Can Direct Latent Model Learning
Solve Linear Quadratic Gaussian Control?

Yi Tian   yitian@mit.edu
Massachusetts Institute of Technology
Kaiqing Zhang   kaiqing@umd.edu
University of Maryland, College Park
Russ Tedrake   russt@mit.edu
Massachusetts Institute of Technology
Suvrit Sra   suvrit@mit.edu
Massachusetts Institute of Technology

Abstract

We study the task of learning state representations from potentially high-dimensional observations, with the goal of controlling an unknown partially observable system. We pursue a *direct latent model learning* approach, where a dynamic model in some latent state space is learned by predicting quantities directly related to planning (e.g., costs) without reconstructing the observations. In particular, we focus on an intuitive cost-driven state representation learning method in solving Linear Quadratic Gaussian (LQG) control, one of the most fundamental partially observable control problems. As our main results, we establish finite-sample guarantees of finding a near-optimal state representation function and a near-optimal controller using the directly learned latent model. To the best of our knowledge, despite various empirical successes, prior to this work it was unclear if such a cost-driven latent model learner enjoys finite-sample guarantees. Our work underscores the value of predicting multi-step costs, an idea that is key to our theory, and notably also an idea that is known to be empirically valuable for learning state representations.

1 Introduction

We consider state representation learning for control in partially observable systems, inspired by the recent successes of *control from pixels* (Hafner et al., 2019b,a). Control from pixels is an everyday task for human beings, but it remains challenging for learning agents. Methods to achieve it generally fall into two main categories: *model-free* and *model-based* ones. Model-free methods directly learn a visuomotor policy, also known as direct reinforcement learning (RL) (Sutton and Barto, 2018). On the other hand, model-based methods, also known as indirect RL (Sutton and Barto, 2018), attempt to learn a *latent model* that is a compact representation of the system, and to synthesize a policy in the latent model. Compared with model-free methods,

1

model-based ones facilitate generalization across tasks and enable efficient planning (Hafner et al., 2020), and are sometimes more sample efficient (Tu and Recht, 2019; Sun et al., 2019; Zhang et al., 2019).

In latent model-based control, the state of the latent model is also referred to as a *state representation* in the deep RL literature, and the mapping from an observed history to a latent state is referred to as the (state) representation function. *Reconstructing the observation* often serves as a supervision for representation learning for control in the empirical RL literature (Hafner et al., 2019b,a, 2020; Fu et al., 2021; Wang et al., 2022). This is in sharp contrast to model-free methods, where the policy improvement step is completely cost-driven. Reconstructing observations provides a powerful supervision signal for learning a task-agnostic world model, but they are high-dimensional and noisy, so the reconstruction requires an expressive reconstruction function; latent states learned by reconstruction contain irrelevant information for control, which can distract RL algorithms (Zhang et al., 2020; Fu et al., 2021; Wang et al., 2022). This is especially the case for practical visuomotor control tasks, e.g., robotic manipulation and self-driving cars, where the visual images contain predominately task-irrelevant objects and backgrounds.

Various empirical attempts (Schrittwieser et al., 2020; Zhang et al., 2020; Okada and Taniguchi, 2021; Deng et al., 2021; Yang et al., 2022) have been made to bypass observation reconstruction. Apart from observation, the interaction involves two other variables: actions (control inputs) and costs. Inverse model methods (Lamb et al., 2022) reconstruct actions; while other methods rely on costs. We argue that since neither the reconstruction function nor the inverse model is used for policy learning, cost-driven state representation learning is the most direct one. In this paper, we aim to examine the soundness of this methodology in linear quