamma}\left(||B|| + 1\right)\left(1 + \zeta\right)\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)},
\end{aligned} \tag{4-20}$$

which holds with a probability of at least $1 - \delta$ for all $t \geq T_\text{w}$ under the event $\mathcal{E}_\mathbf{M}$. Finally, we bound the spectral norm of the term $\Delta_t = \left(\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right)$.

Bounding $||\Delta_t||$

From (2-6), we have the following decompositions:

$$\begin{aligned}
\hat{x}_{t+1|t,\Theta} &= A\hat{x}_{t|t,\Theta} + Bu_t \\
&= A\hat{x}_{t|t,\Theta} - B\hat{K}_t\hat{x}_{t|t,\hat{\Theta}_t} + B\eta_t \\
&= A\hat{x}_{t|t,\Theta} - B\hat{K}_t\hat{x}_{t|t,\Theta} + B\hat{K}_t\hat{x}_{t|t,\Theta} - B\hat{K}_t\hat{x}_{t|t,\hat{\Theta}_t} + B\eta_t \\
&= \left(A - B\hat{K}_t\right)\hat{x}_{t|t,\Theta} - B\hat{K}_t\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) + B\eta_t.
\end{aligned}$$

$$\begin{aligned}
\hat{x}_{t+1|t,\hat{\Theta}_t} &= \hat{A}_t\hat{x}_{t|t,\hat{\Theta}_t} + \hat{B}_t u_t \\
&= \hat{A}_t\hat{x}_{t|t,\hat{\Theta}_t} - \hat{B}_t\hat{K}_t\hat{x}_{t|t,\hat{\Theta}_t} + \hat{B}_t\eta_t \\
&= \left(\hat{A}_t + A - A\right)\hat{x}_{t|t,\hat{\Theta}_t} + \left(-\hat{B}_t\hat{K}_t + B\hat{K}_t - B\hat{K}_t\right)\hat{x}_{t|t,\hat{\Theta}_t} + \hat{B}_t\eta_t \\
&= \left(\hat{A}_t + A - A - \hat{B}_t\hat{K}_t + B\hat{K}_t - B\hat{K}_t\right)\hat{x}_{t|t,\hat{\Theta}_t} + \hat{B}_t\eta_t \\
&= \left(\hat{A}_t + A - A - \hat{B}_t\hat{K}_t + B\hat{K}_t - B\hat{K}_t\right)\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta} + \hat{x}_{t|t,\Theta}\right) + \hat{B}_t\eta_t \\
&= \underbrace{\left(\hat{A}_t - A - \hat{B}_t\hat{K}_t + B\hat{K}_t\right)}_{\delta_{\hat{\Theta}_t}}\hat{x}_{t|t,\Theta} + \left(A - B\hat{K}_t\right)\hat{x}_{t|t,\Theta} \\
&\quad + \left(\hat{A}_t - A - \hat{B}_t\hat{K}_t + B\hat{K}_t\right)\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) + \left(A - B\hat{K}_t\right)\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) + \hat{B}_t\eta_t.
\end{aligned}$$

Thus Δ_{t+1} is,

$$\begin{aligned}
\Delta_{t+1} &= \hat{x}_{t+1|t,\Theta} - \hat{x}_{t+1|t,\hat{\Theta}_t} \\
&= \left(A - B\hat{K}_t\right)\hat{x}_{t|t,\Theta} - B\hat{K}_t\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) + B\eta_t \\
&\quad - \delta_{\hat{\Theta}_t}\hat{x}_{t|t,\Theta} - \delta_{\hat{\Theta}_t}\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) - \left(A - B\hat{K}_t\right)\left(\hat{x}_{t|t,\hat{\Theta}_t} - \hat{x}_{t|t,\Theta}\right) \\
&\quad - \left(A - B\hat{K}_t\right)\hat{x}_{t|t,\Theta} - \hat{B}_t\eta_t \\
&= A\left(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}_t}\right) - \delta_{\hat{\Theta}_t}\hat{x}_{t|t,\Theta} + \delta_{\hat{\Theta}_t}\left(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}_t}\right) + B\eta_t - \hat{B}_t\eta_t \\
&= \left(A + \delta_{\hat{\Theta}_t}\right)\left(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}_t}\right) - \delta_{\hat{\Theta}_t}\hat{x}_{t|t,\Theta} + \left(B - \hat{B}_t\right)\eta_t.
\end{aligned}$$

We will now decompose the term $\left(\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}_t}\right)$. From (2-6), we have [35]:

$$\hat{x}_{t|t,\Theta} - \hat{x}_{t|t,\hat{\Theta}_t} = \left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(L - \hat{L}_t\right)e_t + \hat{L}_t\left(\hat{C}_t - C\right)\hat{x}_{t|t-1,\Theta}.$$

Now, substituting the above expansion into Δ_{t+1}, we get

$$\begin{aligned}
\Delta_{t+1} &= \left(A + \delta_{\hat{\Theta}_t}\right)\left(\left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(L - \hat{L}_t\right)e_t + \hat{L}_t\left(\hat{C}_t - C\right)\hat{x}_{t|t-1,\Theta}\right) \\
&\quad - \delta_{\hat{\Theta}_t}\hat{x}_{t|t,\Theta} + \left(B - \hat{B}_t\right)\eta_t \\
&= \left(A + \delta_{\hat{\Theta}_t}\right)\left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(A + \delta_{\hat{\Theta}_t}\right)\left(L - \hat{L}_t\right)e_t \\
&\quad + \left(A + \delta_{\hat{\Theta}_t}\right)\hat{L}_t\left(\hat{C}_t - C\right)\hat{x}_{t|t-1,\Theta} - \delta_{\hat{\Theta}_t}\hat{x}_{t|t,\Theta} + \left(B - \hat{B}_t\right)\eta_t \\
&= \left(A + \delta_{\hat{\Theta}_t}\right)\left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(A + \delta_{\hat{\Theta}_t}\right)\left(L - \hat{L}_t\right)e_t \\
&\quad + \left(A + \delta_{\hat{\Theta}_t}\right)\hat{L}_t\left(\hat{C}_t - C\right)\hat{x}_{t|t-1,\Theta} - \delta_{\hat{\Theta}_t}\hat{x}_{t|t-1,\Theta} - \delta_{\hat{\Theta}_t}Le_t + \left(B - \hat{B}_t\right)\eta_t \\
&= \left(A + \delta_{\hat{\Theta}_t}\right)\left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(AL - A\hat{L}_t + \delta_{\hat{\Theta}_t}L - \delta_{\hat{\Theta}_t}\hat{L}_t - \delta_{\hat{\Theta}_t}L\right)e_t \\
&\quad + \left(\left(A + \delta_{\hat{\Theta}_t}\right)\hat{L}_t\left(\hat{C}_t - C\right) - \delta_{\hat{\Theta}_t}\right)\hat{x}_{t|t-1,\Theta} + \left(B - \hat{B}_t\right)\eta_t
\end{aligned}$$

$$
\begin{aligned}
&= \left(A + \delta_{\hat{\Theta}_t}\right)\left(I - \hat{L}_t\hat{C}_t\right)\Delta_t + \left(A\left(L - \hat{L}_t\right) - \delta_{\hat{\Theta}_t}\hat{L}_t\right)e_t \\
&\quad + \left(\left(A + \delta_{\hat{\Theta}_t}\right)\hat{L}_t\left(\hat{C}_t - C\right) - \delta_{\hat{\Theta}_t}\right)\hat{x}_{t|t-1,\Theta} + \left(B - \hat{B}_t\right)\eta_t \\
&= \sum_{i=0}^{t}\prod_{j=0}^{t-i-1}\left(\left(A + \delta_{\hat{\Theta}_{t-j}}\right)\left(I - \hat{L}_{t-j}\hat{C}_{t-j}\right)\right)\left(A\left(L - \hat{L}_i\right) - \delta_{\hat{\Theta}_i}\hat{L}_i\right)e_i \\
&\quad + \sum_{i=1}^{t}\prod_{j=0}^{t-i-1}\left(\left(A + \delta_{\hat{\Theta}_{t-j}}\right)\left(I - \hat{L}_{t-j}\hat{C}_{t-j}\right)\right)\left(\left(A + \delta_{\hat{\Theta}_i}\right)\hat{L}_i\left(\hat{C}_i - C\right) - \delta_{\hat{\Theta}_i}\right)\hat{x}_{i|i-1,\Theta} \\
&\quad + \sum_{i=0}^{t}\prod_{j=0}^{t-i-1}\left(\left(A + \delta_{\hat{\Theta}_{t-j}}\right)\left(I - \hat{L}_{t-j}\hat{C}_{t-j}\right)\right)\left(B - \hat{B}_i\right)\eta_i.
\end{aligned}
\tag{4-21}
$$

The last equality in (4-21) comes from the assumption that $\hat{x}_{0|-1,\Theta} = \hat{x}_{0|-1,\hat{\Theta}_{-1}} = 0$. Let us now decompose $\hat{x}_{i|i-1,\Theta}$:

$$
\begin{aligned}
\hat{x}_{i|i-1,\Theta} &= A\hat{x}_{i-1|i-1,\Theta} - B\hat{K}_{i-1}\hat{x}_{i-1|i-1,\hat{\Theta}_{i-1}} + B\eta_{i-1} \\
&= A\hat{x}_{i-1|i-1,\Theta} - B\hat{K}_{i-1}\hat{x}_{i-1|i-1,\Theta} - B\hat{K}_{i-1}\left(\hat{x}_{i-1|i-1,\hat{\Theta}_{i-1}} - \hat{x}_{i-1|i-1,\Theta}\right) + B\eta_{i-1} \\
&= \left(A - B\hat{K}_{i-1}\right)\hat{x}_{i-1|i-1,\Theta} + B\hat{K}_{i-1}\left(\hat{x}_{i-1|i-1,\Theta} - \hat{x}_{i-1|i-1,\hat{\Theta}_{i-1}}\right) + B\eta_{i-1} \\
&= \left(A - B\hat{K}_{i-1}\right)\left(\hat{x}_{i-1|i-2,\Theta} + Le_{i-1}\right) \\
&\quad + B\hat{K}_{i-1}\left(\left(I - \hat{L}_{i-1}\hat{C}_{i-1}\right)\Delta_{i-1} + \left(L - \hat{L}_{i-1}\right)e_{i-1} + \hat{L}_{i-1}\left(\hat{C}_{i-1} - C\right)\hat{x}_{i-1|i-2,\Theta}\right) \\
&\quad + B\eta_{i-1} \\
&= \left(A - B\hat{K}_{i-1}\left(I - \hat{L}_{i-1}\left(\hat{C}_{i-1} - C\right)\right)\right)\hat{x}_{i-1|i-2,\Theta} + B\hat{K}_{i-1}\left(I - \hat{L}_{i-1}\hat{C}_{i-1}\right)\Delta_{i-1} \\
&\quad + \left(\left(A - B\hat{K}_{i-1}\right)L + B\hat{K}_{i-1}\left(L - \hat{L}_{i-1}\right)\right)e_{i-1} + B\eta_{i-1} \\
&= \sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j}\left(A - B\hat{K}_{i-1-k} + B\hat{K}_{i-1-k}\hat{L}_{i-1-k}\left(\hat{C}_{i-1-k} - C\right)\right) \\
&\quad \left(B\hat{K}_j\left(I - \hat{L}_j\hat{C}_j\right)\Delta_j + \left(\left(A - B\hat{K}_j\right)L + B\hat{K}_j\left(L - \hat{L}_j\right)\right)e_j + B\eta_j\right).
\end{aligned}
\tag{4-22}
$$

For brevity of representation, we can define the following terms:

$$
\begin{aligned}
a_i &= \left(A + \delta_{\hat{\Theta}_i}\right)\left(I - \hat{L}_i\hat{C}_i\right) \\
b_i &= \left(A\left(L - \hat{L}_i\right) - \delta_{\hat{\Theta}_i}\hat{L}_i\right) \\
c_i &= \left(A + \delta_{\hat{\Theta}_i}\right)\hat{L}_i\left(\hat{C}_i - C\right) - \delta_{\hat{\Theta}_i} \\
d_i &= \left(B - \hat{B}_i\right)
\end{aligned}
$$

$$\begin{aligned}
f_i &= A - B\hat{K}_i + B\hat{K}_i\hat{L}_i\left(\hat{C}_i - C\right)\\
g_i &= B\hat{K}_i\left(I - \hat{L}_i\hat{C}_i\right)\\
h_i &= \left(A - B\hat{K}_i\right)L + B\hat{K}_i\left(L - \hat{L}_i\right).
\end{aligned} \tag{4-23}$$

Finally, using (4-21) with the equality given in (4-22), and with the definitions in (4-23), we get

$$\begin{aligned}
\Delta_{t+1} = &\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}b_ie_i + \sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}g_j\Delta_j\right)\\
&+\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}h_je_j\right) + \sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}B\eta_j\right)\\
&+\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}d_i\eta_i.
\end{aligned} \tag{4-24}$$

The above expression is similar to the one in the proof of Lemma 4.2 in [36]. The main difference lies in accounting for the additive excitation signal, i.e., the last two terms in the above expression. Since the model parameter estimation error is monotonically decreasing at every time step after the warm-up under the event $\mathcal{E}_{\mathbf{M}}$, we can upper bound the norm of each of the above terms in the decomposition by the bound at the end of the warm-up [35]. The bound for the first three terms in the above expression follows analogously to the bounds given in the proof of Lemma 4.2 in [36]. Further, it must be noted that the norm of the terms in (4-23) are identical to the ones in the proof of Lemma 4.2 in [36]. Hence, in this thesis, a detailed treatment of deriving the bounds on the norm of the terms in (4-23), is not provided.

Bounding the norm of $\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}B\eta_j\right)$

Expanding the term, we see that

$$\begin{aligned}
&\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}B\eta_j\right)\\
&= \left(a_ta_{t-1}...a_2c_1\right)\left(B\eta_0\right) + \left(a_ta_{t-1}...a_3c_2\right)\left(f_1B\eta_0 + B\eta_1\right) + \left(a_ta_{t-1}...a_4c_3\right)\left(f_2f_1B\eta_0 + f_2B\eta_1 + B\eta_2\right)\\
&\quad + \left(a_ta_{t-1}...a_5c_4\right)\left(f_3f_2f_1B\eta_0 + f_2f_1B\eta_1 + f_1B\eta_2 + B\eta_3\right) + ...\\
&\quad + c_t\left(f_{t-1}f_t...f_1B\eta_0 + f_{t-1}f_t...f_2B\eta_1 + ... + B\eta_{t-1}\right)\\
&= \left(a_ta_{t-1}...a_2c_1 + a_ta_{t-1}...a_3c_2f_1 + a_ta_{t-1}...a_4c_3f_2f_1 + ... + c_tf_{t-1}f_t...f_1\right)B\eta_0\\
&\quad + \left(a_ta_{t-1}...a_3c_2 + a_ta_{t-1}...a_4c_3f_2 + a_ta_{t-1}...a_5c_4f_2f_1 + ... + c_tf_{t-1}f_t...f_2\right)B\eta_1\\
&\quad + ... + c_tB\eta_{t-1}.
\end{aligned} \tag{4-25}$$

We will first start with bounding the norm of a_t. Recalling (4-23), we have

$$\begin{aligned}
a_t &= \left(A + \delta_{\hat{\Theta}_t}\right)\left(I - \hat{L}_t\hat{C}_t\right)\\
&= \left(A + \hat{A}_t - A - \hat{B}_t\hat{K}_t + B\hat{K}_t\right)\left(I - \hat{L}_t\hat{C}_t\right)\\
&= \left(\hat{A}_t - \left(\hat{B}_t - B\right)\hat{K}_t\right)\left(I - \hat{L}_t\hat{C}_t\right)\\
&= \left(\hat{A}_t - \hat{A}_t\hat{L}_t\hat{C}_t\right) - \left(\hat{B}_t - B\right)\hat{K}_t\left(I - \hat{L}_t\hat{C}_t\right).
\end{aligned}$$

Let $\sigma \in \mathbb{R}$ satisfy $1 > \sigma > \max\{\rho, \nu\}$. Under the event $\mathcal{E}_{\mathbf{M}}$, we have $||a_t|| \leq \sigma < 1$ for all $t \geq T_{\text{w}} \geq T_a$, where

$$T_a = T_B\left(\frac{\Gamma\left(1 + \zeta\left(||C|| + 1\right)\right)}{(\sigma - \nu)}\right)^2.$$

Similarly, we can bound the norm of f_t. If

$$T_f = T_A\frac{\sigma_n^2(A)}{4}\left(\frac{1 + \Gamma + \Gamma\zeta\,||B||}{\sigma - \rho}\right)^2,$$

we obtain $||f_t|| \leq \sigma < 1$ for all $t \geq T_{\text{w}} \geq T_f$. Furthermore, for $T_{\text{w}} \geq \max\{T_a, T_f\}$, we have that for all $t \geq T_{\text{w}}$, $\max\{||a_t||, ||f_t||\} \leq \sigma < 1$. Further, a bound on the norm of c_t exists under the event $\mathcal{E}_{\mathbf{M}}$:

$$||c_t|| \leq 2\Bigg(\Phi(A) + \beta_A(T_{\text{w}}) + \Gamma\beta_B(T_{\text{w}})\Bigg)\zeta\beta_C(T_{\text{w}}) + 2\Bigg(\beta_A(T_{\text{w}}) + \Gamma\beta_B(T_{\text{w}})\Bigg) := \bar{c}. \tag{4-26}$$

Therefore, from (4-25) and using the above bounds, we have the following under event $\mathcal{E}_{\mathbf{M}}$:

$$\begin{aligned}
&\left|\left|\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}B\eta_j\right)\right|\right|\\
&\leq \left(t\sigma^{t-1} + (t-1)\sigma^{t-2} + ... + 2\sigma + 1\right)\bar{c}||B||\sqrt{\gamma}\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)}\\
&= \bar{c}||B||\sqrt{\gamma}\left((1 + \sigma + \sigma^2 + ... + \sigma^{t-1}) + (\sigma + \sigma^2 + ... + \sigma^{t-1}) + (\sigma^2 + \sigma^3 ... + \sigma^{t-1}) + ... + \sigma^{t-1}\right)\\
&\quad\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)}\\
&\leq \bar{c}||B||\sqrt{\gamma}\left(\frac{1}{1-\sigma} + \frac{\sigma}{1-\sigma} + ...\right)\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)}\\
&= \frac{\bar{c}||B||\sqrt{\gamma}}{(1-\sigma)^2}\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)},
\end{aligned} \tag{4-27}$$

which holds with a probability of at least $1 - \delta$.

Bounding the norm of $\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}d_i\eta_i$

The bound on this term follows a similar procedure as that of the bound in (4-27). It is straightforward to see that $||d_t|| \leq 1$ under the event $\mathcal{E}_{\mathbf{M}}$. Therefore, we have the following under the event $\mathcal{E}_{\mathbf{M}}$:

$$\left|\left|\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}d_i\eta_i\right|\right| \leq \frac{\sqrt{\gamma}}{1-\sigma}\sqrt{2n_u \log\left(\frac{2n_uT}{\delta}\right)}, \tag{4-28}$$

which holds with a probability of at least $1-\delta$.

Bounding the norm of $\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}b_ie_i$ [35]

Similar to the other terms, we can indeed say that a bound on the norm of b_t exists under the event $\mathcal{E}_{\mathbf{M}}$:

$$||b_t|| \leq 2\Phi(A)\beta_L(T_{\mathrm{w}}) + 2\beta_A(T_{\mathrm{w}})\zeta + 2\beta_B(T_{\mathrm{w}})\Gamma\zeta := \bar{b}.$$

Observe that e_t is a $\left(||C||||\Sigma||^{1/2}+\sigma_z\right)$ - sub - Gaussian random variable. Therefore, by using Lemma B.1, for all $t \geq T_{\mathrm{w}}$ with probability of at least $1-\delta$, we have the following under event $\mathcal{E}_{\mathbf{M}}$:

$$\left|\left|\sum_{i=0}^{t}\prod_{j=0}^{t-i-1} a_{t-j}b_ie_i\right|\right| \leq \frac{\bar{b}}{1-\sigma}\left(||C||||\Sigma||^{1/2}+\sigma_z\right)\sqrt{2n_y \log\left(\frac{2n_yT}{\delta}\right)}. \tag{4-29}$$

Bounding the norm of $\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}h_je_j\right)$ [35]

Under the event $\mathcal{E}_{\mathbf{M}}$, we have

$$||h_t|| \leq \left(2\beta_A(T_{\mathrm{w}}) + \rho + 2\beta_B(T_{\mathrm{w}})\Gamma\right)\zeta + 2||B||\Gamma\beta_L(T_{\mathrm{w}}) := \bar{h}.$$

Similar to the previous bound, we have the following under event $\mathcal{E}_{\mathbf{M}}$:

$$\left|\left|\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}h_je_j\right)\right|\right| \leq \frac{\bar{c}\bar{h}}{(1-\sigma)^2}\left(||C||||\Sigma||^{1/2}+\sigma_z\right)\sqrt{2n_y \log\left(\frac{2n_yT}{\delta}\right)}, \tag{4-30}$$

which holds with a probability of at least $1-\delta$.

Bounding the norm of $\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} a_{t-j}c_i\left(\sum_{j=0}^{i-1}\prod_{k=0}^{i-2-j} f_{i-1-k}g_j\Delta_j\right)$ [35]

It can be shown under the event $\mathcal{E}_{\mathbf{M}}$ that, $||g_t|| \leq \bar{g}$ exists. Further, it can be shown through the method of induction, that for all $t \geq T_{\mathrm{w}}$, under the event $\mathcal{E}_{\mathbf{M}}$, we have the following with a probability of at least $1-\delta$:

$$\begin{aligned}\left|\left|\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right|\right| &= ||\Delta_t|| \leq \bar{\Delta},\\ \bar{\Delta} = 10&\left(\frac{\bar{b}}{1-\sigma}+\frac{\bar{c}\bar{h}}{(1-\sigma)^2}\right)\left(||C||||\Sigma||^{1/2}+\sigma_z\right)\sqrt{2n_y\log\left(\frac{2n_yT}{\delta}\right)}\\ &+10\left(\frac{\sqrt{\gamma}}{1-\sigma}+\frac{\bar{c}||B||\sqrt{\gamma}}{(1-\sigma)^2}\right)\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)}.\end{aligned} \tag{4-31}$$

For details on the derivation of the above expression, refer to the proof of Lemma 4.1 in [35].

Putting things together

To recall, from (4-15) we have

$$\begin{aligned}\hat{x}_{t|t,\hat{\Theta}_t} =& \sum_{i=1}^{t}\prod_{j=0}^{t-i-1} N_{t-j}\left(\hat{L}_i C\left(x_i - \hat{x}_{i|i-1,\Theta}\right) + \hat{L}_i C\left(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\hat{\Theta}_{i-1}}\right) + \hat{B}_{i-1}\eta_{i-1}\right. \\ &\left. + \hat{L}_i z_i + \hat{L}_i\left(C - \hat{C}_i\right)\hat{B}_{i-1}\eta_{i-1}\right).\end{aligned}$$

From (4-17), (4-18), (4-20), and (4-31), we have the following under the event $\mathcal{E}_{\mathbf{M}}$:

$$\begin{aligned}\left|\left|\hat{x}_{t|t,\hat{\Theta}_t}\right|\right| \leq& \left|\left|\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} N_{t-j}\left(\hat{L}_i C\left(x_i - \hat{x}_{i|i-1,\Theta}\right) + \hat{L}_i C\left(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\hat{\Theta}_{i-1}}\right) + \hat{B}_{i-1}\eta_{i-1}\right.\right. \\ &\left.\left. + \hat{L}_i z_i + \hat{L}_i\left(C - \hat{C}_i\right)\hat{B}_{i-1}\eta_{i-1}\right)\right|\right| \\ \leq& \max_{1\leq i\leq t}\left|\left|\left(\hat{L}_i C\left(x_i - \hat{x}_{i|i-1,\Theta}\right) + \hat{L}_i C\left(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\hat{\Theta}_{i-1}}\right) + \hat{B}_{i-1}\eta_{i-1}\right.\right. \\ &\left.\left. + \hat{L}_i z_i + \hat{L}_i\left(C - \hat{C}_i\right)\hat{B}_{i-1}\eta_{i-1}\right)\right|\right| \left|\left|\sum_{i=1}^{t}\prod_{j=0}^{t-i-1} N_{t-j}\right|\right| \\ \leq& \left(1 + \frac{1+\rho}{2} + \left(\frac{1+\rho}{2}\right)^2 + \ldots + \left(\frac{1+\rho}{2}\right)^{t-1}\right)\max_{1\leq i\leq t}\left|\left|\left(\hat{L}_i C\left(x_i - \hat{x}_{i|i-1,\Theta}\right)\right.\right.\right. \\ &\left.\left. + \hat{L}_i C\left(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\hat{\Theta}_{i-1}}\right) + \hat{B}_{i-1}\eta_{i-1} + \hat{L}_i z_i + \hat{L}_i\left(C - \hat{C}_i\right)\hat{B}_{i-1}\eta_{i-1}\right)\right|\right| \\ \leq& \frac{2}{1-\rho}\max_{1\leq i\leq t}\left|\left|\left(\hat{L}_i C\left(x_i - \hat{x}_{i|i-1,\Theta}\right) + \hat{L}_i C\left(\hat{x}_{i|i-1,\Theta} - \hat{x}_{i|i-1,\hat{\Theta}_{i-1}}\right) + \hat{B}_{i-1}\eta_{i-1}\right.\right. \\ &\left.\left. + \hat{L}_i z_i + \hat{L}_i\left(C - \hat{C}_i\right)\hat{B}_{i-1}\eta_{i-1}\right)\right|\right| \\ \leq& \bar{\chi},\end{aligned} \tag{4-32}$$

which holds with a probability of at least $1 - 3\delta$, where

$$\bar{\chi} := \frac{2\left(\zeta\left(||C||||\Sigma||^{1/2} + \sigma_z\right)\sqrt{2n_x\log\left(\frac{2n_xT}{\delta}\right)} + \zeta||C||\bar{\Delta} + \sqrt{\gamma}\left(||B|| + 1\right)\left(1 + \zeta\right)\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)}\right)}{1-\rho}. \tag{4-33}$$

Now, using (4-33), we can bound the norm of $\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}$ as follows. From (2-6), we have

$$\hat{x}_{t|t-1,\hat{\Theta}_{t-1}} = \left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1},$$

then

$$\begin{aligned}||\hat{x}_{t|t-1,\hat{\Theta}_{t-1}}|| \leq& \left|\left|\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right|\right|\left|\left|\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}}\right|\right| + \left|\left|\hat{B}_{t-1}\right|\right|||\eta_{t-1}|| \\ \leq& \rho\bar{\chi} + \sqrt{\gamma}\left(||B|| + 1\right)\sqrt{2n_u\log\left(\frac{2n_uT}{\delta}\right)} := X_{\text{est,ac}},\end{aligned} \tag{4-34}$$

which holds with a probability of at least $1-3\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. Now to bound the norm of u_t, we recall the following from (3-9):

$$u_t = -\hat{K}_t \hat{x}_{t|t,\hat{\Theta}_t} + \eta_t,$$

then

$$||u_t|| \leq \Gamma\bar{\chi} + \sqrt{\gamma}\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)} := U_{\text{ac}}, \tag{4-35}$$

which holds with a probability of at least $1-3\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. To derive a bound on the norm of x_t, we do the following:

$$\begin{aligned} x_t &= x_t - \hat{x}_{t|t-1,\Theta} + \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} \\ &= x_t - \hat{x}_{t|t-1,\Theta} + \hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} \\ &\quad + \left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1}. \end{aligned}$$

Then,

$$\begin{aligned} ||x_t|| &\leq \left|\left|x_t - \hat{x}_{t|t-1,\Theta}\right|\right| + \left|\left|\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right|\right| + \left|\left|\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\right|\right| \left|\left|\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}}\right|\right| \\ &\quad + \left|\left|\hat{B}_{t-1}\right|\right| ||\eta_{t-1}|| \\ &\leq \left|\left|x_t - \hat{x}_{t|t-1,\Theta}\right|\right| + \left|\left|\hat{x}_{t|t-1,\Theta} - \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right|\right| + \left|\left|\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right)\right|\right| \left|\left|\hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}}\right|\right| \\ &\quad + \left(||B|| + \left|\left|B - \hat{B}_{t-1}\right|\right|\right) ||\eta_{t-1}|| \\ &\leq ||\Sigma||^{1/2}\sqrt{2n_x \log\left(\frac{2n_x T}{\delta}\right)} + \bar{\Delta} + \rho\bar{\chi} + \sqrt{\gamma}\left(||B|| + 1\right)\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)} := X_{ac}, \end{aligned} \tag{4-36}$$

which holds with a probability of at least $1-3\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. Now for y_t, we have

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

Using similar analysis of x_t, we get the following bound for y_t for all $t \geq T_{\text{w}}$:

$$\begin{aligned} ||y_t|| &\leq \rho||C||\bar{\chi} + \sqrt{\gamma}||C||\left(1 + ||B||\right)\sqrt{2n_u \log\left(\frac{2n_u T}{\delta}\right)} \\ &\quad + \left(||C||||\Sigma||^{1/2} + \sigma_z\right)\sqrt{2n_x \log\left(\frac{2n_x T}{\delta}\right)} + ||C||\bar{\Delta} := Y_{\text{ac}}, \end{aligned} \tag{4-37}$$

which holds with a probability of at least $1-3\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. By re-parameterising $3\delta \to \delta$, the above bounds can be guaranteed with a probability of at least $1-\delta$. This concludes the proof.

4-3-2 Proof of Lemma 4.2

The proof critically requires the 'truncated closed-loop noise evolution parameter'. This parameter captures how the noise and excitation signal sequences influence the input-output data in the vector ϕ. As mentioned previously, this proof is an extension of the result in [35], and the extension lies in accounting for the additive excitation signal. Conceptually, this proof follows similar steps as the one in [35] but, the terms involved are different. Firstly, the expression for the truncated closed-loop noise evolution parameter is derived.

Truncated closed-loop noise evolution parameter

The truncated closed-loop noise evolution parameter derivation is an extension to the one provided in [35], the difference lies in accounting for the additive Gaussian excitation signals. The truncated closed-loop noise evolution parameter, which captures the effect of noises and excitation signals on the vector ϕ_t, will play an important role in establishing the persistence of excitation during the LBC period. Consider a permutation of the vector ϕ_t, i.e., $\bar{\phi}_t = P\phi_t$, where

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

and P is the permutation matrix. The present analysis abstracts away the episodic behaviour of the algorithm, that is, $\hat{\Theta}_{t-1}$ could either be the model parameter estimate being used in the current episode at time step $t-1$ or it could be the model parameter estimated in the previous episode if the time step $t-1$ falls in the previous episode. The analysis is independent of the type of algorithmic behaviour in terms of when the system parameter is being updated.

Following the warm-up period, recall that the LBC policy $u_t = -\hat{K}_t \hat{x}_{t|t,\hat{\Theta}_t} + \eta_t$ is deployed, where

$$\begin{aligned} u_t &= -\hat{K}_t \hat{x}_{t|t,\hat{\Theta}_t} + \eta_t, \\ \eta_t &= \sqrt{\frac{\gamma}{\sqrt{l_k}}} r_0, \;\; r_0 \sim \mathcal{N}(0, I), \gamma > 0, \end{aligned}$$

where l_k is the number of time steps in the k^{th} episode. Recall the following relation from (2-6):

$$\begin{aligned} \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} &= \left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right) \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1}, \\ \hat{x}_{t|t,\hat{\Theta}_t} &= \hat{x}_{t|t-1,\hat{\Theta}_{t-1}} + \hat{L}_t \left(y_t - \hat{C}_t \hat{x}_{t|t-1,\hat{\Theta}_{t-1}}\right) \\ &= \left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right) \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1} \\ &\quad + \hat{L}_t \Bigg(C x_t + z_t - \hat{C}_t \left(\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right) \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1} \right) \Bigg) \\ &= \left(I - \hat{L}_t \hat{C}_t\right) \Bigg(\left(\hat{A}_{t-1} - \hat{B}_{t-1}\hat{K}_{t-1}\right) \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + \hat{B}_{t-1}\eta_{t-1} \Bigg) \\ &\quad + \hat{L}_t \Bigg(C \left(A x_{t-1} - B\hat{K}_{t-1} \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} + B\eta_{t-1} + w_{t-1} \right) + z_t \Bigg). \end{aligned}$$

Now,

$$\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}_t} \end{bmatrix} = \underbrace{\begin{bmatrix} A & -B\hat{K}_{t-1} \\ \hat{L}_tCA & \left(I-\hat{L}_t\hat{C}_t\right)\left(\hat{A}_{t-1}-\hat{B}_{t-1}\hat{K}_{t-1}\right)-\hat{L}_tCB\hat{K}_{t-1} \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})}} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} \end{bmatrix}$$

$$+ \underbrace{\begin{bmatrix} I & 0 & B & 0 \\ \hat{L}_tC & \hat{L}_t & \left(I-\hat{L}_t\hat{C}_t\right)\hat{B}_{t-1}+\hat{L}_tCB & 0 \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t})}} \begin{bmatrix} w_{t-1} \\ z_t \\ \eta_{t-1} \\ \eta_t \end{bmatrix}.$$

Let $f_t = \begin{bmatrix} y_t \\ u_t \end{bmatrix}$. Now, we can express f_t as:

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

$$+ \begin{bmatrix} C & I & CB & 0 \\ -\hat{K}_t\hat{L}_tC & -\hat{K}_t\hat{L}_t & -\hat{K}_t\left(I-\hat{L}_t\hat{C}_t\right)\hat{B}_{t-1}-\hat{K}_t\hat{L}_tCB & I \end{bmatrix} \begin{bmatrix} w_{t-1} \\ z_t \\ \eta_{t-1} \\ \eta_t \end{bmatrix}$$

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

$$= \hat{\mathbf{\Gamma}}_{\mathbf{t}}\hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} \end{bmatrix} + \hat{\mathbf{\Gamma}}_{\mathbf{t}}\hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t})} \begin{bmatrix} w_{t-1} \\ z_t \\ \eta_{t-1} \\ \eta_t \end{bmatrix} + \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} w_{t-1} \\ z_t \\ \eta_{t-1} \\ \eta_t \end{bmatrix}.$$

Rolling back in time $\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}_t} \end{bmatrix}$ for H-time steps, we obtain the following:

$$\begin{aligned} \underbrace{\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}_t} \end{bmatrix}}_{x_t^e} &= \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})} \underbrace{\begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}_{t-1}} \end{bmatrix}}_{x_{t-1}^e} + \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t})} \underbrace{\begin{bmatrix} w_{t-1} \\ z_t \\ \eta_{t-1} \\ \eta_t \end{bmatrix}}_{\eta_{t-1}^e} \\ &= \prod_{i=0}^{H} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-i})} x_{t-H-1}^e + \sum_{i=1}^{H+1} \left(\prod_{j=2}^{i} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-j+2})} \right) \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-i+1})} \eta_{t-i}^e. \end{aligned} \tag{4-38}$$

Now, let us expand x_{t-H-1}^e.

$$x_{t-H-1}^e = \prod_{i=H+1}^{t-1} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-i})} \begin{bmatrix} x_0 \\ \hat{L}_0\hat{C}_0x_0 + \hat{L}_0z_0 \end{bmatrix} + \sum_{i=H+2}^{t} \left(\prod_{j=H+3}^{i} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-j+2})} \right) \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-i+1})} \eta_{t-i}^e. \tag{4-39}$$

The equality comes from the assumption that $\hat{x}_{0|-1,\hat{\Theta}_{-1}} = 0$. Therefore, x^e_{t-H-1} represents the effect of $\begin{bmatrix} w_{i-1} \\ z_i \\ \eta_{i-1} \\ \eta_i \end{bmatrix}$ for $0 \leq i < t-H-1$, which are independent with respect to the time. Now f_t can be rolled back H-time steps backwards, as follows:

$$
\begin{aligned}
f_t &= \hat{\mathbf{\Gamma}}_{\mathbf{t}} x^e_t + \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \eta^e_{t-1} \\
&= \hat{\mathbf{\Gamma}}_{\mathbf{t}} \left(\prod_{i=0}^{H} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-i})} x^e_{t-H-1} + \sum_{i=1}^{H+1} \left(\prod_{j=2}^{i} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-j+2})} \right) \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-i+1})} \eta^e_{t-i} \right) \\
&\quad + \begin{bmatrix} 0 & I & 0 & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \eta^e_{t-1} \\
&= \hat{\mathbf{\Gamma}}_{\mathbf{t}} \sum_{i=2}^{H+1} \left(\prod_{j=2}^{i} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-j+2})} \right) \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-i+1})} \eta^e_{t-i} \\
&\quad + \underbrace{\begin{bmatrix} C & I & CB & 0 \\ -\hat{K}_t \hat{L}_t C & -\hat{K}_t \hat{L}_t & -\hat{K}_t \left(I - \hat{L}_t \hat{C}_t \right) \hat{B}_{t-1} - \hat{K}_t \hat{L}_t CB & I \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{1}}^{(\mathbf{t})}} \eta^e_{t-1} + \mathbf{r}^{\mathbf{c}}_{\mathbf{t}},
\end{aligned}
\tag{4-40}
$$

where $\mathbf{r}^{\mathbf{c}}_{\mathbf{t}}$ is the residual vector that represents the effect of $\begin{bmatrix} w_{i-1} \\ z_i \\ \eta_{i-1} \\ \eta_i \end{bmatrix}$ for $0 \leq i \leq t-H-1$, which are independent with respect to the time. Now f_t can be compactly represented as such:

$$
f_t = \bar{\mathbf{G}}_{\mathbf{t}} \begin{bmatrix} \eta^e_{t-1} \\ \eta^e_{t-2} \\ . \\ . \\ . \\ \eta^e_{t-H-1} \end{bmatrix} + \mathbf{r}^{\mathbf{c}}_{\mathbf{t}},
$$

where

$$
\begin{aligned}
\bar{\mathbf{G}}_{\mathbf{t}} = &\left[\hat{\mathbf{G}}_{\mathbf{1}}^{(\mathbf{t})} \quad \hat{\mathbf{\Gamma}}_{\mathbf{t}} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})} \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-1})} \quad \hat{\mathbf{\Gamma}}_{\mathbf{t}} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-1})} \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-2})} \quad ... \quad \hat{\mathbf{\Gamma}}_{\mathbf{t}} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t})} \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-1})} ... \hat{\mathbf{G}}_{\mathbf{2}}^{(\mathbf{t-H+1})} \hat{\mathbf{G}}_{\mathbf{3}}^{(\mathbf{t-H})} \right] \\
&\in \mathbb{R}^{(n_y+n_u) \times (H+1)(n_x+n_y)}.
\end{aligned}
$$

We can now represent $\bar{\phi}_t$ as follows:

$$
\bar{\phi}_t = \begin{bmatrix} f_{t-1} \\ . \\ . \\ . \\ f_{t-H} \end{bmatrix} + \begin{bmatrix} \mathbf{r}^{\mathbf{c}}_{\mathbf{t-1}} \\ . \\ . \\ . \\ \mathbf{r}^{\mathbf{c}}_{\mathbf{t-H}} \end{bmatrix} = \mathcal{G}^{\mathrm{cl}}_t \begin{bmatrix} \eta^e_{t-2} \\ \eta^e_{t-3} \\ . \\ . \\ . \\ \eta^e_{t-2H-1} \end{bmatrix} + \begin{bmatrix} \mathbf{r}^{\mathbf{c}}_{\mathbf{t-1}} \\ . \\ . \\ . \\ \mathbf{r}^{\mathbf{c}}_{\mathbf{t-H}} \end{bmatrix},
$$

where

$$\mathcal{G}_t^{\mathrm{cl}} = \begin{bmatrix} \mathbf{\bar{G}_{t-1}} & 0 & 0 & 0 & ... \\ 0 & \mathbf{\bar{G}_{t-2}} & 0 & 0 & ... \\ . & & & & \\ . & & & & \\ . & & & & \\ 0 & 0 & 0 & ... & \mathbf{\bar{G}_{t-H}} \end{bmatrix}. \tag{4-41}$$

If the true system parameter is known, the optimal control policy can be deployed. Then, $\mathcal{G}^{cl}$ captures the effect of the process and measurement noises as well as excitation signals on $\bar{\phi}_t$ while using the optimal control policy:

$$\mathcal{G}^{\mathrm{cl}} = \begin{bmatrix} \mathbf{\bar{G}} & 0 & 0 & 0 & ... \\ 0 & \mathbf{\bar{G}} & 0 & 0 & ... \\ . & & & & \\ . & & & & \\ . & & & & \\ 0 & 0 & 0 & ... & \mathbf{\bar{G}} \end{bmatrix},$$

where

$$\mathbf{\bar{G}} = \begin{bmatrix} \mathbf{G_1} & \mathbf{\Gamma G_2 G_3} & \mathbf{\Gamma {G_2}^2 G_3} & ... & \mathbf{\Gamma {G_2}^H G_3} \end{bmatrix} \in \mathbb{R}^{(n_y+n_u)\times(H+1)(n_x+n_y)}$$

with

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

With H chosen such that $\mathbf{\bar{G}}$ is full-row rank, $\mathcal{G}^{cl}$ can be shown to have full-row rank with QR decomposition [35]. Therefore, $\sigma_{\min}\left(\mathcal{G}^{\mathrm{cl}}\right) \geq \sigma_{\mathrm{c}} > 0$.

The proof

The event $\mathcal{E}_{\mathbf{M}}$ will be extensively used in this proof. To recall from (4-16), we have

$$\mathcal{E}_{\mathbf{M}} := \left\{ ||\mathbf{\hat{M}_t} - \mathbf{M}|| \leq 1 \right\}, \tag{4-42}$$

which is assumed to hold with high probability when $t \geq T_{\mathrm{w}} \geq T_{\mathbf{M}}$. This event will indeed be shown to hold with high probability in Theorem 4.3. Now,

$$\mathbb{E}\left[\bar{\phi}_t\bar{\phi}_t^\top\right] = \mathcal{G}_t^{\mathrm{cl}}\Sigma_{w,z,\eta}{\mathcal{G}_t^{\mathrm{cl}}}^\top + \begin{bmatrix}\mathbf{r^c_{t-1}}\\ .\\ .\\ .\\ \mathbf{r^c_{t-H}}\end{bmatrix}\begin{bmatrix}\mathbf{r^c_{t-1}}\\ .\\ .\\ .\\ \mathbf{r^c_{t-H}}\end{bmatrix}^\top,$$

where $\Sigma_{w,z,\eta} \in \mathbb{R}^{2H(n_x+n_y+2n_u)\times 2H(n_x+n_y+2n_u)}$ is

$$\Sigma_{w,z,\eta} = \mathrm{diag}(\sigma_w^2 I, \sigma_z^2 I, \sigma_{\eta_{t-2}}^2 I, \sigma_{\eta_{t-1}}^2 I, ..., \sigma_w^2 I, \sigma_z^2 I, \sigma_{\eta_{t-2H-1}}^2 I, \sigma_{\eta_{t-2H}}^2 I).$$

The notation diag(.) signifies a diagonal matrix, where the diagonal elements are the arguments in this operator. This implies

$$\begin{aligned}\mathbb{E}\left[\bar{\phi}_t\bar{\phi}_t^\top\right] &\geq \mathcal{G}_t^{\mathrm{cl}}\Sigma_{w,z,\eta}{\mathcal{G}_t^{\mathrm{cl}}}^\top\\ \implies \sigma_{\min}\left(\mathbb{E}\left[\bar{\phi}_t\bar{\phi}_t^\top\right]\right) &\geq \sigma_{\min}^2(\mathcal{G}_t^{\mathrm{cl}})\min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta_{t-1}}^2\}.\end{aligned}$$

To proceed, we require a lower bound on $\sigma_{\min}(\mathcal{G}_t^{\mathrm{cl}})$. To obtain such a lower bound, we can first say that the following perturbation bound exists under the event $\mathcal{E}_{\mathbf{M}}$:

$$||\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}}|| \leq \frac{\sigma_{\mathrm{c}}}{2},$$

if $t \geq T_{\mathrm{w}} \geq T_{\mathcal{G}}$ for some $T_{\mathcal{G}} > H$. It is possible to represent $||\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}}||$ as a function of system parameter estimation error whose bound exists under the event $\mathcal{E}_{\mathbf{M}}$. The detailed treatment of this bound is deferred to future work. The above perturbation bound, which is based on the proof of Lemma 3.2 in [35], also fails to provide sufficient details on the derivation of the above-mentioned perturbation bound.

One of the fundamental results of Weyl's inequalities on singular values is as follows:

$$\sigma_j(X) + \sigma_j(Y) \leq \sigma_1(X+Y), \;\; j = 1, 2, ..., \min\{m, n\}, \tag{4-43}$$

holds for any two matrices $X, Y \in \mathbb{R}^{m\times n}$. Taking $j = 1$ and replacing Y with $-Y$, we have

$$\begin{aligned}\sigma_{\min}(X) - \sigma_{\min}(Y) &\leq \sigma_{\max}(X - Y)\\ \implies \sigma_{\min}(X) - \sigma_{\max}(X - Y) &\leq \sigma_{\min}(Y).\end{aligned}$$

Now taking $X = \mathcal{G}^{\mathrm{cl}}$ and $Y = \mathcal{G}_t^{\mathrm{cl}}$, we have

$$\begin{aligned}\sigma_{\min}(\mathcal{G}^{\mathrm{cl}}) - \sigma_{\max}(\mathcal{G}^{\mathrm{cl}} - \mathcal{G}_t^{\mathrm{cl}}) &\leq \sigma_{\min}(\mathcal{G}_t^{\mathrm{cl}})\\ \implies \sigma_{\min}(\mathcal{G}_t^{\mathrm{cl}}) &\geq \sigma_{\min}(\mathcal{G}^{\mathrm{cl}}) - \sigma_{\max}(\mathcal{G}_t^{\mathrm{cl}} - \mathcal{G}^{\mathrm{cl}})\\ \implies \sigma_{\min}(\mathcal{G}_t^{\mathrm{cl}}) &\geq \frac{\sigma_{\mathrm{c}}}{2}.\end{aligned} \tag{4-44}$$

With the above result, we finally have

$$\begin{aligned}\sigma_{\min}\left(\mathbb{E}\left[\bar{\phi}_t\bar{\phi}_t^\top\right]\right) &\geq \sigma_{\min}^2(\mathcal{G}_t^{\mathrm{cl}})\min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta_{t-1}}^2\}\\ &\geq \frac{\sigma_{\mathrm{c}}^2}{4}\min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta_{t-1}}^2\},\end{aligned} \tag{4-45}$$

for $t \geq T_{\text{w}}$. Since singular values do not change under permutation, $\sigma_{\min}\left(\mathbb{E}\left[\bar{\phi}_t\bar{\phi}_t^\top\right]\right) = \sigma_{\min}\left(\mathbb{E}\left[\phi_t\phi_t^\top\right]\right)$. To recall, we need to derive a lower bound for $\sigma_{\min}\left(\sum_{i=T_{\text{w}}}^{t-1}\phi_i\phi_i^\top\right)$. Firstly, we will derive a bound on $\|\ \phi_t\ \|$. From Lemma 4.1, we have

$$\begin{aligned}
||\ \phi_t\ || &= \sqrt{\sum_{i=1}^{H}||y_{t-i}||^2 + ||u_{t-i}||^2} \\
&\leq \sqrt{H\max_{1\leq i\leq H}(||y_{t-i}||^2 + ||u_{t-i}||^2)} \\
&\leq \left(\sqrt{\max_{1\leq i\leq H}||y_{t-i}||^2} + \sqrt{\max_{1\leq i\leq H}||u_{t-i}||^2}\right)\sqrt{H} \rightarrow \text{triangle inequality} \\
&\leq \underbrace{(Y_{\text{ac}} + U_{\text{ac}})}_{\Upsilon_{\text{ac}}}\sqrt{H},
\end{aligned} \tag{4-46}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. We can re-parameterise $\delta \rightarrow \delta/2$. This requires appropriately modifying the $\sqrt{\log(.)}$ terms of Y_{ac} and U_{ac}. The reason for re-parameterising will become apparent shortly. We will now apply Lemma B.2. In Lemma B.2, it becomes evident that the notations $\mathbf{X}_k = \mathbf{A}_k = \phi_i\phi_i^\top$, and

$$\begin{aligned}
\sigma^2 &= ||\sum_{i=T_{\text{w}}}^{t-1}(\phi_i\phi_i^\top)^2|| \\
&\leq (t-T_{\text{w}})\max_{T_{\text{w}}\leq i\leq t-1}||\ \phi_i\ ||^2\ ||\ \phi_i^\top\ ||^2 \\
&= (t-T_{\text{w}})\max_{T_{\text{w}}\leq i\leq t-1}||\ \phi_i\ ||^4\ .
\end{aligned}$$

From Lemma B.2, we can set

$$\begin{aligned}
&\frac{\delta}{2} = H(n_y+n_u)\ \exp\left(\frac{-t^2}{8\sigma^2}\right) \\
\implies &-\log\left(\frac{\delta}{2H(n_y+n_u)}\right) = \frac{t^2}{8\sigma^2} \\
\implies &\log\left(\frac{2H(n_y+n_u)}{\delta}\right) = \frac{t^2}{8\sigma^2} \\
\implies &t = 2\sqrt{2}\sigma\sqrt{\log\left(\frac{2H(n_y+n_u)}{\delta}\right)} \\
\implies &t = 2\sqrt{2(t-T_{\text{w}})}\max_{T_{\text{w}}\leq i\leq t-1}||\phi_i||^2\sqrt{\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}.
\end{aligned}$$

Finally by using Lemma B.2, we have

$$\lambda_{\max}\left(\sum_{i=T_{\text{w}}}^{t-1}\phi_i\phi_i^\top - \mathbb{E}\left[\phi_i\phi_i^\top\right]\right) \leq 2\sqrt{2(t-T_{\text{w}})}\Upsilon_{\text{ac}}^2 H\sqrt{\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}, \tag{4-47}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. Since $\phi_i\phi_i^\top$ is a symmetric matrix, its singular values are the absolute values of its eigenvalues. Now using Weyl's inequality as described in (4-43), we get

$$\begin{aligned}
&\sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top\right) \geq \sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\mathbb{E}[\phi_i\phi_i^\top]\right) - \left|\lambda_{\max}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top - \mathbb{E}[\phi_i\phi_i^\top]\right)\right| \\
\implies &\sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top\right) \geq (t-T_{\mathrm{w}})\frac{\sigma_{\mathrm{c}}^2}{4}\min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta_{t-1}}^2\} - 2\sqrt{2(t-T_{\mathrm{w}})}\Upsilon_{\mathrm{ac}}^2 H\sqrt{\log\left(\frac{2H(n_y+n_u)}{\delta}\right)},
\end{aligned} \tag{4-48}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. Now we need to determine the minimum number of time steps to ensure $\sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top\right) > 0$. Equating the LHS of (4-48) to 0 we get,

$$\begin{aligned}
&2\sqrt{2(t-T_{\mathrm{w}})}\Upsilon_{\mathrm{ac}}^2 H\sqrt{\log\left(\frac{2H(n_y+n_u)}{\delta}\right)} \geq (t-T_{\mathrm{w}})\frac{\sigma_{\mathrm{c}}^2}{4}\min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta_{t-1}}^2\} \\
\implies &8(t-T_{\mathrm{w}})\Upsilon_{\mathrm{ac}}^4 H^2\log\left(\frac{2H(n_y+n_u)}{\delta}\right) \geq \frac{\sigma_{\mathrm{c}}^4}{16}(t-T_{\mathrm{w}})^2\min\{\sigma_w^4,\sigma_z^4,\sigma_{\eta_{t-1}}^4\} \\
\implies &(t-T_{\mathrm{w}}) \geq \frac{128\Upsilon_{\mathrm{ac}}^4 H^2\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_{\mathrm{c}}^4\min\{\sigma_w^4,\sigma_z^4,\sigma_{\eta_{t-1}}^4\}}.
\end{aligned}$$

Therefore, for all $(t-T_{\mathrm{w}}) \geq T_{\mathrm{ac}}$ where

$$T_{\mathrm{ac}} = \frac{512\Upsilon_{\mathrm{ac}}^4 H^2\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_{\mathrm{c}}^4\min\{\sigma_w^4,\sigma_z^4,\sigma_{\eta_{t-1}}^4\}}, \tag{4-49}$$

we have

$$\begin{aligned}
&\sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top\right) \geq \frac{128\Upsilon_{\mathrm{ac}}^4 H^2\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_{\mathrm{c}}^2\min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta_{t-1}}^2\}} - \frac{64\Upsilon_{\mathrm{ac}}^4 H^2\log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_{\mathrm{c}}^2\min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta_{t-1}}^2\}} \\
\implies &\sigma_{\min}\left(\sum_{i=T_{\mathrm{w}}}^{t-1}\phi_i\phi_i^\top\right) \geq (t-T_{\mathrm{w}})\frac{\sigma_{\mathrm{c}}^2\min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta_{t-1}}^2\}}{8},
\end{aligned} \tag{4-50}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. This completes the proof.

4-3-3 Proof of Theorem 4.3

Recalling from Section 3-1, we have for a single input-output trajectory $\{y_t, u_t\}_{t=0}^t$:

$$Y_t = \Phi_t\mathbf{M}^\top + E_t + N_t,$$

where

$$
\begin{aligned}
\mathbf{M} &= \begin{bmatrix} CF & C\bar{A}F & \dots & C\bar{A}^{H-1}F & CB & C\bar{A}B & \dots & C\bar{A}^{H-1}B \end{bmatrix} \in \mathbb{R}^{n_y \times (n_y+n_u)H}, \\
Y_t &= \begin{bmatrix} y_H & y_{H+1} & \dots & y_t \end{bmatrix}^\top \in \mathbb{R}^{(t-H+1)\times n_y}, \\
\Phi_t &= \begin{bmatrix} \phi_H & \phi_{H+1} & \dots & \phi_t \end{bmatrix}^\top \in \mathbb{R}^{(t-H+1)\times (n_y+n_u)H}, \\
E_t &= \begin{bmatrix} e_H & e_{H+1} & \dots & e_t \end{bmatrix}^\top \in \mathbb{R}^{(t-H+1)\times n_y}, \\
N_t &= \begin{bmatrix} C\bar{A}^H\hat{x}_{0|-1,\Theta} & C\bar{A}^H\hat{x}_{1|0,\Theta} & \dots & C\bar{A}^H\hat{x}_{t-H|t-H-1,\Theta} \end{bmatrix}^\top \in \mathbb{R}^{(t-H+1)\times n_y},
\end{aligned}
$$

where $\bar{A} = A - FC$. Further, recall that

$$
\hat{\mathbf{M}}_{\mathbf{t}}^\top = (\Phi_t^\top \Phi_t + \lambda I)^{-1}\Phi_t^\top Y_t.
$$

This implies

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

Now consider the following:

$$
\begin{aligned}
|\mathrm{Tr}(X(\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})^\top)| &= |\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t) + \mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t) \\
&\quad - \lambda\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)| \\
&\leq |\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t)| + |\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t)| \\
&\quad + \lambda|\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)|,
\end{aligned}
$$

where X is some matrix. Let M_1, M_2, M_3 be three matrices. Using the property $|\mathrm{Tr}(M_1M_2M_3^\top)| \leq \sqrt{\mathrm{Tr}(M_1M_2M_1^\top)\mathrm{Tr}(M_3M_2M_3^\top)}$ for a positive definite M_2, we have

$$
\begin{aligned}
|\mathrm{Tr}(X(\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})^\top)| &\leq \sqrt{\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\mathrm{Tr}(E_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t)} \\
&\quad + \sqrt{\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\mathrm{Tr}(N_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t)} \\
&\quad + \lambda\sqrt{\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)\mathrm{Tr}(\mathbf{M}(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)} \\
&= \sqrt{\mathrm{Tr}(X(\Phi_t^\top\Phi_t + \lambda I)^{-1}X^\top)}\times \\
&\quad \Bigg[\sqrt{\mathrm{Tr}(E_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top E_t)} + \sqrt{\mathrm{Tr}(N_t^\top\Phi_t(\Phi_t^\top\Phi_t + \lambda I)^{-1}\Phi_t^\top N_t)} \\
&\quad + \lambda\sqrt{\mathrm{Tr}(\mathbf{M}(\Phi_t^\top\Phi_t + \lambda I)^{-1}\mathbf{M}^\top)}\Bigg].
\end{aligned}
$$

Substituting $X = (\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})(\Phi_t^\top\Phi_t + \lambda I)$ and $V_t = (\Phi_t^\top\Phi_t + \lambda I)$, we get

$$
\begin{aligned}
|\mathrm{Tr}((\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})V_t(\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})^\top)| \leq \sqrt{\mathrm{Tr}((\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})V_t^\top(\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})^\top)} \times \\
\Bigg[\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} + \sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \\
+ \lambda\sqrt{\mathrm{Tr}(\mathbf{M}V_t^{-1}\mathbf{M}^\top)}\Bigg].
\end{aligned}
$$

Since V_t is a symmetric positive definite matrix, the above expression reduces to

$$
\begin{aligned}
\sqrt{\mathrm{Tr}((\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})V_t(\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M})^\top)} &\leq \Bigg[\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} + \sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \\
&\quad + \lambda\sqrt{\mathrm{Tr}(\mathbf{M}V_t^{-1}\mathbf{M}^\top)}\Bigg] \\
&\leq \Bigg[\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} + \sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \\
&\quad + \lambda\sqrt{||V_t^{-1}||}\sqrt{\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top)}\Bigg] \\
&\leq \Bigg[\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} + \sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \\
&\quad + \sqrt{\lambda}\bar{S}\Bigg],
\end{aligned}
$$

where $||\mathbf{M}||_{\mathrm{F}} \leq \bar{S}$. Now we provide bounds for each of the terms in the above expression.

Bounding $\sqrt{\mathbf{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)}$

Since e_t is $||C\Sigma C^\top + \sigma_z^2 I||$ - sub-Gaussian vector, from Theorem B.1 we have

$$
\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} \leq \sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)}, \tag{4-51}
$$

which holds with a probability of at least $1 - \delta$. Here, $V = \lambda I$. For the sake of convenience, define the event $\mathcal{E}_{E_t}$:

$$
\mathcal{E}_{E_t} := \left\{\sqrt{\mathrm{Tr}(E_t^\top \Phi_t V_t^{-1} \Phi_t^\top E_t)} \leq \sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)}\right\},
$$

which holds with a probability of at least $1 - \delta$.

Bounding $\sqrt{\mathbf{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)}$

Now, we provide a bound for the second term.

$$\begin{aligned}\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} &\leq \frac{1}{\sqrt{\lambda}} ||N_t^\top \Phi_t||_{\mathrm{F}} \\ &\leq \sqrt{\frac{n_y}{\lambda}} \left|\left| \sum_{i=H}^{t} \phi_i (C\bar{A}^H \hat{x}_{i-H|i-H-1,\Theta})^\top \right|\right|.\end{aligned}$$

The last inequality comes from the following property,

$$\sigma_{\max}(X) \leq ||X||_{\mathrm{F}} \leq \sqrt{\min\{m,n\}}\sigma_{\max}(X),$$

for any matrix $X \in \mathbb{R}^{m \times n}$. Now,

$$\begin{aligned}\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} &\leq \sqrt{\frac{n_y}{\lambda}} \left|\left| \sum_{i=H}^{t} \phi_i (C\bar{A}^H \hat{x}_{i-H|i-H-1,\Theta})^\top \right|\right| \\ &\leq (t-H+1)\sqrt{\frac{n_y}{\lambda}} \max_{H \leq i \leq t} \left|\left| \phi_i (C\bar{A}^H \hat{x}_{i-H|i-H-1,\Theta})^\top \right|\right| \\ &\leq t\sqrt{\frac{n_y}{\lambda}} ||C|| \nu^H \max_{H \leq i \leq t} ||\phi_i|| \, ||\hat{x}_{i-H|i-H-1,\Theta}||.\end{aligned}$$

During the warm-up phase, i.e., $t \leq T_{\mathrm{w}} - 1$, we do not have a bound on $||\hat{x}_{t|t-1}||$ since there is no model of the system during this phase. Therefore, during the warm-up phase, we can bound $||\hat{x}_{t|t-1}|| = ||x_t|| \leq X_{\mathrm{w}}$. Further, recall from Lemma A.1 that $||\phi_t|| \leq \Upsilon_{\mathrm{w}}\sqrt{H}$ with a probability of at least $1 - \delta/2$. Therefore, during the warm-up phase, we have

$$\max_{H \leq i \leq T_{\mathrm{w}}-1} ||\phi_i|| \, ||\hat{x}_{i-H|i-H-1,\Theta}|| \leq \Upsilon_{\mathrm{w}} X_{\mathrm{w}} \sqrt{H},$$

which holds with a probability of at least $1 - \delta/2$. Define an event $\mathcal{E}_{\phi,\mathrm{warm}}$:

$$\mathcal{E}_{\phi,\mathrm{warm}} := \left\{ ||\phi_t|| \leq \Upsilon_{\mathrm{w}}\sqrt{H} \right\},$$

which holds with a probability of at least $1 - \delta/2$. Further, define an event $\mathcal{E}_{\text{PE, warm}}$:

$$\mathcal{E}_{\text{PE, warm}} := \left\{ \sigma_{\min}\left(\sum_{i=H}^{T_{\mathrm{w}}-1} \phi_i \phi_i^\top \right) \geq (T_{\mathrm{w}} - H) \frac{\sigma_{\mathrm{o}}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}}{2} \middle| \mathcal{E}_{\phi,\mathrm{warm}} \right\},$$

which holds with a probability of at least $1 - \delta/2$ if $T_{\mathrm{w}} \geq T_{\mathrm{o}}$, where T_{o} is as defined in (A-5). This event is a consequence of Lemma A.2.

Now, during the LBC phase, recall from (4-46) that $||\phi_t|| \leq \Upsilon_{\mathrm{ac}}\sqrt{H}$, with a probability of at least $1 - \delta$ under the event $\mathcal{E}_{\mathbf{M}}$. To recall from (4-16), we have

$$\mathcal{E}_{\mathbf{M}} := \left\{ ||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}|| \leq 1 \right\}, \tag{4-52}$$

which is assumed to hold with high probability. Therefore, during the LBC phase, we have from (4-34):

$$\max_{T_\mathrm{w} \leq i \leq t} ||\phi_i|| \, ||\hat{x}_{i-H|i-H-1,\Theta}|| \leq \Upsilon_\mathrm{ac} X_\mathrm{est,ac}\sqrt{H},$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_\mathbf{M}$. Just like in the proof of Lemma 4.2, δ can be re-parameterised to $\delta \to \delta/2$, in the above result. Define an event $\mathcal{E}_{\phi,\mathrm{ac}}$:

$$\mathcal{E}_{\phi,\mathrm{ac}} \coloneqq \left\{ ||\phi_t|| \leq \Upsilon_\mathrm{ac}\sqrt{H} \middle| \, \mathcal{E}_\mathbf{M} \right\},$$

which holds with a probability of at least $1-\delta/2$. Further, from Lemma 4.2 we can define an event $\mathcal{E}_\text{PE, ac}$, where

$$\mathcal{E}_\text{PE, ac} \coloneqq \left\{ \sigma_\mathrm{min}\left(\sum_{i=T_\mathrm{w}}^{t-1} \phi_i\phi_i^\top \right) \geq (t-T_\mathrm{w})\frac{\sigma_\mathrm{c}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta t}^2\}}{8} \middle| \, \mathcal{E}_{\phi,\mathrm{ac}} \right\},$$

which holds with a probability of at least $1-\delta/2$ if $t - T_\mathrm{w} \geq T_\mathrm{ac}$, where T_ac is as defined in (4-49).

By setting

$$H = \frac{\log\left(\frac{1}{\sqrt{n_y/\lambda}||C|| \max\{\Upsilon_\mathrm{w}X_\mathrm{w}, \Upsilon_\mathrm{ac}X_\mathrm{est,ac}\}T^2}\right)}{\log(\nu)} = \frac{\log\left(\sqrt{n_y/\lambda}||C|| \max\{\Upsilon_\mathrm{w}X_\mathrm{w}, \Upsilon_\mathrm{ac}X_\mathrm{est,ac}\}T^2\right)}{\log(1/\nu)}, \tag{4-53}$$

we have

$$\sqrt{\mathrm{Tr}(N_t^\top \Phi_t V_t^{-1} \Phi_t^\top N_t)} \leq \frac{t}{T^2}\sqrt{H}, \tag{4-54}$$

which holds under the event $\mathcal{E}_{\phi,\mathrm{warm}} \cap \mathcal{E}_{\phi,\mathrm{ac}}$.

Putting things together

Now, combining (4-51) and (4-54), we get

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

Now,

$$\begin{aligned} &\mathrm{Tr}((\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})V_t(\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M})^\top) \geq \sigma_\mathrm{min}(V_t)||\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}||_\mathrm{F}^2 \\ \implies &\sigma_\mathrm{min}(V_t)||\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}||_\mathrm{F}^2 \leq \left(\sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)} + \frac{t}{T^2}\sqrt{H} + \sqrt{\lambda}\bar{S} \right)^2. \end{aligned}$$

From Lemma A.2 and Lemma 4.2, we have

$$\begin{aligned} \sigma_\mathrm{min}(V_t) &= \sigma_\mathrm{min}\left(\sum_{t=H}^{t} \phi_t\phi_t^\top + \lambda I \right) \\ &\geq \sigma_\mathrm{min}\left(\sum_{t=H}^{t} \phi_t\phi_t^\top \right) \\ &\geq (T_\mathrm{w} - H)\frac{\sigma_\mathrm{o}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}}{2} + (t - T_\mathrm{w} + 1)\frac{\sigma_\mathrm{c}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_{\eta t}^2\}}{8}. \end{aligned}$$

Therefore, from Lemma B.4, we have

$$
\begin{aligned}
||\hat{\mathbf{M}}_{\mathbf{t}} - \mathbf{M}||_{\mathrm{F}} &\leq \frac{\sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)} + \frac{t}{T^2}\sqrt{H} + \sqrt{\lambda}\bar{S}}{\sqrt{(T_{\mathrm{w}} - H)\frac{\sigma_{\mathrm{o}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2} + (t - T_{\mathrm{w}} + 1)\frac{\sigma_{\mathrm{c}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta t}^2\}}{8}}} \\
&\leq \frac{\sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \left(\log(1/\delta) + \frac{H(n_u+n_y)}{2}\log\left(\frac{\lambda(n_u+n_y)H + (t-H+1)\max\{\Upsilon_{\mathrm{w}}^2,\Upsilon_{\mathrm{ac}}^2\}}{\lambda(n_u+n_y)H}\right)\right)} + \frac{t}{T^2}\sqrt{H} + \sqrt{\lambda}\bar{S}}{\sqrt{(T_{\mathrm{w}} - H)\frac{\sigma_{\mathrm{o}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2} + (t - T_{\mathrm{w}} + 1)\frac{\sigma_{\mathrm{c}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta t}^2\}}{8}}} \\
&\leq \frac{\sqrt{n_y||C\Sigma C^\top + \sigma_z^2 I|| \left(\log(1/\delta) + \frac{H(n_u+n_y)}{2}\log\left(\frac{\lambda(n_u+n_y)H + T\max\{\Upsilon_{\mathrm{w}}^2,\Upsilon_{\mathrm{ac}}^2\}}{\lambda(n_u+n_y)H}\right)\right)} + \frac{\sqrt{H}}{T} + \sqrt{\lambda}\bar{S}}{\sqrt{t - H + 1}\sqrt{\min\left\{\frac{\sigma_{\mathrm{o}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_u^2\}}{2}, \frac{\sigma_{\mathrm{c}}^2 \min\{\sigma_w^2,\sigma_z^2,\sigma_{\eta t}^2\}}{8}\right\}}},
\end{aligned}
\tag{4-55}
$$

which holds under the event $\mathcal{E}_{\text{PE, warm}} \cap \mathcal{E}_{\text{PE, ac}} \cap \mathcal{E}_{E_t}$. This concludes the proof.

4-3-4 Proof of Theorem 4.4

To recall from (2-16), we are trying to minimise the following definition of regret:

$$
\begin{aligned}
R(T) &= \sum_{t=0}^{T-1} c_t - TJ_*, \text{ where} \\
c_t &= y_t^\top Q y_t + u_t^\top R u_t.
\end{aligned}
$$

Since the LBC policy is deployed in an episodic fashion, as described in Algorithm 2, the regret is also analysed episode-wise, i.e., the cumulative difference between the (sub)optimal cost incurred by the LBC policy and the optimal long-term average expected cost J_*, is upper bounded for every episode. This bound is then summed over the number of episodes to obtain the final regret upper bound. The relation between the long-term average expected cost and the solution to the Lyapunov equation is derived in Appendix A-3. This relation as described in (A-15), is critical for establishing the regret upper bound. To recall, the estimated model parameter is maintained during the length of each episode. Therefore, for our present analysis of the regret, let us denote $\hat{\Theta}_k = \hat{\Theta}$, $\hat{K}_k = \hat{K}$, and $\sigma_{\eta_k}^2 = \sigma_\eta^2$ for the sake of brevity, where k is the episode number.

We will first decompose the cost. Consider the following decomposition of the cost at time step t:

$$
\begin{aligned}
y_t^\top Q y_t + u_t^\top R u_t &= (Cx_t + z_t)^\top Q (Cx_t + z_t) + u_t^\top R u_t \\
&= x_t^\top C^\top Q C x_t + u_t^\top R u_t + 2z_t^\top Q C x_t + z_t^\top Q z_t \\
&= \underbrace{x_t^\top C^\top Q C x_t + \hat{x}_{t|t,\hat{\Theta}}^\top \hat{K}^\top R \hat{K} \hat{x}_{t|t,\hat{\Theta}}}_{c_{t,1}} + \underbrace{\eta_t^\top R \eta_t - 2\eta_t^\top R \hat{K} \hat{x}_{t|t,\hat{\Theta}} + 2z_t^\top Q C x_t + z_t^\top Q z_t}_{c_{t,2}}.
\end{aligned}
\tag{4-56}
$$

We will upper bound $\sum_{t=0}^{t-1} c_{t,1}$ and $\sum_{t=0}^{t-1} c_{t,2}$ separately. To avoid ambiguity in the present analysis, it must be noted that we are not analysing the 0th episode: the time step starts at 0 just for the sake of convenience.

Upper bounding $\sum_{t=0}^{t-1} c_{t,1}$

From (2-6), we have

$$\underbrace{\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}} \end{bmatrix}}_{\bar{x}_t} = \underbrace{\begin{bmatrix} A & -B\hat{K} \\ \hat{L}CA & \left(I-\hat{L}\hat{C}\right)\left(\hat{A}-\hat{B}\hat{K}\right)-\hat{L}CB\hat{K} \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{1}}} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}} \end{bmatrix}$$
$$+ \underbrace{\begin{bmatrix} I & 0 \\ \hat{L}C & \hat{L} \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{2}}} \underbrace{\begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix}}_{\bar{\epsilon}_{t-1}} + \underbrace{\begin{bmatrix} B \\ \left(I-\hat{L}\hat{C}\right)\hat{B}+\hat{L}CB \end{bmatrix}}_{\hat{\mathbf{G}}_{\mathbf{3}}} \eta_{t-1}$$
$$\implies \bar{x}_t = \hat{\mathbf{G}}_{\mathbf{1}}\bar{x}_{t-1} + \hat{\mathbf{G}}_{\mathbf{2}}\bar{\epsilon}_{t-1} + \hat{\mathbf{G}}_{\mathbf{3}}\eta_{t-1}.$$

This implies,

$$\bar{x}_t = \hat{\mathbf{G}}_{\mathbf{1}}^t\bar{x}_0 + \sum_{i=0}^{t-1}\hat{\mathbf{G}}_{\mathbf{1}}^{t-i-1}\hat{\mathbf{G}}_{\mathbf{2}}\bar{\epsilon}_i + \sum_{i=0}^{t-1}\hat{\mathbf{G}}_{\mathbf{1}}^{t-i-1}\hat{\mathbf{G}}_{\mathbf{3}}\eta_i.$$

For simplicity of exposition, let us define for $l \in \mathbb{N}$ and $j, l \geq i$,

$$\mathrm{Col}_{i,j}(A) := \begin{bmatrix} \mathbb{I}_{i\geq 1}A^{i-1} \\ \mathbb{I}_{i\geq 2}A^{i-2} \\ \cdot \\ \cdot \\ \cdot \\ \mathbb{I}_{i\geq j}A^{i-j} \end{bmatrix}, \quad \mathrm{Toep}_{i,j,l}(A) := \begin{bmatrix} A^i\mathbb{I}_{i\geq 0} & A^{i+1}\mathbb{I}_{i\geq -1} & \dots & A^{i+l}\mathbb{I}_{i\geq -l} \\ A^{i-1}\mathbb{I}_{i\geq 1} & A^i\mathbb{I}_{i\geq 0} & \dots & A^{i+l-1}\mathbb{I}_{i\geq 1-l} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ A^{i-j}\mathbb{I}_{i\geq j} & A^{i-j+1}\mathbb{I}_{i\geq j-1} & \dots & A^{i+l-j}\mathbb{I}_{i\geq j-l} \end{bmatrix},$$
$$\text{and } \mathrm{diag}_t(A) := I_t \otimes A, \tag{4-57}$$

where $\mathbb{I}$ is the indicator function and with a slight abuse of notations, we have I_t as the identity matrix with t rows. This implies

$$\underbrace{\begin{bmatrix} \bar{x}_{t-1} \\ \bar{x}_{t-2} \\ \cdot \\ \cdot \\ \cdot \\ \bar{x}_0 \end{bmatrix}}_{\bar{x}_{[t-1:0]}} = \begin{bmatrix} \hat{\mathbf{G}}_{\mathbf{1}}^{t-1} \\ \hat{\mathbf{G}}_{\mathbf{1}}^{t-2} \\ \cdot \\ \cdot \\ \cdot \\ I \end{bmatrix}\bar{x}_0 + \begin{bmatrix} 0 & I & \hat{\mathbf{G}}_{\mathbf{1}} & \hat{\mathbf{G}}_{\mathbf{1}}^2 & . & . & . & \hat{\mathbf{G}}_{\mathbf{1}}^{t-2} \\ 0 & 0 & I & \hat{\mathbf{G}}_{\mathbf{1}} & . & . & . & \hat{\mathbf{G}}_{\mathbf{1}}^{t-3} \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ 0 & 0 & 0 & 0 & . & . & . & I \\ 0 & 0 & 0 & 0 & . & . & . & 0 \end{bmatrix} I_t \otimes \hat{\mathbf{G}}_{\mathbf{2}} \underbrace{\begin{bmatrix} \bar{\epsilon}_{t-1} \\ \bar{\epsilon}_{t-2} \\ \cdot \\ \cdot \\ \cdot \\ \bar{\epsilon}_0 \end{bmatrix}}_{\bar{\epsilon}_{[t-1:0]}}$$
$$+ \begin{bmatrix} 0 & I & \hat{\mathbf{G}}_{\mathbf{1}} & \hat{\mathbf{G}}_{\mathbf{1}}^2 & . & . & . & \hat{\mathbf{G}}_{\mathbf{1}}^{t-2} \\ 0 & 0 & I & \hat{\mathbf{G}}_{\mathbf{1}} & . & . & . & \hat{\mathbf{G}}_{\mathbf{1}}^{t-3} \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ \cdot & & & & & & & \\ 0 & 0 & 0 & 0 & . & . & . & I \\ 0 & 0 & 0 & 0 & . & . & . & 0 \end{bmatrix} I_t \otimes \hat{\mathbf{G}}_{\mathbf{3}} \underbrace{\begin{bmatrix} \eta_{t-1} \\ \eta_{t-2} \\ \cdot \\ \cdot \\ \cdot \\ \eta_0 \end{bmatrix}}_{\eta_{[t-1:0]}}$$

$$\begin{aligned}\implies \bar{x}_{[t-1:0]} = \mathrm{Col}_{t,t}(\hat{\mathbf{G}}_{\mathbf{1}})\bar{x}_0 &+ \mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\hat{\mathbf{G}}_{\mathbf{2}})\bar{\epsilon}_{[t-1:0]} \\ &+ \mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\hat{\mathbf{G}}_{\mathbf{3}})\eta_{[t-1:0]}.\end{aligned} \tag{4-58}$$

Finally, $\sum_{t=0}^{T-1} c_{t,1}$ can be decomposed as:

$$\begin{aligned}\sum_{t=0}^{t-1} c_{t,1} &= \sum_{t=0}^{t-1} x_t^\top C^\top Q C x_t + \hat{x}_{t|t,\hat{\Theta}}^\top \hat{K}^\top R \hat{K} \hat{x}_{t|t,\hat{\Theta}} \\ &= \sum_{t=0}^{t-1} \bar{x}_t^\top \bar{\mathbf{W}} \bar{x}_t \\ &= \bar{x}_{[t-1:0]}^\top \mathrm{diag}_t(\bar{\mathbf{W}}) \bar{x}_{[t-1:0]} \\ &= \bar{x}_0^\top \underbrace{\mathrm{Col}_{t,t}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Col}_{t,t}(\hat{\mathbf{G}}_{\mathbf{1}})}_{\Lambda_{\bar{x}_0}} \bar{x}_0 \\ &\quad + \bar{\epsilon}_{[t-1:0]}^\top \underbrace{\mathrm{diag}_t^\top(\hat{\mathbf{G}}_{\mathbf{2}})\mathrm{Toep}_{-1,t-1,t-1}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\hat{\mathbf{G}}_{\mathbf{2}})}_{\Lambda_{\bar{\epsilon}}} \bar{\epsilon}_{[t-1:0]} \\ &\quad + \eta_{[t-1:0]}^\top \underbrace{\mathrm{diag}_t^\top(\hat{\mathbf{G}}_{\mathbf{3}})\mathrm{Toep}_{-1,t-1,t-1}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\hat{\mathbf{G}}_{\mathbf{3}})}_{\Lambda_{\eta}} \eta_{[t-1:0]} \\ &\quad + 2\bar{\epsilon}_{[t-1:0]}^\top \underbrace{\mathrm{diag}_t^\top(\hat{\mathbf{G}}_{\mathbf{2}})\mathrm{Toep}_{-1,t-1,t-1}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Col}_{t,t}(\hat{\mathbf{G}}_{\mathbf{1}})}_{\Lambda_{\mathrm{cross},1}} \bar{x}_0 \\ &\quad + 2\eta_{[t-1:0]}^\top \underbrace{\mathrm{diag}_t^\top(\hat{\mathbf{G}}_{\mathbf{3}})\mathrm{Toep}_{-1,t-1,t-1}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Col}_{t,t}(\hat{\mathbf{G}}_{\mathbf{1}})}_{\Lambda_{\mathrm{cross},2}} \bar{x}_0 \\ &\quad + 2\bar{\epsilon}_{[t-1:0]}^\top \underbrace{\mathrm{diag}_t^\top(\hat{\mathbf{G}}_{\mathbf{2}})\mathrm{Toep}_{-1,t-1,t-1}^\top(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\mathrm{diag}_t(\hat{\mathbf{G}}_{\mathbf{3}})}_{\Lambda_{\mathrm{cross},3}} \eta_{[t-1:0]},\end{aligned} \tag{4-59}$$

where $\bar{\mathbf{W}} = \begin{bmatrix} C^\top Q C & 0 \\ 0 & \hat{K}^\top R \hat{K} \end{bmatrix}$. We will now upper bound each of the above terms individually.

Bounding $\bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\bar{\epsilon}} \bar{\epsilon}_{[t-1:0]}$

This term can be upper-bounded using the Hanson-Wright inequality. To do so, firstly we require the following bounds. From Lemma B.12 we have

$$\begin{aligned}||\Lambda_{\bar{\epsilon}}|| &\leq \left|\left|\hat{\mathbf{G}}_{\mathbf{2}}\right|\right|^2 \left|\left|\bar{\mathbf{W}}\right|\right| \left|\left|\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_{\mathbf{1}})\right|\right|^2 \\ &\leq \left|\left|\hat{\mathbf{G}}_{\mathbf{2}}\right|\right|^2 \left|\left|\bar{\mathbf{W}}\right|\right| \left|\left|\hat{\mathbf{G}}_{\mathbf{1}}\right|\right|_{\mathcal{H}_\infty}^2 .\end{aligned}$$

The above bound depends on the estimated system parameter $\hat{\Theta}$. To use the Hanson-Wright inequality, we need a bound on $||\Lambda_{\bar{\epsilon}}||$ that is not a random variable. Now,

$$\left|\left|\hat{\mathbf{G}}_{\mathbf{2}}\right|\right| \leq \left|\left|\hat{\mathbf{G}}_{\mathbf{2}} - \mathbf{G}_{\mathbf{2}}\right|\right| + ||\mathbf{G}_{\mathbf{2}}||$$

$$\begin{aligned}
&\leq \left|\left|\begin{bmatrix} 0 & 0 \\ (\hat{L}-L)C & (\hat{L}-L) \end{bmatrix}\right|\right| + \left|\left|\begin{bmatrix} I & 0 \\ LC & L \end{bmatrix}\right|\right| \\
&\leq c_{\hat{\mathbf{G}}_\mathbf{2}}.
\end{aligned}$$

The above bound exists under the event $\mathcal{E}_\mathbf{M}$. This event, which was defined in (4-16), is recalled here for convenience:

$$\mathcal{E}_\mathbf{M} = \left\{ ||\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}|| \leq 1 \right\},$$

where $\mathbb{P}\{\mathcal{E}_\mathbf{M}\} \geq 1 - 2\delta$ if $t \geq T_\mathrm{w} \geq T_\mathbf{M}$ with $T_\mathbf{M}$ as defined in Theorem 4.1. This event is a direct consequence of Theorem 4.3. Under this event, we have

1. $\Theta \in \mathcal{C}_A(t) \times \mathcal{C}_B(t) \times \mathcal{C}_C(t) \times \mathcal{C}_L(t)$ for all $t \geq T_\mathrm{w}$.
2. $||\hat{C}_t - C||, ||\hat{B}_t - B||, ||\hat{F}_t - F|| \leq \beta_B(T_\mathrm{w}) = 1$ when $T_\mathrm{w} \geq T_B$.
3. $\left|\left|\hat{A}_t - A\right|\right| \leq \beta_A(T_\mathrm{w}) = \sigma_{n_x}(A)/2$ when $T_\mathrm{w} \geq T_A$.
4. $\left|\left|\hat{L}_t - L\right|\right| \leq \beta_L(T_\mathrm{w})$,

where the similarity transformation matrix $\mathbf{T} = I$ without loss of generality. Similarly, we can bound $\left|\left|\hat{\mathbf{G}}_\mathbf{1}\right|\right|_{\mathcal{H}_\infty} \leq c_{\hat{\mathbf{G}}_\mathbf{1}}$ under the event $\mathcal{E}_\mathbf{M}$. For Hanson-Wright inequality, we also require the following bound, which exists under the event $\mathcal{E}_\mathbf{M}$:

$$\begin{aligned}
||\Lambda_{\bar{\epsilon}}||_\mathrm{F} &= \sqrt{\mathrm{Tr}(\Lambda_{\bar{\epsilon}}\Lambda_{\bar{\epsilon}}^\top)} \\
&\lesssim \sqrt{t(n_x + n_y)} := \bar{c}_{\Lambda_{\bar{\epsilon}}}.
\end{aligned}$$

Consider the following:

$$\begin{aligned}
\left|\left|\bar{\mathbf{W}}\right|\right| &= \left|\left|\begin{bmatrix} C^\top Q C & 0 \\ 0 & \hat{K}^\top R \hat{K} \end{bmatrix}\right|\right| \\
&\leq \left|\left|C^\top Q C + \hat{K}^\top R \hat{K}\right|\right| \\
&\leq ||C||^2 \, ||Q|| + \Gamma^2 \, ||R|| := c_{\bar{\mathbf{W}}}.
\end{aligned}$$

Finally, from the Hanson-Wright inequality as defined in Theorem B.3, we have

$$\mathbb{P}\left\{ \bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\bar{\epsilon}} \bar{\epsilon}_{[t-1:0]} - \mathbb{E}\left[\bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\bar{\epsilon}} \bar{\epsilon}_{[t-1:0]} \right] > t \right\} \leq 2 \exp\left[-c \min\left(\frac{t^2}{a^4 \bar{c}_{\Lambda_{\bar{\epsilon}}}^2}, \frac{t}{a^2 c_{\hat{\mathbf{G}}_\mathbf{2}}^2 c_{\hat{\mathbf{G}}_\mathbf{1}}^2 c_{\bar{\mathbf{W}}}} \right) \right],$$

where c is an absolute positive constant. Since, $\bar{\epsilon}_{[t-1:0]}$ consists of Gaussian random variables, the constant a exists. Now we can simplify the above expression as follows. Let,

$$\begin{aligned}
\frac{\delta}{2} &= \exp\left[-c\min\left(\frac{t^2}{a^4\bar{c}^2_{\Lambda_{\bar{\epsilon}}}}, \frac{t}{a^2c^2_{\hat{\mathbf{G}}_2}c^2_{\hat{\mathbf{G}}_1}c_{\bar{\mathbf{W}}}}\right)\right] \\
\frac{1}{c}\log\left(\frac{2}{\delta}\right) &= \min\left(\frac{t^2}{a^4\bar{c}^2_{\Lambda_{\bar{\epsilon}}}}, \frac{t}{a^2c^2_{\hat{\mathbf{G}}_2}c^2_{\hat{\mathbf{G}}_1}c_{\bar{\mathbf{W}}}}\right) \\
\implies t &= a^2\bar{c}_{\Lambda_{\bar{\epsilon}}}\sqrt{\frac{1}{c}\log\left(\frac{2}{\delta}\right)} \text{ or } t = \frac{a^2c^2_{\hat{\mathbf{G}}_2}c^2_{\hat{\mathbf{G}}_1}c_{\bar{\mathbf{W}}}}{c}\log\left(\frac{2}{\delta}\right).
\end{aligned}$$

This implies with a probability of at least $1-\delta$,

$$\begin{aligned}
\bar{\epsilon}^\top_{[t-1:0]}\Lambda_{\bar{\epsilon}}\bar{\epsilon}_{[t-1:0]} &\leq \mathbb{E}\left[\bar{\epsilon}^\top_{[t-1:0]}\Lambda_{\bar{\epsilon}}\bar{\epsilon}_{[t-1:0]}\right] + \mathcal{O}\left(\bar{c}_{\Lambda_{\bar{\epsilon}}}\sqrt{\log\left(\frac{2}{\delta}\right)} + c^2_{\hat{\mathbf{G}}_2}c^2_{\hat{\mathbf{G}}_1}c_{\bar{\mathbf{W}}}\log\left(\frac{2}{\delta}\right)\right) \\
\implies \bar{\epsilon}^\top_{[t-1:0]}\Lambda_{\bar{\epsilon}}\bar{\epsilon}_{[t-1:0]} &\leq \mathrm{Tr}\left(\Lambda_{\bar{\epsilon}}\mathrm{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right) \\
&\quad + \mathcal{O}\left(\bar{c}_{\Lambda_{\bar{\epsilon}}}\sqrt{\log\left(\frac{2}{\delta}\right)} + c^2_{\hat{\mathbf{G}}_2}c^2_{\hat{\mathbf{G}}_1}c_{\bar{\mathbf{W}}}\log\left(\frac{2}{\delta}\right)\right).
\end{aligned} \tag{4-60}$$

Bounding Tr $\left(\Lambda_{\bar{\epsilon}}\mathbf{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right)$

$$\begin{aligned}
&\mathrm{Tr}\left(\Lambda_{\bar{\epsilon}}\mathrm{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right) \\
&= \mathrm{Tr}\left(\mathrm{diag}_t^\top(\hat{\mathbf{G}}_2)\mathrm{Toep}^\top_{-1,t-1,t-1}(\hat{\mathbf{G}}_1)\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_1)\mathrm{diag}_t(\hat{\mathbf{G}}_2)\mathrm{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right).
\end{aligned}$$

Now, consider the following:

$$\begin{aligned}
&\mathrm{Tr}\left(\mathrm{Toep}^\top_{-1,t-1,t-1}(\hat{\mathbf{G}}_1)\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_1)\right) \\
&= \mathrm{Tr}\left(\left[\mathrm{Col}_{0,t}(\hat{\mathbf{G}}_1) \quad \dots \quad \mathrm{Col}_{t-1,t}(\hat{\mathbf{G}}_1)\right]^\top \mathrm{diag}_t(\bar{\mathbf{W}})\left[\mathrm{Col}_{0,t}(\hat{\mathbf{G}}_1) \quad \dots \quad \mathrm{Col}_{t-1,t}(\hat{\mathbf{G}}_1)\right]\right) \\
&= \sum_{i=0}^{t-1}\mathrm{Col}^\top_{i,t}(\hat{\mathbf{G}}_1)\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Col}_{i,t}(\hat{\mathbf{G}}_1) \\
&\leq t\cdot\mathrm{dlyap}\left(\hat{\mathbf{G}}_1, \bar{\mathbf{W}}\right),
\end{aligned}$$

where the last inequality comes from Corollary B.11. Further, we have the following relation. For any positive semi-definite matrices X, Y, and any matrix P, if $X \leq Y$ then, $P^\top XP \leq P^\top YP \implies \mathrm{Tr}(P^\top XP) \leq \mathrm{Tr}(P^\top YP)$. Further, for another diagonal matrix Z, $\mathrm{Tr}(P^\top XPZ) = \mathrm{Tr}(Z^{1/2}P^\top XPZ^{1/2}) \leq \mathrm{Tr}(Z^{1/2}P^\top YPZ^{1/2})$. Considering these relations, we have

$$
\begin{aligned}
&\mathrm{Tr}\left(\Lambda_{\bar{\epsilon}}\mathrm{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right)\\
&= \mathrm{Tr}\left(\mathrm{diag}_t^\top(\hat{\mathbf{G}}_\mathbf{2})\mathrm{Toep}^\top_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})\mathrm{diag}_t(\hat{\mathbf{G}}_\mathbf{2})\mathrm{diag}_t\left(\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right)\\
&\leq t\mathrm{Tr}\left(\hat{\mathbf{G}}_\mathbf{2}^\top\mathrm{dlyap}\left(\hat{\mathbf{G}}_\mathbf{1},\bar{\mathbf{W}}\right)\hat{\mathbf{G}}_\mathbf{2}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\\
&= t\mathrm{Tr}\left(\hat{\mathbf{G}}_\mathbf{2}^\top S\hat{\mathbf{G}}_\mathbf{2}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\\
&= tJ_s(\hat{\Theta}),
\end{aligned}
$$

where the last equality comes from (A-15) and to recall, $J_s(\hat{\Theta})$ is an alternate formulation of the LQG control problem defined in (A-14) as

$$
\begin{aligned}
J_s(\hat{\Theta}) &= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_c x_t + u_t^\top R u_t\right]\\
&= \lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\tilde{x}_t^\top \underbrace{\begin{bmatrix}Q_c & 0\\ 0 & \hat{K}^\top R\hat{K}\end{bmatrix}}_{\bar{\mathbf{W}}}\tilde{x}_t\right] \quad \text{s.t.}\\
&\quad x_{t+1} = Ax_t + Bu_t + w_t,\ \ w_t\sim\mathcal{N}(0,\sigma_w^2 I),\\
&\quad y_t = Cx_t + z_t,\ \ z_t\sim\mathcal{N}(0,\sigma_z^2 I),\\
&\quad \hat{x}_{t|t,\hat{\Theta}} = (I-\hat{L}\hat{C})\hat{x}_{t|t-1,\hat{\Theta}} + \hat{L}y_t,\\
&\quad \hat{x}_{t+1|t,\hat{\Theta}} = \hat{A}\hat{x}_{t|t,\hat{\Theta}} + \hat{B}u_t,\\
&\quad u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}},
\end{aligned}
$$

where $Q_c = C^\top QC$, $\hat{K}$ stabilises the true system and $\hat{A}-\hat{F}\hat{C}$ is asymptotically stable.

Putting things together

From (4-60), we have the following under the event $\mathcal{E}_\mathbf{M}$:

$$
\begin{aligned}
\bar{\epsilon}^\top_{[t-1:0]}\Lambda_{\bar{\epsilon}}\bar{\epsilon}_{[t-1:0]} &\leq tJ_s(\hat{\Theta}) + \mathcal{O}\left(\sqrt{t(n_x+n_y)\log\left(\frac{2}{\delta}\right)} + \left(c^2_{\hat{\mathbf{G}}_\mathbf{2}}c^2_{\hat{\mathbf{G}}_\mathbf{1}}c_{\bar{\mathbf{W}}}\right)\log\left(\frac{2}{\delta}\right)\right)\\
&\leq tJ_s(\hat{\Theta}) + \mathcal{O}\left[\left(\sqrt{t(n_x+n_y)\log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_\mathbf{2}}c^2_{\hat{\mathbf{G}}_\mathbf{1}}c_{\bar{\mathbf{W}}}\right)\right],
\end{aligned}
\tag{4-61}
$$

which holds with a probability of at least $1-\delta$.

Bounding $\eta^\top_{[t-1:0]}\Lambda_\eta\eta_{[t-1:0]}$

This term can also be upper-bounded using the Hanson-Wright inequality. Under the event $\mathcal{E}_\mathbf{M}$, and from Lemma B.12, we have

$$\begin{aligned}||\Lambda_\eta|| &\leq \left|\left|\hat{\mathbf{G}}_\mathbf{3}\right|\right|^2 \left|\left|\bar{\mathbf{W}}\right|\right| \left|\left|\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})\right|\right|^2 \\ &\leq c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}.\end{aligned}$$

Now from Theorem B.3, we have

$$\eta^\top_{[t-1:0]}\Lambda_\eta\eta_{[t-1:0]} \leq \mathrm{Tr}\left(\Lambda_\eta \mathrm{diag}_t\left(\sigma^2_\eta I\right)\right) + \mathcal{O}\left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}} \log\left(\frac{2}{\delta}\right)\right), \quad (4\text{-}62)$$

which holds with a probability of at least $1 - \delta$.

Bounding Tr $\left(\Lambda_\eta \mathbf{diag}_t\left(\sigma^2_\eta I\right)\right)$

It is a standard fact that, for any positive semi-definite matrix X and any matrix Y,

$$\mathrm{Tr}(XY) \leq \mathrm{Tr}(X)||Y||.$$

Now from Corollary B.11, we have the following under the event $\mathcal{E}_\mathbf{M}$:

$$\begin{aligned}\mathrm{Tr}\left(\Lambda_\eta \mathrm{diag}_t\left(\sigma^2_\eta I\right)\right) &\leq \mathrm{Tr}\left(\Lambda_\eta\right)\sigma^2_\eta \\ &= \mathrm{Tr}\left(\mathrm{diag}_t(\hat{\mathbf{G}}_\mathbf{3})^\top \mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})^\top \mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})\mathrm{diag}_t(\hat{\mathbf{G}}_\mathbf{3})\right)\sigma^2_\eta \\ &\leq t\mathrm{Tr}\left(\hat{\mathbf{G}}_\mathbf{3}^\top \mathrm{dlyap}\left(\hat{\mathbf{G}}_\mathbf{1}, \bar{\mathbf{W}}\right)\hat{\mathbf{G}}_\mathbf{3}\right)\sigma^2_\eta \\ &\leq tn_u \left|\left|\hat{\mathbf{G}}_\mathbf{3}^\top S\hat{\mathbf{G}}_\mathbf{3}\right|\right|\sigma^2_\eta \\ &\leq tn_u c^2_{\hat{\mathbf{G}}_\mathbf{3}} \left|\left|S\right|\right| \sigma^2_\eta.\end{aligned}$$

Putting things together

Under the event $\mathcal{E}_\mathbf{M}$, and from (4-62), we have

$$\begin{aligned}\eta^\top_{[t-1:0]}\Lambda_\eta\eta_{[t-1:0]} &\leq tn_u c^2_{\hat{\mathbf{G}}_\mathbf{3}} \left|\left|S\right|\right| \sigma^2_\eta + \mathcal{O}\left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + \left(c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}\right) \log\left(\frac{2}{\delta}\right)\right) \\ &\leq tn_u c^2_{\hat{\mathbf{G}}_\mathbf{3}} \left|\left|S\right|\right| \sigma^2_\eta + \mathcal{O}\left[\left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}\right)\right],\end{aligned} \quad (4\text{-}63)$$

which holds with a probability of at least $1 - \delta$.

Bounding $\bar{x}_0^\top \Lambda_{\bar{x}_0}\bar{x}_0$

From Corollary B.11, we have

$$\begin{aligned}\bar{x}_0^\top \Lambda_{\bar{x}_0}\bar{x}_0 &= \bar{x}_0^\top \mathrm{Col}^\top_{t,t}(\hat{\mathbf{G}}_\mathbf{1})\mathrm{diag}_t(\bar{\mathbf{W}})\mathrm{Col}^\top_{t,t}(\hat{\mathbf{G}}_\mathbf{1})\bar{x}_0 \\ &\leq \bar{x}_0^\top \mathrm{dlyap}\left(\hat{\mathbf{G}}_\mathbf{1}, \bar{\mathbf{W}}\right)\bar{x}_0 \\ &= \bar{x}_0^\top S\bar{x}_0 \\ &\leq ||\bar{x}_0||^2||S|| \\ &\leq \left(X^2_{\mathrm{ac}} + \bar{\chi}^2\right)||S||,\end{aligned} \quad (4\text{-}64)$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}}$. The last inequality comes from Lemma 4.1. It must be noted that although $x_0 = 0$ is assumed, we do not consider that for the above bound. The reason is that x_0 here represents the state of the system at the start of each episode. For instance, the above bound will make sense if we consider the 2nd episode. For the sake of convenience, define the following event:

$$\mathcal{E}_x \coloneqq \{ ||x_t|| \leq X_{\text{ac}} | \, \mathcal{E}_{\mathbf{M}} \} ,$$

which holds with a probability of at least $1 - \delta$. Under the event $\mathcal{E}_x$, the bound in (4-64) becomes deterministic.

Bounding $2\bar{\epsilon}_{[t-1:0]}^{\top} \Lambda_{\textbf{cross,1}} \bar{x}_0$

Firstly, notice that

$$\begin{aligned} \mathbb{E}\left[2\bar{\epsilon}_{[t-1:0]}^{\top} \Lambda_{\text{cross},1} \bar{x}_0\right] &= 0, \\ \text{Var}\left(2\bar{\epsilon}_{[t-1:0]}^{\top} \Lambda_{\text{cross},1} \bar{x}_0\right) &= 4\bar{x}_0^{\top} \Lambda_{\text{cross},1}^{\top} \text{diag}_t \left(\begin{bmatrix} \sigma_w^2 I & 0 \\ 0 & \sigma_z^2 I \end{bmatrix} \right) \Lambda_{\text{cross},1} \bar{x}_0. \end{aligned}$$

It can be verified that there exists a matrix X such that $XX^{\top} = \Lambda_{\bar{\epsilon}}$ and a matrix Y such that $YY^{\top} = \Lambda_{\bar{x}_0}$ to obtain $\Lambda_{\text{cross},1} = XY^{\top}$ [58].

$$\begin{aligned} ||\Lambda_{\text{cross},1} \bar{x}_0|| &= \sqrt{\bar{x}_0^{\top} Y X^{\top} X Y^{\top} \bar{x}_0} \\ &\leq \sqrt{||X^{\top} X|| \cdot \bar{x}_0^{\top} Y Y^{\top} \bar{x}_0} \\ &\leq \sqrt{||\Lambda_{\bar{\epsilon}}|| \cdot \bar{x}_0^{\top} S \bar{x}_0}, \end{aligned}$$

which holds since $\Lambda_{\bar{\epsilon}}$ is symmetric positive semi-definite. The last inequality comes from (4-64). Now, from arithmetic mean-geometric mean inequality, we have the following under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$:

$$\begin{aligned} ||\Lambda_{\text{cross},1} \bar{x}_0|| &\lesssim ||\Lambda_{\bar{\epsilon}}|| + \bar{x}_0^{\top} S \bar{x}_0 \\ &\lesssim c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} + \bar{x}_0^{\top} S \bar{x}_0 \\ &\lesssim c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} + ||\bar{x}_0||^2 ||S|| \\ &\lesssim c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} + \left(X_{\text{ac}}^2 + \bar{\chi}^2\right) ||S||. \end{aligned}$$

Now observe that $2\bar{\epsilon}_{[t-1:0]}^{\top} \Lambda_{\text{cross},1} \bar{x}_0$ is $c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} + (X_{\text{ac}}^2 + \bar{\chi}^2) ||S||$ - Lipschitz. Using Lemma B.5, we have

$$\begin{aligned} & 2\bar{\epsilon}_{[t-1:0]}^{\top} \Lambda_{\text{cross},1} \bar{x}_0 \\ & \lesssim 2\sqrt{2 \max\{\sigma_w^2, \sigma_z^2\}} \left(c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} \sqrt{\log\left(\frac{2}{\delta}\right)} + \left(X_{\text{ac}}^2 + \bar{\chi}^2\right) ||S|| \sqrt{\log\left(\frac{2}{\delta}\right)} \right) \\ & \lesssim \max\{\sigma_w, \sigma_z\} \left(c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} \sqrt{\log\left(\frac{2}{\delta}\right)} + \left(X_{\text{ac}}^2 + \bar{\chi}^2\right) ||S|| \sqrt{\log\left(\frac{2}{\delta}\right)} \right), \end{aligned} \tag{4-65}$$

which holds with a probability of at least $1 - \delta$ under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$.

Bounding $2\eta_{[t-1:0]}^\top \Lambda_{\textbf{cross,2}} \bar{x}_0$

In a similar fashion to the previous cross-term, we obtain the following bound under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$:

$$\begin{aligned}&2\eta_{[t-1:0]}^\top \Lambda_{\text{cross},2} \bar{x}_0 \\ &\lesssim \sigma_\eta c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} \sqrt{\log\left(\frac{2}{\delta}\right)} + \sigma_\eta \left(X_{\text{ac}}^2 + \bar{\chi}^2\right) ||S|| \sqrt{\log\left(\frac{2}{\delta}\right)},\end{aligned} \tag{4-66}$$

which holds with a probability of at least $1-\delta$.

Bounding $2\bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\textbf{cross,3}} \eta_{[t-1:0]}$

Recalling from (4-59), we have

$$\begin{aligned}&2\bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\text{cross},3} \eta_{[t-1:0]} \\ &= 2\bar{\epsilon}_{[t-1:0]}^\top \text{diag}_t(\hat{\mathbf{G}}_\mathbf{2})^\top \text{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1})^\top \text{diag}_t(\bar{\mathbf{W}}) \text{Toep}_{-1,t-1,t-1}(\hat{\mathbf{G}}_\mathbf{1}) \text{diag}_t(\hat{\mathbf{G}}_\mathbf{3}) \eta_{[t-1:0]}.\end{aligned}$$

Now in a similar fashion as the previous cross terms, we have the following from the arithmetic mean-geometric mean inequality:

$$\begin{aligned}\left|\left|\Lambda_{\text{cross},3} \eta_{[t-1:0]}\right|\right| &\leq \sqrt{||\Lambda_{\bar{\epsilon}}|| \cdot \eta_{[t-1:0]}^\top \Lambda_\eta \eta_{[t-1:0]}} \\ &\lesssim ||\Lambda_{\bar{\epsilon}}|| + \eta_{[t-1:0]}^\top \Lambda_\eta \eta_{[t-1:0]}.\end{aligned}$$

From (4-63), we have

$$\begin{aligned}&\left|\left|\Lambda_{\text{cross},3} \eta_{[t-1:0]}\right|\right| \\ &\lesssim c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} + t n_u c_{\hat{\mathbf{G}}_3}^2 ||S|| \sigma_\eta^2 + \left(\sqrt{t n_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) \left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right),\end{aligned}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$. The above bound is indeed deterministic since t here represents the number of time steps in a particular episode, which is known a priori. Now using Lemma B.5 on the entire term, we have

$$\begin{aligned}&2\bar{\epsilon}_{[t-1:0]}^\top \Lambda_{\text{cross},3} \eta_{[t-1:0]} \\ &\lesssim \max\{\sigma_w, \sigma_z\} c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}} \sqrt{\log\left(\frac{2}{\delta}\right)} + t n_u c_{\hat{\mathbf{G}}_3}^2 ||S|| \sigma_\eta^2 \max\{\sigma_w, \sigma_z\} \sqrt{\log\left(\frac{2}{\delta}\right)} \\ &\quad + \left(\sqrt{t n_u} \log\left(\frac{2}{\delta}\right) + \log^2\left(\frac{2}{\delta}\right)\right) \left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \max\{\sigma_w, \sigma_z\},\end{aligned} \tag{4-67}$$

which holds with a probability of at least $1-2\delta$ under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$.

Putting things together

Finally, we have the following bound under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$:

$$\begin{aligned}
\sum_{t=0}^{t-1} c_{t,1} \lesssim\, & tJ_s(\hat{\Theta}) + tn_u c^2_{\hat{\mathbf{G}}_\mathbf{3}} \left|\left|S\right|\right| \sigma_\eta^2 \left(1 + \max\{\sigma_w, \sigma_z\}\sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\
& + \max\{\sigma_w, \sigma_z, \sigma_\eta\}\left(X_{\mathrm{ac}}^2 + \bar{\chi}^2\right)\left|\left|S\right|\right|\left(1 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\
& + \left(\sqrt{\left(t(n_x + n_y) + \max\{\sigma_w^2, \sigma_z^2\}\right)\log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_\mathbf{2}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}\right) \\
& + \left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + \left(\sqrt{tn_u \max\{\sigma_w^2, \sigma_z^2\}} + 1\right)\log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}\right) \\
& + \left(\sigma_\eta\sqrt{\log\left(\frac{2}{\delta}\right)} + \max\{\sigma_w, \sigma_z\}\log^2\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_\mathbf{3}} c^2_{\hat{\mathbf{G}}_\mathbf{1}} c_{\bar{\mathbf{W}}}\right),
\end{aligned} \tag{4-68}$$

which holds with a probability of at least $1 - 6\delta$.

Upper bounding $\sum_{t=0}^{t-1} c_{t,2}$

To recall, from (4-56) we have

$$\sum_{t=0}^{t-1} c_{t,2} = \eta_t^\top R \eta_t - 2\eta_t^\top R\hat{K}\hat{x}_{t|t,\hat{\Theta}} + 2z_t^\top QCx_t + z_t^\top Qz_t.$$

Upper bounding $\sum_{t=0}^{t-1} 2z_t^\top QCx_t$

Bounding this term follows analogously to the previous cross-terms. Under the event $\mathcal{E}_x$, we have:

$$\begin{aligned}
\left|\left|\mathrm{diag}_t(QC)x_{[t-1:0]}\right|\right| &\leq \sqrt{\left|\left|Q\right|\right| \cdot x_{[t-1:0]}^\top \mathrm{diag}_t(C^\top QC)x_{[t-1:0]}} \\
&\leq \sqrt{\left|\left|Q\right|\right| \cdot \left|\left|\mathrm{diag}_t(C^\top QC)\right|\right| \cdot \left|\left|x_{[t-1:0]}\right|\right|^2} \\
&\leq \sqrt{\left|\left|Q\right|\right| \cdot \left|\left|C^\top QC\right|\right|} \cdot \sqrt{t}X_{\mathrm{ac}}.
\end{aligned}$$

Now from Lemma B.5, we have

$$\sum_{t=0}^{t-1} 2z_t^\top QCx_t \lesssim \sqrt{t}\sigma_z \left|\left|Q\right|\right| \cdot \left|\left|C\right|\right| X_{\mathrm{ac}}\sqrt{\log\left(\frac{2}{\delta}\right)}, \tag{4-69}$$

which holds with a probability of at least $1 - \delta$ under the event $\mathcal{E}_x$.

Upper bounding $-\sum_{t=0}^{t-1} 2\eta_t^\top R\hat{K}\hat{x}_{t|t,\hat{\Theta}}$

Let us define, $\hat{x}_{[t-1:0]} := \begin{bmatrix}\hat{x}_{t-1|t-1,\hat{\Theta}}^\top & \cdots & \hat{x}_{0|0,\hat{\Theta}}^\top\end{bmatrix}^\top$. Bounding this term is again similar to the previous cross-terms. Under the event $\mathcal{E}_x$, we have:

$$\begin{aligned}\left|\left|\text{diag}_t(R\hat{K})\hat{x}_{[t-1:0]}\right|\right| &\leq \sqrt{||R|| \cdot \hat{x}_{[t-1:0]}^\top \text{diag}_t(\hat{K}^\top R\hat{K})\hat{x}_{[t-1:0]}} \\ &\leq \sqrt{||R|| \cdot \left|\left|\text{diag}_t(\hat{K}^\top R\hat{K})\right|\right| \cdot ||\hat{x}_{[t-1:0]}||^2} \\ &\leq ||R||\,\Gamma\sqrt{t}\bar{\chi}.\end{aligned}$$

From Lemma B.5, we have

$$\begin{aligned}\sum_{t=0}^{t-1} -2\eta_t^\top R\hat{K}\hat{x}_{t|t,\hat{\Theta}} &= 2(-\eta_{[t-1:0]})^\top \text{diag}_t(R\hat{K})\hat{x}_{[t-1:0]} \\ &\lesssim \sigma_\eta\sqrt{t}\,||R||\,\Gamma\bar{\chi}\sqrt{\log\left(\frac{2}{\delta}\right)},\end{aligned} \tag{4-70}$$

which holds with a probability of at least $1-\delta$ under the event $\mathcal{E}_x$.

Upper bounding $\sum_{t=0}^{t-1} \eta_t^\top R\eta_t$

From Hanson-Wright inequality (refer to Theorem B.3), we have

$$\begin{aligned}&\sum_{t=0}^{t-1} \eta_t^\top R\eta_t = \eta_{[t-1:0]}^\top \text{diag}_t(R)\eta_{[t-1:0]} \\ &\leq tn_u\sigma_\eta^2 \text{Tr}\,(R) + \mathcal{O}\left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + ||R||\log\left(\frac{2}{\delta}\right)\right) \\ &\leq tn_u\sigma_\eta^2 \text{Tr}\,(R) + \mathcal{O}\left(\sqrt{tn_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)||R||,\end{aligned} \tag{4-71}$$

which holds with a probability of at least $1-\delta$.

Upper bounding $\sum_{t=0}^{t-1} z_t^\top Qz_t$

From Hanson-Wright inequality, we have

$$\begin{aligned}&\sum_{t=0}^{t-1} z_t^\top Qz_t = z_{[t-1:0]}^\top \text{diag}_t(Q)z_{[t-1:0]} \\ &\leq tn_y\sigma_z^2 \text{Tr}\,(Q) + \mathcal{O}\left(\sqrt{tn_x \log\left(\frac{2}{\delta}\right)} + ||Q||\log\left(\frac{2}{\delta}\right)\right) \\ &\leq tn_y\sigma_z^2 \text{Tr}\,(Q) + \mathcal{O}\left(\sqrt{tn_x \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)||Q||,\end{aligned} \tag{4-72}$$

which holds with a probability of at least $1-\delta$.

Putting things together

We have the following bound under the event $\mathcal{E}_x$:

$$\begin{aligned}\sum_{t=0}^{t-1} c_{t,2} \lesssim\ & \sqrt{t}\sigma_z ||Q|| \cdot ||C|| X_{\mathrm{ac}} \sqrt{\log\left(\frac{2}{\delta}\right)} + \sigma_\eta \sqrt{t}\, ||R||\, \Gamma \bar{\chi} \sqrt{\log\left(\frac{2}{\delta}\right)} \\ & + t n_u \sigma_\eta^2 \mathrm{Tr}\left(R\right) + \left(\sqrt{t n_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||R|| \\ & + t n_y \sigma_z^2 \mathrm{Tr}\left(Q\right) + \left(\sqrt{t n_x \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||Q||,\end{aligned} \tag{4-73}$$

which holds with a probability of at least $1 - 4\delta$.

Final upper-bound on the cumulative cost

Combining (4-68) and (4-73), we have

$$\begin{aligned}&\sum_{t=0}^{t-1} y_t^\top Q y_t + u_t^\top R u_t \\ &\lesssim t J_s(\hat{\Theta}) + t n_u c_{\hat{\mathbf{G}}_3}^2 \, ||S|| \, \sigma_\eta^2 \left(1 + \max\{\sigma_w, \sigma_z\} \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ &\quad + \max\{\sigma_w, \sigma_z, \sigma_\eta\} \left(X_{\mathrm{ac}}^2 + \bar{\chi}^2\right) ||S|| \left(1 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ &\quad + \left(\sqrt{\left(t(n_x + n_y) + \max\{\sigma_w^2, \sigma_z^2\}\right) \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) \left(c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ &\quad + \left(\sqrt{t n_u \log\left(\frac{2}{\delta}\right)} + \left(\sqrt{t n_u \max\{\sigma_w^2, \sigma_z^2\}} + 1\right) \log\left(\frac{2}{\delta}\right)\right) \left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ &\quad + \left(\sigma_\eta \sqrt{\log\left(\frac{2}{\delta}\right)} + \max\{\sigma_w, \sigma_z\} \log^2\left(\frac{2}{\delta}\right)\right) \left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ &\quad + \sqrt{t}\sigma_z ||Q|| \cdot ||C|| X_{\mathrm{ac}} \sqrt{\log\left(\frac{2}{\delta}\right)} + \sigma_\eta \sqrt{t}\, ||R||\, \Gamma \bar{\chi} \sqrt{\log\left(\frac{2}{\delta}\right)} \\ &\quad + t n_u \sigma_\eta^2 \mathrm{Tr}\left(R\right) + \left(\sqrt{t n_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||R|| \\ &\quad + t n_y \sigma_z^2 \mathrm{Tr}\left(Q\right) + \left(\sqrt{t n_x \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||Q|| \coloneqq c_{\mathrm{cost},k},\end{aligned} \tag{4-74}$$

which holds with a probability of at least $1 - 10\delta$ under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x$. Further, we can say that there exists an event $\mathcal{E}_{\mathrm{cost}}$, which holds with a probability of at least $1 - 10\delta$ such that, on $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x \cap \mathcal{E}_{\mathrm{cost}}$ the following bound holds:

$$\sum_{t=0}^{t-1} y_t^\top Q y_t + u_t^\top R u_t \lesssim c_{\text{cost},k}.$$

Now, $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x \cap \mathcal{E}_{\text{cost}}$ holds with a probability of at least $1 - 13\delta$. We can re-parametrise $\delta \to \frac{\delta}{13\log_2(T)}$. The reason for doing so will become apparent in the following section. The bound in (4-74) holds for one episode. This bound can be used to determine the upper bound on the cumulative cost for the entire horizon by summing the established bound in (4-74) over the number of episodes, i.e., by taking the union bound. This will be addressed in the following section.

Regret upper bound

To recall, the regret is defined as such:

$$R(T) = \sum_{t=0}^{T-1} (y_t^\top Q y_t + u_t^\top R u_t - J_*).$$

To recall, the system parameters are estimated at the start of each episode. Hence, the system parameter being used during the k^{th} episode is denoted as $\hat{\Theta}_k$. Further, as a reminder, the number of time steps in each episode is double the previous episode. Hence, we can approximate the number of episodes to be $\log_2(T)$. Also, to recall l_k is the number of time steps in the k^{th} episode. From (4-74), we have

$$\begin{aligned} R(T) \lesssim \sum_{k=0}^{\log_2(T)-1} & l_k \left(J_s(\hat{\Theta}_k) - J_*\right) + l_k n_u c_{\hat{\mathbf{G}}_3}^2 \, ||S|| \, \sigma_{\eta_k}^2 \left(1 + \max\{\sigma_w, \sigma_z\}\sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ & + \max\{\sigma_w, \sigma_z, \sigma_{\eta_k}\} \left(X_{\text{ac}}^2 + \bar{\chi}^2\right) ||S|| \left(1 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ & + \left(\sqrt{\left(l_k(n_x+n_y) + \max\{\sigma_w^2, \sigma_z^2\}\right)\log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)\left(c_{\hat{\mathbf{G}}_2}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ & + \left(\sqrt{l_k n_u \log\left(\frac{2}{\delta}\right)} + \left(\sqrt{l_k n_u \max\{\sigma_w^2, \sigma_z^2\}} + 1\right)\log\left(\frac{2}{\delta}\right)\right)\left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ & + \left(\sigma_{\eta_k}\sqrt{\log\left(\frac{2}{\delta}\right)} + \max\{\sigma_w, \sigma_z\}\log^2\left(\frac{2}{\delta}\right)\right)\left(c_{\hat{\mathbf{G}}_3}^2 c_{\hat{\mathbf{G}}_1}^2 c_{\bar{\mathbf{W}}}\right) \\ & + \sqrt{l_k}\sigma_z ||Q|| \cdot ||C|| X_{\text{ac}}\sqrt{\log\left(\frac{2}{\delta}\right)} + \sigma_{\eta_k}\sqrt{l_k}\, ||R||\, \Gamma\bar{\chi}\sqrt{\log\left(\frac{2}{\delta}\right)} \\ & + l_k n_u \sigma_{\eta_k}^2 \operatorname{Tr}(R) + \left(\sqrt{l_k n_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||R|| \\ & + l_k n_y \sigma_z^2 \operatorname{Tr}(Q) + \left(\sqrt{l_k n_x \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right) ||Q||, \end{aligned} \tag{4-75}$$

which holds under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x \cap \mathcal{E}_{\text{cost}}$. Now, we will refine the above bound. To recall, $\sigma_{\eta_k}^2 = \frac{\gamma}{\sqrt{l_k}}$. Combining with the result in Theorem 4.2, we have the following result from Theorem 4 in [43]:

$$\begin{aligned} J(\hat{\Theta}_k) - J_* &= J_s(\hat{\Theta}_k) - J_s(\Theta) \\ &\leq c_\Theta \underbrace{\left(\max\{||\hat{A}_t - \mathbf{T}^\top A\mathbf{T}||_{\mathrm{F}}, ||\hat{B}_t - \mathbf{T}^\top B||_{\mathrm{F}}, ||\hat{C}_t - C\mathbf{T}||_{\mathrm{F}}, ||\hat{L}_t - \mathbf{T}^\top L||_{\mathrm{F}}\}\right)^2}_{\epsilon^2}, \end{aligned}$$

where c_Θ is some constant dependent on the true system parameter. Now, we have

$$\begin{aligned} J_s(\hat{\Theta}_k) - J_* &= J_s(\hat{\Theta}_k) - J_s(\Theta) - \sigma_z^2 n_y \mathrm{Tr}(Q) \\ &\lesssim c_\Theta \left(\frac{1}{\sqrt{l_k}}\right)^2 - \sigma_z^2 n_y \mathrm{Tr}(Q). \end{aligned}$$

Combining the above result with the bound in (4-75), we obtain the final regret upper bound:

$$\begin{aligned} R(T) \lesssim\ & \log_2(T) c_\Theta + \sqrt{T}\gamma n_u c^2_{\hat{\mathbf{G}}_3} ||S|| \left(1 + \max\{\sigma_w, \sigma_z\}\sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ & + \log_2(T)\max\{\sigma_w, \sigma_z, T^{-1/4}\}\left(X^2_{\mathrm{ac}} + \bar{\chi}^2\right)||S|| \left(1 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right) \\ & + \left(\sqrt{\left(T(n_x + n_y) + \max\{\sigma_w^2, \sigma_z^2\}\right)\log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_2} c^2_{\hat{\mathbf{G}}_1} c_{\bar{\mathbf{W}}}\right) \\ & + \left(\sqrt{Tn_u \log\left(\frac{2}{\delta}\right)} + \left(\sqrt{Tn_u \max\{\sigma_w^2, \sigma_z^2\}} + 1\right)\log\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_3} c^2_{\hat{\mathbf{G}}_1} c_{\bar{\mathbf{W}}}\right) \\ & + \log_2(T)\left(\sqrt{n_u\gamma}T^{-1/4}\sqrt{\log\left(\frac{2}{\delta}\right)} + \max\{\sigma_w, \sigma_z\}\log^2\left(\frac{2}{\delta}\right)\right)\left(c^2_{\hat{\mathbf{G}}_3} c^2_{\hat{\mathbf{G}}_1} c_{\bar{\mathbf{W}}}\right) \\ & + \sqrt{T}\sigma_z ||Q|| \cdot ||C|| X_{\mathrm{ac}}\sqrt{\log\left(\frac{2}{\delta}\right)} + \sqrt{T\gamma}\,||R||\,\Gamma\bar{\chi}\sqrt{\log\left(\frac{2}{\delta}\right)} \\ & + \sqrt{T}\gamma n_u \mathrm{Tr}\left(R\right) + \left(\sqrt{Tn_u \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)||R|| \\ & + \left(\sqrt{Tn_x \log\left(\frac{2}{\delta}\right)} + \log\left(\frac{2}{\delta}\right)\right)||Q||, \end{aligned} \tag{4-76}$$

which holds under the event $\mathcal{E}_{\mathbf{M}} \cap \mathcal{E}_x \cap \mathcal{E}_{\mathrm{cost}}$. The regret bound in (4-76) suggests that $R(T) = \tilde{\mathcal{O}}(\sqrt{T})$ with a probability of at least $1 - \delta$. This concludes the proof.

Chapter 5

Conclusions and Future Work

5-1 Conclusions

In this thesis, we have analysed learning and control in one coherent framework. In particular, we have focused on the learning and control of unknown partially observable LTI systems in the LQG setting. With a focus on designing computationally efficient LBC algorithms, we proposed LQG-NAIVE and LQG-IF2E in Chapter 3. LQG-NAIVE, which is based on the naive exploration strategy, is argued to be computationally efficient with the ability to guarantee a regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$. On the other hand, LQG-IF2E extends the setting of 'open-loop' additive excitation signal in LQG-NAIVE to a setting of 'closed-loop' additive excitation by incorporating FIM in designing the covariance of the external signal. The FIM, which has a significant presence in the field of system identification, is argued to show potential in adapting the magnitude of external signal to the degree of informativity in the output signal. In Chapter 4, we proceeded to derive finite-time guarantees on the persistence of excitation of the input-output signal and regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$ for LQG-NAIVE, which matches the rate of regret growth in the LQR setting up to poly-logarithmic constants. We further validated the finite-time regret guarantee of LQG-NAIVE with numerical simulations. Providing finite-time guarantees for LQG-IF2E is however significantly more challenging: although the optimal rate of growth of the FIM has been shown to be $\mathcal{O}(\sqrt{T})$ in the 'open-loop' setting, proving a similar finite-time result in the 'closed-loop' is significantly more challenging due to the additive excitation signal not being i.i.d. Therefore, in this thesis, we presented sufficient numerical results for LQG-IF2E, showing its potential to perform competitively with LQG-NAIVE, with a hope to engender sufficient motivation to pursue deriving finite-time guarantees for FIM-based LBC strategies such as LQG-IF2E.

5-2 Possible future directions

Firstly, the regret guarantee of LQG-NAIVE as detailed in (4-76), can be refined further by representing the terms in the regret upper bound as a function of the solution to DARE or a function of the controllability/observability matrices. This makes it easier to understand the relative difficulty in learning to control a specific instance of the system parameter. In the LQR setting, the solution to the DARE in (2-9), is generally present in the regret upper bound (cf. [58]). Unlike the LQR setting, the LQG setting requires the solution to two DAREs, (2-5) and (2-9). It would be an interesting direction to pursue to further refine the stated regret upper bound in terms of such quantities.

Secondly, the LQG-IF2E algorithm lacks finite-time guarantees on the persistence of excitation and regret growth. These guarantees require analysing the correlations in the external signal through mathematical tools that cater to such settings. Given the significance of the FIM-based input signal in system identification and regret minimisation, bolstered by the empirical results of LQG-IF2E in this thesis, addressing this challenge of deriving finite-time guarantees is a promising direction to pursue in the future.

Finally, it must be noted that the LBC algorithms proposed in this thesis rely on the fact that a regret growth of $\tilde{\mathcal{O}}(\sqrt{T})$ can be guaranteed if the additive excitation signal to the CEC diminishes at a rate $\mathcal{O}(\frac{1}{\sqrt{t}})$. Although this rate of regret growth is optimal, it is more intuitive to design the exploration signal directly by minimising the regret. Recently, the work in [24] addresses this challenge in the LQR setting by first decomposing the regret as a sum of an 'exploitation cost' and an 'exploration cost' followed by, determining the optimal value of η_t by directly minimising over this alternative regret formulation. However, this optimisation problem is non-convex. Reformulating the regret in a similar fashion and convexifying the resultant optimisation problem for the LQG setting is a promising direction to pursue.

Appendix A

Technical Background

A-1 Finite-time guarantees during the warm-up period

Lemma A.1 [36] Let $\Phi(A)$ be as defined in Section 2-9. For any $\delta \in (0, 1/6)$, with a probability of at least $1 - \delta/6$, the following bounds hold when controlling the system as defined in (2-1) with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for all $t \in [0, T_\mathrm{w} - 1]$:

$$||x_t|| \leq X_\mathrm{w},\ ||u_t|| \leq U_\mathrm{w},\ ||z_t|| \leq Z, \tag{A-1}$$

where

$$\begin{aligned} X_\mathrm{w} &:= (\sigma_w + \sigma_u ||B||) \frac{\Phi(A)\rho(A)}{\sqrt{1-\rho(A)^2}} \sqrt{2n_x \, \log(12 n_x T_\mathrm{w}/\delta)}, \\ U_\mathrm{w} &:= \sigma_u \sqrt{2n_u \, \log(12 n_u T_\mathrm{w}/\delta)}, \\ Z &:= \sigma_z \sqrt{2n_y \, \log(12 n_y T_\mathrm{w}/\delta)}. \end{aligned} \tag{A-2}$$

As a consequence of the above bounds, we have

$$||\phi_t|| \leq \underbrace{(||C|| X_\mathrm{w} + Z + U_\mathrm{w})}_{\Upsilon_\mathrm{w}} \sqrt{H}, \tag{A-3}$$

which holds with a probability of at least $1 - \delta/2$ with $\delta \in (0, 1/2)$, for all $t \in [H, T_\mathrm{w} - 1]$.

Lemma A.2 [35] For some $\sigma_\mathrm{o} > 0$, if the warm-up duration $T_\mathrm{w} \geq T_\mathrm{o}$, then for all $t \in [T_\mathrm{o}, T_\mathrm{w} - 1]$, and for any $\delta \in (0, 1)$, with a probability of at least $1 - \delta$, we have

$$\sigma_{\min}\left(\sum_{i=H}^{T_\mathrm{w}-1} \phi_i \phi_i^\top\right) \geq (T_\mathrm{w} - H) \frac{\sigma_\mathrm{o}^2 \min\{\sigma_w^2, \sigma_z^2, \sigma_u^2\}}{2}, \tag{A-4}$$

where

$$T_\mathrm{o} := \frac{32 \Upsilon_w^4 H \log\left(\frac{2H(n_y+n_u)}{\delta}\right)}{\sigma_o^4 \min\{\sigma_w^4, \sigma_z^4, \sigma_u^4\}}. \tag{A-5}$$

A-2 Confidence set construction

Proof of Theorem 4.2 For brevity, we have the following notation $\mathbf{O} = \mathbf{O}(\bar{A}, C, d_1)$, $\hat{\mathbf{O}}_\mathbf{t} = \mathbf{O}(\hat{\bar{A}}_t, \hat{C}_t, d_1)$, $\mathbf{C_F} = \mathbf{C}(\bar{A}, F, d_2+1)$, $\hat{\mathbf{C}}_{\mathbf{F_t}} = \mathbf{C}(\hat{\bar{A}}_t, \hat{F}_t, d_2+1)$, $\mathbf{C_B} = \mathbf{C}(\bar{A}, B, d_2+1)$, $\hat{\mathbf{C}}_{\mathbf{B_t}} = \mathbf{C}(\hat{\bar{A}}_t, \hat{B}_t, d_2+1)$. Let $T_N = T_\mathbf{M}\frac{8H}{\sigma^2_{n_x}(\mathcal{H})}$, then for $T_\text{w} \geq T_N$, we have $\sigma_{\min}(\mathcal{N}) \geq 2\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|$. Here, $\mathcal{N}$ denotes the best rank - n_x approximation of $\mathcal{H}$. Applying Lemma B.9 with $t \geq T_\text{w} \geq T_N$, the following result can be obtained:

$$\begin{aligned}
\left|\left|\hat{\mathbf{O}}_\mathbf{t} - \mathbf{OT}\right|\right|_\text{F}^2 + \left|\left|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}] - \mathbf{T}^\top[\mathbf{C_F} \;\; \mathbf{C_B}]\right|\right|_\text{F}^2 &\leq \frac{5n_x\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|^2}{\sigma_{n_x}(\mathcal{N}) - \left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|} \\
&\leq \frac{5n_x\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|^2}{2\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right| - \left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|} \\
&= \frac{10n_x\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|^2}{\sigma_{n_x}(\mathcal{N})}.
\end{aligned}$$

Here, $\sigma_{n_x}(\mathcal{N}) = \sigma_{\min}(\mathcal{N})$. Since $\hat{C}_t - \tilde{C}\mathbf{T}$ is a submatrix of $\hat{\mathbf{O}}_\mathbf{t} - \mathbf{OT}$, $\hat{B}_t - \mathbf{T}^\top\tilde{B}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{B_t}} - \mathbf{T}^\top\mathbf{C_B}$, and $\hat{F}_t - \mathbf{T}^\top\tilde{F}$ is a submatrix of $\hat{\mathbf{C}}_{\mathbf{F_t}} - \mathbf{T}^\top\mathbf{C_F}$, we get

$$||\hat{C}_t - \tilde{C}\mathbf{T}||, ||\hat{B}_t - \mathbf{T}^\top\tilde{B}||, ||\hat{F}_t - \mathbf{T}^\top\tilde{F}|| \leq \sqrt{\frac{10n_x\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|^2}{\sigma_{n_x}(\mathcal{N})}}.$$

Now applying Lemma B.8 with $d_1, d_2 \geq H/2$, we get

$$\begin{aligned}
\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right| &\leq 2\sqrt{\min\{d_1, d_2\}}\left|\left|\hat{\mathbf{M}}_t - \mathbf{M}\right|\right| \\
&\leq 2\sqrt{H/2}\left|\left|\hat{\mathbf{M}}_t - \mathbf{M}\right|\right| \\
&\leq \sqrt{2H}\left|\left|\hat{\mathbf{M}}_t - \mathbf{M}\right|\right|.
\end{aligned}$$

This implies

$$\begin{aligned}
||\hat{C}_t - \tilde{C}\mathbf{T}||, ||\hat{B}_t - \mathbf{T}^\top\tilde{B}||, ||\hat{F}_t - \mathbf{T}^\top\tilde{F}|| &\leq \frac{\sqrt{20n_xH}\left|\left|\hat{\mathbf{M}}_t - \mathbf{M}\right|\right|}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \\
&= \frac{\sqrt{20n_xH}\left|\left|\hat{\mathbf{M}}_t - \mathbf{M}\right|\right|}{\sqrt{\sigma_{n_x}(\mathcal{H})}}.
\end{aligned} \tag{A-6}$$

Equation (A-6) provides the advertised bounds in the theorem. From Theorem 4.3, we observe that $||\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}||_\text{F} = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$. That is, the estimation error is monotonically decreasing. Therefore, if $T_B = T_\mathbf{M}\frac{20n_xH}{\sigma_{n_x}(\mathcal{H})}$ and $T_\text{w} \geq T_B$, we have

$$||\hat{C}_t - \tilde{C}\mathbf{T}||, ||\hat{B}_t - \mathbf{T}^\top\tilde{B}||, ||\hat{F}_t - \mathbf{T}^\top\tilde{F}|| \leq 1.$$

Before we can bound $||\hat{A}_t - \mathbf{T}^\top \tilde{A}\mathbf{T}||$, we will first bound $||\hat{\hat{A}}_t - \mathbf{T}^\top \bar{\tilde{A}}\mathbf{T}||$. To recall, $\hat{\hat{A}}_t = \hat{A}_t - \hat{F}_t\hat{C}_t$ and $\bar{\tilde{A}} = \tilde{A} - \tilde{F}\tilde{C}$. Let $X = \mathbf{OT}$ and $Y = \mathbf{T}^\top[\mathbf{C_F} \ \ \mathbf{C_B}]$. Therefore,

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

We will now provide a bound for each of the above terms. The first and the third terms can be bounded from the perturbation bound presented in [68] and [45], as follows:

$$\left|\left|\mathbf{\hat{O}_t^\dagger} - X^\dagger\right|\right|_\mathrm{F} \leq \left|\left|\mathbf{\hat{O}_t} - X\right|\right|_\mathrm{F} \max\left\{||X^\dagger||^2, ||\mathbf{\hat{O}_t^\dagger}||^2\right\} \leq ||\mathcal{N} - \hat{\mathcal{N}}_t||\sqrt{\frac{10n_x}{\sigma_{n_x}(\mathcal{N})}}\max\left\{||X^\dagger||^2, ||\mathbf{\hat{O}_t^\dagger}||^2\right\}.$$

Since $\sigma_{n_x}(\mathcal{N}) \geq 2||\mathcal{N} - \hat{\mathcal{N}}_t||$, we have $||\hat{\mathcal{N}}_t|| \leq 2||\mathcal{N}||$ and $2\sigma_{n_x}(\hat{\mathcal{N}}_t) \geq \sigma_{n_x}(\mathcal{N})$ from Lemma B.10. Using this result, let us now analyse the following term:

$$\begin{aligned} \max\left\{||X^\dagger||^2, ||\mathbf{\hat{O}_t^\dagger}||^2\right\} &= \max\left\{\left|\left|\left(\mathbf{U\Sigma}^{1/2}\mathbf{T}^\top\right)^\dagger\right|\right|^2, \left|\left|\left(\mathbf{U_t\Sigma_t}^{1/2}I^\top\right)^\dagger\right|\right|^2\right\} \\ &= \max\left\{\left|\left|\mathbf{U\Sigma}^{-1/2}\mathbf{T}^\top\right|\right|^2, \left|\left|\mathbf{U_t\Sigma_t}^{-1/2}I^\top\right|\right|^2\right\} \\ &= \max\left\{\frac{1}{\sigma_{n_x}(\mathcal{N})}, \frac{1}{\sigma_{n_x}(\hat{\mathcal{N}}_t)}\right\} \\ &\leq \frac{2}{\sigma_{n_x}(\mathcal{N})}. \end{aligned}$$

In a similar fashion, the third term can also be bounded. Therefore,

$$\left|\left|\mathbf{\hat{O}_t^\dagger} - X^\dagger\right|\right|_\mathrm{F}, \left|\left|[\mathbf{\hat{C}_{F_t}} \ \ \mathbf{\hat{C}_{B_t}}]^\dagger - Y^\dagger\right|\right|_\mathrm{F} \leq \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right|\sqrt{\frac{40n_x}{\sigma_{n_x}^3(\mathcal{N})}}.$$

Using the above individual bounds, we obtain the following:

$$\begin{aligned} \left|\left|X^\dagger\left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\right)[\mathbf{\hat{C}_{F_t}} \ \ \mathbf{\hat{C}_{B_t}}]^\dagger\right|\right|_\mathrm{F} &\leq \left|\left|X^\dagger\right|\right|\left|\left|\left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\right)\right|\right|_\mathrm{F}\left|\left|[\mathbf{\hat{C}_{F_t}} \ \ \mathbf{\hat{C}_{B_t}}]^\dagger\right|\right| \\ &\leq \frac{2}{\sigma_{n_x}(\mathcal{N})}\sqrt{n_x}\left|\left|\left(\hat{\mathcal{H}}_t^+ - \mathcal{H}^+\right)\right|\right|, \\ \left|\left|X^\dagger\mathcal{H}^+\left([\mathbf{\hat{C}_{F_t}} \ \ \mathbf{\hat{C}_{B_t}}]^\dagger - Y^\dagger\right)\right|\right|_\mathrm{F} &\leq \left|\left|X^\dagger\right|\right|\left|\left|\mathcal{H}^+\right|\right|\left|\left|\left([\mathbf{\hat{C}_{F_t}} \ \ \mathbf{\hat{C}_{B_t}}]^\dagger - Y^\dagger\right)\right|\right|_\mathrm{F} \\ &\leq \sqrt{\frac{1}{\sigma_{n_x}(\mathcal{N})}}\sqrt{\frac{40n_x}{\sigma_{n_x}^3(\mathcal{N})}}\left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right|\left|\left|\mathcal{H}^+\right|\right| \\ &= \frac{2\sqrt{10n_x}}{\sigma_{n_x}^2(\mathcal{N})}\left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right|\left|\left|\mathcal{H}^+\right|\right|, \end{aligned}$$

$$\begin{aligned}
\left|\left|\left(\hat{\mathbf{O}}_{\mathbf{t}}^{\dagger} - X^{\dagger}\right)\hat{\mathcal{H}}_t^{+}[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^{\dagger}\right|\right|_{\mathrm{F}} &\leq \left|\left|\hat{\mathbf{O}}_{\mathbf{t}}^{\dagger} - X^{\dagger}\right|\right|_{\mathrm{F}} \left|\left|\hat{\mathcal{H}}_t^{+}\right|\right| \left|\left|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^{\dagger}\right|\right| \\
&\leq \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \sqrt{\frac{40n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left|\left|\hat{\mathcal{H}}_t^{+}\right|\right| \left|\left|[\hat{\mathbf{C}}_{\mathbf{F_t}} \;\; \hat{\mathbf{C}}_{\mathbf{B_t}}]^{\dagger}\right|\right| \\
&= \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \sqrt{\frac{40n_x}{\sigma_{n_x}^3(\mathcal{N})}} \frac{1}{\sqrt{\sigma_{n_x}(\hat{\mathcal{N}}_t)}} \left|\left|\hat{\mathcal{H}}_t^{+}\right|\right| \\
&\leq \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \sqrt{\frac{40n_x}{\sigma_{n_x}^3(\mathcal{N})}} \sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} \left|\left|\hat{\mathcal{H}}_t^{+}\right|\right| \\
&= \frac{4\sqrt{5n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\hat{\mathcal{H}}_t^{+}\right|\right| \\
&\leq \frac{4\sqrt{5n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left(\left|\left|\mathcal{H}^{+}\right|\right| + \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right|\right) \rightarrow \text{Triangle inequality.}
\end{aligned}$$

Combining these, we obtain

$$\begin{aligned}
||\hat{\bar{A}}_t - \mathbf{T}^{\top}\bar{\tilde{A}}\mathbf{T}||_{\mathrm{F}} \leq\; & \frac{4\sqrt{5n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left(\left|\left|\mathcal{H}^{+}\right|\right| + \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right|\right) + \frac{2}{\sigma_{n_x}(\mathcal{N})}\sqrt{n_x} \left|\left|\left(\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right)\right|\right| \\
& + \frac{2\sqrt{10n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^{+}\right|\right| \\
=\; & \frac{4\sqrt{5n_x} + 2\sqrt{10n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^{+}\right|\right| + \frac{4\sqrt{5n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \\
+\; & \frac{2}{\sigma_{n_x}(\mathcal{N})}\sqrt{n_x} \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \\
\leq\; & \frac{4\sqrt{5n_x} + 2\sqrt{10n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^{+}\right|\right| + \frac{4\sqrt{5n_x}}{2\left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \sigma_{n_x}(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \\
+\; & \frac{2}{\sigma_{n_x}(\mathcal{N})}\sqrt{n_x} \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \rightarrow \text{Using the condition in Lemma B.9} \\
=\; & \frac{4\sqrt{5n_x} + 2\sqrt{10n_x}}{\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^{+}\right|\right| + \frac{2\sqrt{5n_x}}{\sigma_{n_x}(\mathcal{N})} \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \\
& + \frac{2}{\sigma_{n_x}(\mathcal{N})}\sqrt{n_x} \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right| \\
\leq\; & \frac{31\sqrt{n_x}}{2\sigma_{n_x}^2(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^{+}\right|\right| + \frac{13\sqrt{n_x}}{2\sigma_{n_x}(\mathcal{N})} \left|\left|\hat{\mathcal{H}}_t^{+} - \mathcal{H}^{+}\right|\right|.
\end{aligned}$$

Now consider $\hat{A}_t = \hat{\bar{A}}_t + \hat{F}_t\hat{C}_t$.

$$\begin{aligned}
\left|\left|\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T}\right|\right|_{\mathrm{F}} &= \left|\left|\hat{\bar{A}}_t + \hat{F}_t\hat{C}_t - \mathbf{T}^{\top}\bar{\tilde{A}}\mathbf{T} - \mathbf{T}^{\top}\tilde{F}\tilde{C}\mathbf{T}\right|\right|_{\mathrm{F}} \\
&\leq \left|\left|\hat{\bar{A}}_t - \mathbf{T}^{\top}\bar{\tilde{A}}\mathbf{T}\right|\right|_{\mathrm{F}} + \left|\left|\hat{F}_t\hat{C}_t - \mathbf{T}^{\top}\tilde{F}\tilde{C}\mathbf{T}\right|\right|_{\mathrm{F}} \\
&= \left|\left|\hat{\bar{A}}_t - \mathbf{T}^{\top}\bar{\tilde{A}}\mathbf{T}\right|\right|_{\mathrm{F}} + \left|\left|\hat{F}_t\hat{C}_t - \mathbf{T}^{\top}\tilde{F}\tilde{C}\mathbf{T} - \mathbf{T}^{\top}\tilde{F}\hat{C}_t + \mathbf{T}^{\top}\tilde{F}\hat{C}_t\right|\right|_{\mathrm{F}} \\
&\leq \left|\left|\hat{\bar{A}}_t - \mathbf{T}^{\top}\bar{\tilde{A}}\mathbf{T}\right|\right|_{\mathrm{F}} + \left|\left|\left(\hat{F}_t - \mathbf{T}^{\top}\tilde{F}\right)\hat{C}_t\right|\right|_{\mathrm{F}} + \left|\left|\mathbf{T}^{\top}\tilde{F}\left(\hat{C}_t - \tilde{C}\mathbf{T}\right)\right|\right|_{\mathrm{F}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \left|\left|\hat{\bar{A}}_t - \mathbf{T}^\top \bar{\tilde{A}}\mathbf{T}\right|\right|_{\mathrm{F}} + \left|\left|\left(\hat{F}_t - \mathbf{T}^\top \tilde{F}\right)\right|\right|_{\mathrm{F}} \left|\left|\hat{C}_t - \tilde{C}\mathbf{T}\right|\right|_{\mathrm{F}} + \left|\left|\left(\hat{F}_t - \mathbf{T}^\top \tilde{F}\right)\right|\right|_{\mathrm{F}} \left|\left|\tilde{C}\mathbf{T}\right|\right| \\
&\quad + \left|\left|\mathbf{T}^\top \tilde{F}\right|\right| \left|\left|\left(\hat{C}_t - \tilde{C}\mathbf{T}\right)\right|\right|_{\mathrm{F}} \rightarrow \text{Triangle inequality} \\
&\leq \frac{31\sqrt{n_x}}{2\sigma^2_{n_x}(\mathcal{N})} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right| \left|\left|\mathcal{H}^+\right|\right| + \frac{13\sqrt{n_x}}{2\sigma_{n_x}(\mathcal{N})} \left|\left|\hat{\mathcal{H}}^+_t - \mathcal{H}^+\right|\right| + \frac{10n_x \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right|^2}{\sigma_{n_x}(\mathcal{N})} \\
&\quad + \left(\left|\left|\tilde{C}\mathbf{T}\right|\right| + \left|\left|\mathbf{T}^\top \tilde{F}\right|\right|\right) \sqrt{\frac{10n_x}{\sigma_{n_x}(\mathcal{N})}} \left|\left|\mathcal{N} - \hat{\mathcal{N}}_t\right|\right|
\end{aligned}
$$

Finally, from Lemma B.8 we have

$$
\begin{aligned}
\left|\left|\hat{A}_t - \mathbf{T}^\top \tilde{A}\mathbf{T}\right|\right|_{\mathrm{F}} \leq \frac{31\sqrt{2n_xH}}{2\sigma^2_{n_x}(\mathcal{N})} \left|\left|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\right|\right| \left|\left|\mathcal{H}\right|\right| + \frac{13\sqrt{n_xH}}{2\sqrt{2}\sigma_{n_x}(\mathcal{N})} \left|\left|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\right|\right| \\
+ \frac{20n_xH \left|\left|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\right|\right|^2}{\sigma_{n_x}(\mathcal{N})} + \left(\left|\left|\tilde{C}\right|\right| + \left|\left|\tilde{F}\right|\right|\right) \sqrt{\frac{20n_xH}{\sigma_{n_x}(\mathcal{N})}} \left|\left|\hat{\mathbf{M}}_\mathbf{t} - \mathbf{M}\right|\right|.
\end{aligned} \tag{A-7}
$$

Similar to the bounds in (A-6), define T_A such that $\left|\left|\hat{A}_t - \mathbf{T}^\top \tilde{A}\mathbf{T}\right|\right| \leq \sigma_{n_x}(\tilde{A})/2$ when $T_{\mathrm{w}} \geq T_A$, where

$$
T_A = T_\mathbf{M} \left(\frac{\frac{62\sqrt{2n_xH}}{2\sigma^2_{n_x}(\mathcal{N})} \left|\left|\mathcal{H}\right|\right| + \frac{26\sqrt{n_xH}}{2\sqrt{2}\sigma_{n_x}(\mathcal{N})} + \sqrt{\frac{40n_xH\sigma_{n_x}(\tilde{A})}{\sigma_{n_x}(\mathcal{N})}} + \frac{\sqrt{80n_xH}}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \left(\left|\left|\tilde{F}\right|\right| + \left|\left|\tilde{C}\right|\right|\right)}{\sigma_{n_x}(\tilde{A})} \right)^2.
$$

Now we will focus on $\left|\left|\hat{L}_t - \mathbf{T}^\top \tilde{L}\right|\right|_{\mathrm{F}}$.

$$
\begin{aligned}
\left|\left|\hat{L}_t - \mathbf{T}^\top \tilde{L}\right|\right|_{\mathrm{F}} &= \left|\left|\hat{A}^\dagger_t \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{O}^\dagger \mathcal{H}^-\right|\right|_{\mathrm{F}} \\
&= \Big|\Big|\hat{A}^\dagger_t \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t + \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} X^\dagger \hat{\mathcal{H}}^-_t \\
&\qquad + \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} X^\dagger \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{O}^\dagger \mathcal{H}^-\Big|\Big|_{\mathrm{F}} \\
&= \Big|\Big|\hat{A}^\dagger_t \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t + \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} X^\dagger \hat{\mathcal{H}}^-_t \\
&\qquad + \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} X^\dagger \hat{\mathcal{H}}^-_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} \mathbf{X}^\dagger \mathcal{H}^-\Big|\Big|_{\mathrm{F}} \rightarrow \text{Since } X = \mathbf{OT} \\
&\leq \left|\left|(\hat{A}^\dagger_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T}) \hat{\mathbf{O}}^\dagger_t \hat{\mathcal{H}}^-_t\right|\right|_{\mathrm{F}} + \left|\left|\mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} (\hat{\mathbf{O}}^\dagger_t - X^\dagger) \hat{\mathcal{H}}^-_t\right|\right|_{\mathrm{F}} \\
&\quad + \left|\left|\mathbf{T}^\top \tilde{A}^\dagger \mathbf{T} X^\dagger (\hat{\mathcal{H}}^-_t - \mathcal{H}^-)\right|\right|_{\mathrm{F}} \\
&\leq \left|\left|(\hat{A}^\dagger_t - \mathbf{T}^\top \tilde{A}^\dagger \mathbf{T})\right|\right|_{\mathrm{F}} \left|\left|\hat{\mathbf{O}}^\dagger_t\right|\right| \left|\left|\hat{\mathcal{H}}^-_t\right|\right| + \left|\left|\mathbf{T}^\top \tilde{A}^\dagger \mathbf{T}\right|\right| \left|\left|(\hat{\mathbf{O}}^\dagger_t - X^\dagger)\right|\right|_{\mathrm{F}} \left|\left|\hat{\mathcal{H}}^-_t\right|\right| \\
&\quad + \left|\left|\mathbf{T}^\top \tilde{A}^\dagger \mathbf{T}\right|\right| \left|\left|X^\dagger\right|\right| \left|\left|(\hat{\mathcal{H}}^-_t - \mathcal{H}^-)\right|\right|_{\mathrm{F}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \left\|(\hat{A}_t^{\dagger} - \mathbf{T}^{\top}\tilde{A}^{\dagger}\mathbf{T})\right\|_{\mathrm{F}} \left\|\hat{\mathbf{O}}_t^{\dagger}\right\| \left\|\hat{\mathcal{H}}_t^{-}\right\| + \left\|\tilde{A}^{\dagger}\right\| \left\|(\hat{\mathbf{O}}_t^{\dagger} - X^{\dagger})\right\|_{\mathrm{F}} \left\|\hat{\mathcal{H}}_t^{-}\right\| \\
&\quad + \sqrt{n_x}\left\|\tilde{A}^{\dagger}\right\| \left\|X^{\dagger}\right\| \left\|(\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-})\right\| \\
&\leq \left(\left\|(\hat{A}_t^{\dagger} - \mathbf{T}^{\top}\tilde{A}^{\dagger}\mathbf{T})\right\|_{\mathrm{F}} \sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} + \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| \sqrt{\frac{40 n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|\right) \left\|\hat{\mathcal{H}}_t^{-}\right\| \\
&\quad + \sqrt{n_x}\left\|\tilde{A}^{\dagger}\right\| \frac{1}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \left\|(\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-})\right\| \\
&\leq \left(\left\|(\hat{A}_t^{\dagger} - \mathbf{T}^{\top}\tilde{A}^{\dagger}\mathbf{T})\right\|_{\mathrm{F}} \sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} + \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| \sqrt{\frac{40 n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|\right) \left(||\mathcal{H}^{-}|| + \left\|\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-}\right\|\right) \\
&\quad + \sqrt{n_x}\left\|\tilde{A}^{\dagger}\right\| \frac{1}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \left\|(\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-})\right\|.
\end{aligned}
$$

From Lemma B.10, we have $\sigma_{n_x}(\hat{A}_t) \geq \sigma_{n_x}(\tilde{A})/2$. Now using the perturbation bounds of the Moore-Penrose inverse under the Frobenius norm [45], we obtain the following:

$$
\begin{aligned}
\left\|(\hat{A}_t^{\dagger} - \mathbf{T}^{\top}\tilde{A}^{\dagger}\mathbf{T})\right\|_{\mathrm{F}} &\leq \left\|(\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T})\right\|_{\mathrm{F}} \max\left\{||\mathbf{T}^{\top}\tilde{A}^{\dagger}\mathbf{T}||^2, ||\hat{A}_t^{\dagger}||^2\right\} \\
&= \left\|(\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T})\right\|_{\mathrm{F}} \max\left\{\frac{1}{\sigma_{n_x}^2(\tilde{A})}, \frac{1}{\sigma_{n_x}^2(\hat{A}_t)}\right\} \\
&\leq \frac{2}{\sigma_{n_x}^2(\tilde{A})} \left\|(\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T})\right\|_{\mathrm{F}}.
\end{aligned}
$$

Now using Lemma B.8, we have

$$
\begin{aligned}
\left\|\hat{L}_t - \mathbf{T}^{\top}\tilde{L}\right\|_{\mathrm{F}} \leq &\left(\frac{2}{\sigma_{n_x}^2(\tilde{A})} \left\|(\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T})\right\|_{\mathrm{F}} \sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} + \left\|\mathcal{N} - \hat{\mathcal{N}}_t\right\| \sqrt{\frac{40 n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|\right) \\
&\left(||\mathcal{H}^{-}|| + \left\|\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-}\right\|\right) + \sqrt{n_x}\left\|\tilde{A}^{\dagger}\right\| \frac{1}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \left\|(\hat{\mathcal{H}}_t^{-} - \mathcal{H}^{-})\right\| \\
\leq &\left(\frac{2}{\sigma_{n_x}^2(\tilde{A})} \left\|(\hat{A}_t - \mathbf{T}^{\top}\tilde{A}\mathbf{T})\right\|_{\mathrm{F}} \sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} + \sqrt{2H}\left\|\hat{\mathbf{M}}_t - \mathbf{M}\right\| \sqrt{\frac{40 n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|\right) \\
&\left(||\mathcal{H}|| + \sqrt{\frac{H}{2}}\left\|\hat{\mathbf{M}}_t - \mathbf{M}\right\|\right) + \sqrt{n_x}\sqrt{\frac{H}{2}} \frac{1}{\sqrt{\sigma_{n_x}(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\| \left\|\hat{\mathbf{M}}_t - \mathbf{M}\right\|.
\end{aligned}
\tag{A-8}
$$

With $T_{\mathrm{w}} \geq T_A$, we have

$$
\left\|\hat{L}_t - \mathbf{T}^{\top}\tilde{L}\right\|_{\mathrm{F}} \leq \left(\sigma_{n_x}(\tilde{A})\sqrt{\frac{2}{\sigma_{n_x}(\mathcal{N})}} + \sqrt{\frac{80 H n_x}{\sigma_{n_x}^3(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|\right) \left(||\mathcal{H}|| + \sqrt{\frac{H}{2}}\right) + \sqrt{\frac{H n_x}{2\sigma_{n_x}(\mathcal{N})}} \left\|\tilde{A}^{\dagger}\right\|.
\tag{A-9}
$$

This concludes the proof.

A-3 Regret minimisation

Proof of Corollary 2.1

From (2-18),

$$\begin{aligned}
&\sum_{t=0}^{T-1} c_{t,*} - TJ_* = \mathcal{O}(T^{1/2}\text{log}T) \\
\implies &\sum_{t=0}^{T-1} c_{t,*} + \sum_{t=0}^{T-1} c_t - \sum_{t=0}^{T-1} c_t - TJ_* = \mathcal{O}(T^{1/2}\text{log}T) \\
\implies &R(T) - \bar{R}(T) = \mathcal{O}(T^{1/2}\text{log}T).
\end{aligned}$$

If $\bar{R}(T) = \tilde{\mathcal{O}}(T^{1/2})$ then,

$$\begin{aligned}
&R(T) - \bar{R}(T) = \mathcal{O}(T^{1/2}\text{log}T) \\
\implies &R(T) = \mathcal{O}(T^{1/2}\text{log}T) + \tilde{\mathcal{O}}(T^{1/2}) \\
\implies &R(T) = \tilde{\mathcal{O}}(T^{1/2}).
\end{aligned}$$

The proof showing $\bar{R}(T) = \tilde{\mathcal{O}}(T^{1/2})$ when $R(T) = \tilde{\mathcal{O}}(T^{1/2})$ can be derived in the same way as above.

Representing the (sub)optimal long-term average expected cost as the solution of a Lyapunov equation

The following analysis will prove to be a critical component in establishing the finite-time regret upper bound. Roughly speaking, the regret is analysed episode-wise i.e., the cumulative difference between the (sub)optimal cost incurred by the LBC policy and the optimal long-term average expected cost J_* incurred by the optimal control policy (assuming the full knowledge of Θ) during each episode, is upper bounded. This bound on the cumulative difference in the cost incurred in each episode is then summed over the number of episodes to obtain the final regret upper bound. The final piece in establishing the regret upper bound requires bounding the sub-optimality gap $\Delta_{\hat{\Theta}_k}$ as defined in (2-15), where k is the episode number. This inherently requires a way to represent the (sub)optimal long-term average expected cost incurred during the LBC phase denoted by $J(\hat{\Theta}_k)$. The following exposition addresses this problem through a Lyapunov equation.

Since the estimated system parameter is maintained during each episode, for the sake of brevity, we will consider $\hat{\Theta}_k = \hat{\Theta}$ and $\hat{K}_k = \hat{K}$. To recall, the LBC policy as described in (3-9), is of the form

$$u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}} + \eta_t.$$

From (2-6), we have

$$\begin{aligned}
\hat{x}_{t|t-1,\hat{\Theta}} &= \left(\hat{A} - \hat{B}\hat{K}\right)\hat{x}_{t-1|t-1,\hat{\Theta}} + \hat{B}\eta_{t-1}, \\
\hat{x}_{t|t,\hat{\Theta}} &= \hat{x}_{t|t-1,\hat{\Theta}} + \hat{L}\left(y_t - \hat{C}\hat{x}_{t|t-1,\hat{\Theta}}\right) \\
&= \left(\hat{A} - \hat{B}\hat{K}\right)\hat{x}_{t-1|t-1,\hat{\Theta}} + \hat{B}\eta_{t-1} \\
&\quad + \hat{L}\Bigg(Cx_t + z_t - \hat{C}\left(\left(\hat{A} - \hat{B}\hat{K}\right)\hat{x}_{t-1|t-1,\hat{\Theta}} + \hat{B}\eta_{t-1}\right)\Bigg) \\
&= \left(I - \hat{L}\hat{C}\right)\Bigg(\left(\hat{A} - \hat{B}\hat{K}\right)\hat{x}_{t-1|t-1,\hat{\Theta}} + \hat{B}\eta_{t-1}\Bigg) \\
&\quad + \hat{L}\Bigg(C\left(Ax_{t-1} - B\hat{K}\hat{x}_{t-1|t-1,\hat{\Theta}} + B\eta_{t-1} + w_{t-1}\right) + z_t\Bigg).
\end{aligned}$$

Now,

$$\begin{aligned}
\underbrace{\begin{bmatrix} x_t \\ \hat{x}_{t|t,\hat{\Theta}} \end{bmatrix}}_{\bar{x}_t} &= \underbrace{\begin{bmatrix} A & -B\hat{K} \\ \hat{L}CA & \left(I - \hat{L}\hat{C}\right)\left(\hat{A} - \hat{B}\hat{K}\right) - \hat{L}CB\hat{K} \end{bmatrix}}_{\hat{\mathbf{G}}_\mathbf{1}} \begin{bmatrix} x_{t-1} \\ \hat{x}_{t-1|t-1,\hat{\Theta}} \end{bmatrix} \\
&\quad + \underbrace{\begin{bmatrix} I & 0 \\ \hat{L}C & \hat{L} \end{bmatrix}}_{\hat{\mathbf{G}}_\mathbf{2}} \underbrace{\begin{bmatrix} w_{t-1} \\ z_t \end{bmatrix}}_{\bar{\epsilon}_{t-1}} + \underbrace{\begin{bmatrix} B \\ \left(I - \hat{L}\hat{C}\right)\hat{B} + \hat{L}CB \end{bmatrix}}_{\hat{\mathbf{G}}_\mathbf{3}} \eta_{t-1} \\
\implies \bar{x}_t &= \hat{\mathbf{G}}_\mathbf{1}\bar{x}_{t-1} + \hat{\mathbf{G}}_\mathbf{2}\bar{\epsilon}_{t-1} + \hat{\mathbf{G}}_\mathbf{3}\eta_{t-1}.
\end{aligned}$$

Let us consider a case where $u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}}$. Then, $\tilde{x}_t = \hat{\mathbf{G}}_\mathbf{1}\tilde{x}_{t-1} + \hat{\mathbf{G}}_\mathbf{2}\bar{\epsilon}_{t-1}$. Now consider an alternative formulation of the *finite-horizon* LQG control problem:

$$\begin{aligned}
\bar{J}_s(\hat{\Theta}) &= \mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_c x_t + u_t^\top R u_t + \tilde{x}_T^\top Q_f \tilde{x}_T\right] \\
&= \mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{x}_t^\top \underbrace{\begin{bmatrix} Q_c & 0 \\ 0 & \hat{K}^\top R\hat{K} \end{bmatrix}}_{\bar{\mathbf{W}}} \tilde{x}_t + \tilde{x}_T^\top Q_f \tilde{x}_T\right] \quad \text{s.t.} \\
& x_{t+1} = Ax_t + Bu_t + w_t, \;\; w_t \sim \mathcal{N}(0, \sigma_w^2 I), \\
& y_t = Cx_t + z_t, \;\; z_t \sim \mathcal{N}(0, \sigma_z^2 I), \\
& \hat{x}_{t|t,\hat{\Theta}} = (I - \hat{L}\hat{C})\hat{x}_{t|t-1,\hat{\Theta}} + \hat{L}y_t, \\
& \hat{x}_{t+1|t,\hat{\Theta}} = \hat{A}\hat{x}_{t|t,\hat{\Theta}} + \hat{B}u_t, \\
& u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}},
\end{aligned} \tag{A-10}$$

where Q_f is the terminal cost weighting matrix, $Q_c = C^\top QC$, $\hat{K}$ stabilises the true system, and $\hat{A} - \hat{F}\hat{C}$ is asymptotically stable. The reason for considering this alternate formulation (A-10) becomes clear in the regret analysis, as detailed in Section 4-3-4. Define the finite-horizon value function as

$$\hat{V}(\tilde{x},k) = \mathbb{E}_{\{w_k\},\{z_k\}}\left[\sum_{t=k}^{T-1} \tilde{x}_t^\top \bar{\mathbf{W}}\tilde{x}_t + \tilde{x}_T^\top Q_f \tilde{x}_T \mid \tilde{x}_k = \tilde{x}\right], \tag{A-11}$$

and given that it can be assumed to have a quadratic form as $\hat{V}(\tilde{x},k) = \tilde{x}^\top S_k \tilde{x} + q_k$ with $q_T = 0$ [12], we can deduce from Bellman's principle of optimality that

$$\begin{aligned}
\hat{V}(\tilde{x},k) &= \tilde{x}_k^\top \bar{\mathbf{W}}\tilde{x}_k + \mathbb{E}_{\tilde{x}_{k+1}}\left[V^{\hat{K}}(\tilde{x}_{k+1},k+1) \mid \tilde{x}_k = \tilde{x}\right]\\
\implies \tilde{x}_k^T S_k \tilde{x}_k + q_k &= \tilde{x}_k^\top \bar{\mathbf{W}}\tilde{x}_k + \mathbb{E}_{\tilde{x}_{k+1}}\left[\tilde{x}_{k+1}^\top S_{k+1}\tilde{x}_{k+1} + q_{k+1}\right]\\
\implies \tilde{x}_k^\top S_k \tilde{x}_k + q_k &= \tilde{x}_k^\top \bar{\mathbf{W}}\tilde{x}_k + \tilde{x}_k^\top \hat{\mathbf{G}}_{\mathbf{1}}^\top S_{k+1}\hat{\mathbf{G}}_{\mathbf{1}}\tilde{x}_k + \mathbb{E}\left[\bar{\epsilon}_k^\top \hat{\mathbf{G}}_{\mathbf{2}}^\top S_{k+1}\hat{\mathbf{G}}_{\mathbf{2}}\bar{\epsilon}_k\right] + q_{k+1}\\
\implies S_k &= \bar{\mathbf{W}} + \hat{\mathbf{G}}_{\mathbf{1}}^\top S_{k+1}\hat{\mathbf{G}}_{\mathbf{1}}\\
\implies q_k &= \operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{k+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right) + q_{k+1}.
\end{aligned} \tag{A-12}$$

We find that $\hat{\mathbf{G}}_{\mathbf{1}}$ is a stable matrix since it is assumed that the control law $u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}}$ stabilises the true system and $\hat{A} - \hat{F}\hat{C}$ is stable. Further, $\bar{\mathbf{W}}$ can be verified to be symmetric positive semi-definite. This implies that (A-12) converges to a unique symmetric positive semi-definite solution S such that [12]:

$$S = \bar{\mathbf{W}} + \hat{\mathbf{G}}_{\mathbf{1}}^\top S \hat{\mathbf{G}}_{\mathbf{1}} \tag{A-13}$$

From Definition B.1 we have, $\operatorname{dlyap}(\hat{\mathbf{G}}_{\mathbf{1}}, \bar{\mathbf{W}}) = S$. Now, the expected cumulative cost given by $\bar{J}_s(\hat{\Theta}) = \hat{V}(\tilde{x},0)$ can be expressed as:

$$\begin{aligned}
\bar{J}_s(\hat{\Theta}) &= \mathbb{E}\left[\tilde{x}_0^\top S_0 \tilde{x}_0\right] + q_0\\
&= \mathbb{E}\left[\operatorname{tr}\left(S_0 \tilde{x}_0 \tilde{x}_0^\top\right)\right] + \sum_{t=0}^{T-1}\operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{t+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\\
&= \mathbb{E}\left[\operatorname{tr}\left(S_0 \begin{bmatrix}x_0\\ \hat{x}_{0|0,\hat{\Theta}}\end{bmatrix}\begin{bmatrix}x_0\\ \hat{x}_{0|0,\hat{\Theta}}\end{bmatrix}^\top\right)\right] + \sum_{t=0}^{T-1}\operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{t+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\\
&= \mathbb{E}\left[\operatorname{tr}\left(S_0 \begin{bmatrix}x_0 x_0^\top & x_0 \hat{x}_{0|0,\hat{\Theta}}^\top\\ \hat{x}_{0|0,\hat{\Theta}} x_0^\top & \hat{x}_{0|0,\hat{\Theta}}\hat{x}_{0|0,\hat{\Theta}}^\top\end{bmatrix}\right)\right] + \sum_{t=0}^{T-1}\operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{t+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right)\\
&= \operatorname{tr}\left(S_0 \begin{bmatrix}\Sigma & \Sigma C^\top \hat{L}^\top\\ \hat{L}C\Sigma & \hat{L}(C\Sigma C^\top + \sigma_z^2 I)\hat{L}^\top\end{bmatrix}\right) + \sum_{t=0}^{T-1}\operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{t+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0\\ 0 & \sigma_z^2 I\end{bmatrix}\right),
\end{aligned}$$

where the last equality comes from the assumption that $x_0 \sim \mathcal{N}(0,\Sigma)$ and $\hat{x}_{0|-1,\hat{\Theta}} = 0$. Now consider the following *infinite-horizon* setting of $\bar{J}_s(\hat{\Theta})$:

$$\begin{aligned}
J_s(\hat{\Theta}) &= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} x_t^\top Q_c x_t + u_t^\top R u_t\right] \\
&= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{x}_t^\top \underbrace{\begin{bmatrix} Q_c & 0 \\ 0 & \hat{K}^\top R\hat{K}\end{bmatrix}}_{\bar{\mathbf{W}}} \tilde{x}_t\right] \text{ s.t.} \\
& x_{t+1} = Ax_t + Bu_t + w_t,\;\; w_t \sim \mathcal{N}(0, \sigma_w^2 I), \\
& y_t = Cx_t + z_t,\;\; z_t \sim \mathcal{N}(0, \sigma_z^2 I), \\
& \hat{x}_{t|t,\hat{\Theta}} = (I - \hat{L}\hat{C})\hat{x}_{t|t-1,\hat{\Theta}} + \hat{L}y_t, \\
& \hat{x}_{t+1|t,\hat{\Theta}} = \hat{A}\hat{x}_{t|t,\hat{\Theta}} + \hat{B}u_t, \\
& u_t = -\hat{K}\hat{x}_{t|t,\hat{\Theta}}.
\end{aligned} \tag{A-14}$$

Since $S_t \to S$ as $T \to \infty$, we have

$$\begin{aligned}
J_s(\hat{\Theta}) = \lim_{T\to\infty}\frac{1}{T}&\left[\operatorname{tr}\left(S_0\begin{bmatrix}\Sigma & \Sigma C^\top \hat{L}^\top \\ \hat{L}C\Sigma & \hat{L}(C\Sigma C^\top + \sigma_z^2 I)\hat{L}^\top\end{bmatrix}\right) + \sum_{t=0}^{T-1}\operatorname{tr}\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S_{t+1}\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0 \\ 0 & \sigma_z^2 I\end{bmatrix}\right)\right] \\
= \operatorname{tr}&\left(\hat{\mathbf{G}}_{\mathbf{2}}^\top S\hat{\mathbf{G}}_{\mathbf{2}}\begin{bmatrix}\sigma_w^2 I & 0 \\ 0 & \sigma_z^2 I\end{bmatrix}\right).
\end{aligned} \tag{A-15}$$

This concludes the proof.

Appendix B

Technical Tools

Definition B.1 [58] (Discrete Lyapunov equation) Let $X, Y \in \mathbb{R}^{m \times m}$ with $Y = Y^\top$ and $\rho(X) < 1$. We let $\mathcal{T}_X[P] \coloneqq X^\top P X + Y$, and let $\text{dlyap}(X, Y)$ denote the unique positive semi-definite solution $\mathcal{T}_X[P] = P$.

Lemma B.1 [3] Let $v \in \mathbb{R}^d$ be an entry-wise R-sub-Gaussian random variable. Then with probability of at least $1 - \delta$, $||v|| \leq R\sqrt{2d \, \log(2d/\delta)}$.

Lemma B.2 [60] Consider a self-adjoint matrix martingale $\{\mathbf{Y}_k : k = 1, .., n\}$ in dimension d, and let $\{\mathbf{X}_k\}$ be the associated difference equation. Consider also a fixed sequence $\{\mathbf{A}_k\}$ of self-adjoint matrices that satisfy

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

Compute the variance parameter

$$\sigma^2 :=|| \sum_k \mathbf{A}_k^2 || \,.$$

Then for all $t \geq 0$,

$$\mathbb{P}\left\{\lambda_{\max}(\mathbf{Y}_n - \mathbb{E}\mathbf{Y}_n) \geq t\right\} \leq d \, e^{\frac{-t^2}{8\sigma^2}}.$$

Lemma B.3 [3] Let $X_1, ..., X_t$ be random variables. Let $a \in \mathbb{R}$. Let $S_t = \sum_{s=1}^t X_s$ and $\tilde{S}_t = \sum_{s=1}^t \mathbb{I}_{X_s \leq a} X_s$ where $\mathbb{I}_{X_s \leq a} X_s$ denotes the truncated version of X_s. Then it holds that

$$\mathbb{P}\left\{S_t > x\right\} \leq \mathbb{P}\left\{\max_{1 \leq s \leq t} X_s \geq a\right\} + \mathbb{P}\left\{\tilde{S}_t > x\right\}.$$

Lemma B.4 (Regularised design matrix Lemma) [2] When the covariates satisfy $||z_t|| \leq D$, with some $D > 0$ with probability 1 then,

$$\log \frac{\det(V_t)}{\det(\lambda I)} \leq d \log \left(\frac{\lambda d + tD^2}{\lambda d} \right),$$

where $V_t = \lambda I + \sum_{i=1}^{t} z_i z_i^\top$ for $z_i \in \mathbb{R}^d$.

Lemma B.5 (Gaussian concentration inequality) Let $X = \Big[X_1, ..., X_n \Big]^\top$ be a vector with i.i.d. standard Gaussian entries and $F : \mathbb{R}^n \to \mathbb{R}$ a L-Lipschitz function ($|F(x) - F(y)| \leq L||x - y||$, for all $x, y \in \mathbb{R}^n$). Then, for every $t \geq 0$

$$\mathbb{P}\left\{|F(X) - \mathbb{E}[F(X)]| \geq t\right\} \leq 2 \exp\left(\frac{-t^2}{2L^2} \right).$$

Lemma B.6 (Chain rule for Fisher information) [69] For a density p, consider the following FIM for a bivariate density $p_\theta(x, y)$:

$$\bar{I}_{p(x,y)}(\theta) = \int \int \nabla_\theta \log p_\theta(x, y) \left(\nabla_\theta \log p_\theta(x, y)\right)^\top p_\theta(x, y) dx dy.$$

Define the conditional FIM as

$$\bar{I}_{p(x|y)}(\theta) = \int \int \nabla_\theta \left(\log p_\theta(x|y)\right) \left[\nabla_\theta \left(\log p_\theta(x|y)\right)\right]^\top p_\theta(x|y) dx \; p_\theta(y) dy.$$

Then

$$\bar{I}_{p(x,y)}(\theta) = \bar{I}_{p(x|y)}(\theta) + \bar{I}_{p(y)}(\theta),$$

assuming that $\nabla_\theta \log p_\theta(x|y)$ and $\nabla_\theta \log p_\theta(y)$ have mean zero.

Lemma B.7 [70] Let $\mu : \mathbb{R}^{d_\theta} \to \mathbb{R}^d$ and $V : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d \times d}$, with $V > 0$ for all $\theta \in \mathbb{R}^{d_\theta}$, and define

$$\gamma_\theta(x) = \frac{1}{\sqrt{(2\pi)^d \det(V(\theta))}} \exp\left(-\frac{1}{2}(x - \mu(\theta))^\top V(\theta)^{-1}(x - \mu(\theta)) \right).$$

Then

$$\bar{I}_\gamma(\theta) = \left(\mathtt{D}_\theta \mu(\theta)\right)^\top V(\theta)^{-1} \left(\mathtt{D}_\theta \mu(\theta)\right) + \frac{1}{2} \left(\mathtt{D}_\theta \mathtt{vec}\left(V(\theta)\right)\right)^\top \left(I \otimes V(\theta)^{-2}\right) \mathtt{D}_\theta \mathtt{vec}\left(V(\theta)\right).$$

Lemma B.8 [50] Let $\mathcal{H}, \hat{\mathcal{H}}_t$ and $\mathcal{N}, \hat{\mathcal{N}}_t$ be as defined in Algorithm 1. They satisfy the following perturbation bounds,

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

$$\left|\left| \mathcal{N} - \hat{\mathcal{N}}_t \right|\right| \leq 2 \left|\left| \mathcal{H}^- - \hat{\mathcal{H}}_t^- \right|\right| \leq 2\sqrt{\min\{d_1, d_2\}} \left|\left| \hat{\mathbf{M}}_t - \mathbf{M} \right|\right|.$$

Lemma B.9 [50] Let $\mathcal{N}$ and $\hat{\mathcal{N}_t}$ be as defined in Algorithm 1. Suppose $\sigma_{\min}(\mathcal{N}) \geq 2\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|$. Let rank n_x matrices $\mathcal{N}, \hat{\mathcal{N}_t}$ have SVDs $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ and $\hat{\mathbf{U}}_t\hat{\mathbf{\Sigma}}_t\hat{\mathbf{V}}_t^T$ respectively. There exists a unitary matrix $\mathbf{T} \in \mathbb{R}^{n_x \times n_x}$ such that

$$\left|\left|\mathbf{U}\mathbf{\Sigma}^{1/2} - \hat{\mathbf{U}}_t\hat{\mathbf{\Sigma}}_t^{1/2}\mathbf{T}\right|\right|_{\mathrm{F}}^2 + \left|\left|\mathbf{V}\mathbf{\Sigma}^{1/2} - \hat{\mathbf{V}}_t\hat{\mathbf{\Sigma}}_t^{1/2}\mathbf{T}\right|\right|_{\mathrm{F}}^2 \leq \frac{5n_x\left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|^2}{\sigma_{\min}(\mathcal{N}) - \left|\left|\mathcal{N} - \hat{\mathcal{N}_t}\right|\right|}.$$

Lemma B.10 [50] Let $\mathcal{N}$ and $\hat{\mathcal{N}_t}$ be as defined in Algorithm 1. Suppose $\sigma_{\min}(\mathcal{N}) \geq 2||\mathcal{N} - \hat{\mathcal{N}_t}||$. Then, $||\hat{\mathcal{N}_t}|| \leq 2||\mathcal{N}||$ and $2\sigma_{\min}(\hat{\mathcal{N}_t}) \geq \sigma_{\min}(\mathcal{N})$.

Lemma B.11 [58] Let $\mathrm{Toep}_{i,j,l}(X)$ and $\mathrm{Col}_{i,j}(X)$ be as defined in (4-57) for $X \in \mathbb{R}^{m \times m}$. For any $i \leq j, l$, and for $Y \in \mathbb{R}^{m \times m}$, and $\mathrm{diag}_{j-i}(Y)$ denoting a $j - i$ block matrix with blocks Y on the diagonal, we have the bound

$$\mathrm{Tr}\left(\mathrm{Col}_{i,j}(X)^\top \mathrm{diag}_{j-i}(Y)\mathrm{Col}_{i,j}(X)\right) \leq \mathrm{Tr}(\mathrm{dlyap}(X, Y)).$$

Corollary B.11 [58] Let $\mathrm{Toep}_{i,j,l}(X)$ and $\mathrm{Col}_{i,j}(X)$ be as defined in (4-57) for $X \in \mathbb{R}^{m \times m}$. For any $i \leq j, l$, and for $Y \in \mathbb{R}^{m \times m}$, and $\mathrm{diag}_{j-i}(Y)$ denoting a $j-i$ block matrix with blocks Y on the diagonal, we have the bound

$$\mathrm{Col}_{i,j}(X)^\top \mathrm{diag}_{j-i}(Y)\mathrm{Col}_{i,j}(X) \leq \mathrm{dlyap}(X, Y).$$

Lemma B.12 [58] Let $\mathrm{Toep}_{i,j,l}(X)$ and $\mathrm{Col}_{i,j}(X)$ be as defined in (4-57) for $X \in \mathbb{R}^{m \times m}$. For any $i \leq j, l$, we have $||\mathrm{Col}_{i,j}(X)|| \leq ||\mathrm{Toep}_{i,j,l}(X)|| \leq ||X||_{\mathcal{H}_\infty}$.

Theorem B.1 [3] 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

$$V_t = \sum_{k=1}^{t} m_{k-1}m_{k-1}^T, \quad \bar{V_t} = V + V_t, \ \ t \geq 0.$$

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

$$S_t^\top V_t^{-1} S_t \leq 2R^2 \log\left(\frac{\det(\bar{V_t})^{1/2}\det(V)^{-1/2}}{\delta}\right) \ \ \forall t \geq 0.$$

Theorem B.2 (Azuma's inequality) Assume that $(X_s; s \geq 0)$ is a supermartingale and $|X_s - X_{s-1}| \leq c_s$ almost surely. Then for all $t > 0$ and $\epsilon > 0$,

$$\mathbb{P}\left\{|X_t - X_0| \geq \epsilon\right\} \leq 2\text{exp}\left(\frac{-\epsilon^2}{2\sum_{s=1}^{t} c_s^2}\right).$$

Theorem B.3 (Hanson-Wright inequality) [55] Let $X = (X_1, ..., X_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfy $\mathbb{E}[x_i] = 0$ and $||X_i||_{\psi_2} \leq k$ for all $i = 1, .., n$, where $||.||_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}[.]^p)^{1/p}$ is the sub-Gaussian norm. Let A be an $n \times n$ matrix. Then, for every $t \geq 0$,

$$\mathbb{P}\left\{|X^\top AX - \mathbb{E}X^\top AX| > t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{k^4||A||_\text{F}^2}, \frac{t}{k^2||A||}\right)\right],$$

where c is a positive absolute constant.

Bibliography

[1] Yasin Abbasi-Yadkori. *Online learning for linearly parametrized control problems.* PhD thesis, University of Alberta, Edmonton, AB, Canada, 2013.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

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

[4] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11, 2015.

[5] Marc Abeille and Alessandro Lazaric. Thompson sampling for LQ control problems. In *Artificial Intelligence and Statistics*, pages 1246–1254, 2017.

[6] Marc Abeille and Alessandro Lazaric. Improved regret bounds for Thompson sampling in LQ control problems. In *International Conference on Machine Learning*, pages 1–9, 2018.

[7] Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in LQRs via Lagrangian relaxation. In *International Conference on Machine Learning*, pages 23–31, 2020.

[8] Karl J Åström and Björn Wittenmark. *Adaptive Control.* Courier Corporation, 2013.

[9] Karl Johan Åström and Richard M Murray. *Feedback systems: An Introduction for Scientists and Engineers.* Princeton University Press, 2021.

[10] Karl Johan Åström and Björn Wittenmark. On self tuning regulators. *Automatica*, 9(2):185–199, 1973.

[11] Archith Athrey. Exploration techniques in the LBC of unknown linear systems in the LQ paradigm. Technical report, TU Delft, 2023.

[12] Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*, volume 4. Athena Scientific, 2012.

[13] Matthieu Blanke and Marc Lelarge. Online greedy identification of linear dynamical systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5363–5368, 2022.

[14] Xavier Bombois, Michel Gevers, and Gérard Scorletti. Open-loop versus closed-loop identification of Box-Jenkins models: a new variance analysis. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 3117–3122, 2005.

[15] Xavier Bombois, Gérard Scorletti, Michel Gevers, Paul M.J. Van den Hof, and Roland Hildebrand. Least costly identification experiment for control. *Automatica*, 42(10):1651–1662, 2006.

[16] Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning LQRs efficiently. In *International Conference on Machine Learning*, pages 1328–1337, 2020.

[17] Alon Cohen, Tomer Koren, and Yishay Mansour. Learning LQRs efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309, 2019.

[18] Kévin Colin, Mina Ferizbegovic, and Håkan Hjalmarsson. Regret minimization for LQ adaptive controllers using Fisher feedback exploration. *IEEE Control Systems Letters*, 6:2870–2875, 2022.

[19] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the LQR. *Advances in Neural Information Processing Systems*, 31, 2018.

[20] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time analysis of optimal adaptive policies for LQ systems. *arXiv preprint arXiv:1711.07230*, 2017.

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

[22] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On optimality of adaptive LQRs. *arXiv preprint arXiv:1806.10749*, 2018.

[23] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive LQRs. *Automatica*, 117:108982, 2020.

[24] Mina Ferizbegovic. *Dual control concepts for linear dynamical systems.* PhD thesis, KTH Royal Institute of Technology, 2022.

[25] Mukul Gagrani, Sagar Sudhakara, Aditya Mahajan, Ashutosh Nayyar, and Yi Ouyang. A modified Thompson sampling-based learning algorithm for unknown linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6658–6665, 2022.

[26] László Gerencsér, Håkan Hjalmarsson, and Lirong Huang. Adaptive input design for LTI systems. *IEEE Transactions on Automatic Control*, 62(5):2390–2405, 2016.

[27] Håkan Hjalmarsson. System identification of complex and structured systems. *European Journal of Control*, 15(3-4):275–310, 2009.

[28] BL Ho and Rudolf E Kálmán. Effective construction of linear state-variable models from input/output functions: Die konstruktion von linearen modeilen in der darstellung durch zustandsvariable aus den beziehungen für ein-und ausgangsgrößen. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

[29] Yassir Jedra and Alexandre Proutiere. Minimal expected regret in LQ control. In *International Conference on Artificial Intelligence and Statistics*, pages 10234–10321, 2022.

[30] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Thompson sampling for partially observable LQ control. In *2023 American Control Conference (ACC)*, pages 4561–4568, 2023.

[31] Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\mathcal{O}(\sqrt{T})$ regret in LQ control. In *Conference on Learning Theory*, pages 3235–3284, 2022.

[32] Tze Leung Lai. Asymptotically efficient adaptive control in stochastic regression models. *Advances in Applied Mathematics*, 7(1):23–45, 1986.

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

[34] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Logarithmic regret bound in partially observable linear dynamical systems. *Advances in Neural Information Processing Systems*, 33:20876–20888, 2020.

[35] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret bound of adaptive control in LQG systems. *arXiv preprint 2003.05999*, 2020.

[36] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Regret minimization in partially observable LQ control. *arXiv preprint arXiv:2002.00082*, 2020.

[37] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Anima Anandkumar. Adaptive control and regret minimization in LQG setting. In *2021 American Control Conference (ACC)*, pages 2517–2522, 2021.

[38] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pages 5354–5390, 2022.

[39] Christian A Larsson, Afrooz Ebadat, Cristian R Rojas, Xavier Bombois, and Håkan Hjalmarsson. An application-oriented approach to dual control with excitation for closed-loop identification. *European Journal of Control*, 29:1–16, 2016.

[40] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[41] Lennart Ljung. *System Identification (2nd Ed.): Theory for the User*. Prentice Hall PTR, USA, 1999.

[42] Asif Iqbal Malik and Biswajit Sarkar. Optimizing a multi-product continuous-review inventory model with uncertain demand, quality improvement, setup cost reduction, and variation control in lead time. *IEEE Access*, 6:36176–36187, 2018.

[43] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for LQ control. *Advances in Neural Information Processing Systems*, 32, 2019.

[44] Nikolai Matni, Alexandre Proutiere, Anders Rantzer, and Stephen Tu. From self-tuning regulators to reinforcement learning and back again. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3724–3740, 2019.

[45] 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.

[46] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[47] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

[48] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

[49] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with Thompson sampling. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1198–1205, 2017.

[50] Samet Oymak and Necmiye Ozay. Revisiting Ho–Kalman-based system identification: Robustness and finite-sample analysis. *IEEE Transactions on Automatic Control*, 67(4):1914–1928, 2021.

[51] Javad Parsa and Håkan Hjalmarsson. Optimal input design through infinity norm minimization using proximal mapping. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 4442–4447, 2021.

[52] Jan Willem Polderman. On the necessity of identifying the true parameter in adaptive LQ control. *Systems & Control Letters*, 8(2):87–91, 1986.

[53] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

[54] Francisco Rubio, Francisco Valero, and Carlos Llopis-Albert. A review of mobile robots: Concepts, methods, theoretical framework, and applications. *International Journal of Advanced Robotic Systems*, 16(2):1729881419839596, 2019.

[55] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[56] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Nonparametric finite time LTI system identification. *arXiv preprint arXiv:1902.01848*, 2019.

[57] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[58] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In *International Conference on Machine Learning*, pages 8937–8948, 2020.

[59] Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436, 2020.

[60] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[61] Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654, 2019.

[62] Anastasios Tsiamis and George J. Pappas. Linear systems can be hard to learn. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2903–2910, 2021.

[63] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the LQR: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083, 2019.

[64] Sundarapandian Vaidyanathan. Anti-synchronization of the generalized Lotka-Volterra three-species biological systems via adaptive control. *International Journal of PharmTech Research*, 8(8):141–156, 2015.

[65] Michel Verhaegen and Vincent Verdult. *Filtering and System Identification: A Least Squares Approach.* Cambridge university press, 2007.

[66] Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582, 2020.

[67] Andrew J. Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652, 2021.

[68] Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232, 1973.

[69] Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive LQR with unknown B-matrix. In *Learning for Dynamics and Control*, pages 213–226, 2021.

[70] Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning LQG systems. *arXiv preprint arXiv:2201.01680*, 2022.

Glossary

List of Acronyms

ARX	Auto Regressive Model
CEC	Certainty Equivalence Controller
DARE	Discrete Algebraic Riccati Equation
FIM	Fisher Information Matrix
IF2E	Inverse Fisher Feedback Exploration
LBC	Learning-Based Control
LQ	Linear Quadratic
LQG	Linear Quadratic Gaussian
LQR	Linear Quadratic Regulator
LTI	Linear Time Invariant
OFU	Optimism in the Face of Uncertainty
RMSE	Root Mean Squared Error
SISO	Single-Input-Single-Output
TS	Thompson Sampling
VAF	Variance Accounted For

List of Symbols

δ	Probability of an event not occurring
$\Delta_{\hat{\Theta}}$	Sub-optimality gap in the long-term average expected cost
$\hat{\Theta}$	Estimated system parameter
$\Omega(.)$	Big - Omega notation
Σ	Solution to the DARE in (2-5)

Θ	True system parameter
$\tilde{\Omega}(.)$	Big - Omega notation ignoring constants and poly-logarithmic terms
$\bar{R}(T)$	Secondary formulation of cumulative regret in the LQ setting
$\hat{\mathbf{M}}_{\mathbf{t}}$	Markov parameters estimated at time step t
$\mathbf{M}$	Markov parameters of the true system with parameter Θ
$\tilde{\mathcal{O}}(.)$	Big - O notation ignoring constants and poly-logarithmic terms
$\mathcal{I}_t$	The observations available to the controller until time step t
$\mathcal{O}(.)$	Big - O notation
$\mathcal{S}$	A set of system parameters of interest (refer section 2-9)
$\mathbf{C}(A, B, n_x)$	Controllability matrix with n_x-block columns
$\mathbf{O}(A, C, n_x)$	Observability matrix with n_x-block rows
$A \in \mathbb{R}^{n_x \times n_x}$	State matrix of the true system
$B \in \mathbb{R}^{n_x \times n_u}$	Input matrix of the true system
$C \in \mathbb{R}^{n_y \times n_x}$	Output matrix of the true system
c_{tol}	Tolerance value to switch to the FIM-based LBC strategy
c_t	Cost incurred by the true system at time step t
F	Optimal Kalman gain in the innovations form
H	Length of the input-output data history to construct the ϕ vector
$I_T(\theta)$	Fisher Information Matrix (FIM) after T time steps evaluated on θ
$J(\hat{\Theta})$	Long-term average expected cost incurred when using the control law computed from $\hat{\Theta}$ on the true system with parameter Θ
J_*	Optimal long-term average expected cost of the system with parameter Θ
$J_*(\tilde{\Theta}_t)$	Optimal long-term average expected cost of the system with parameter $\tilde{\Theta}_t$
K	Optimal feedback gain for the true system parameter
L	Optimal Kalman gain for the measurement and the time update of the state estimate
l_k	Length of the k^{th} episode
P	Solution to the DARE in (2-9)
$Q \in \mathbb{R}^{n_y \times n_y}$	Output weighting matrix
$R \in \mathbb{R}^{n_u \times n_u}$	Input weighting matrix
$R(T)$	Cumulative regret in the LQ setting
$u_t \in \mathbb{R}^{n_u}$	Input at time step t
$w_t \in \mathbb{R}^{n_x}$	Process noise at time step t
$x_t \in \mathbb{R}^{n_x}$	State of the system at time step t
$y_t \in \mathbb{R}^{n_y}$	System output at time step t
$z_t \in \mathbb{R}^{n_y}$	Measurement noise at time step t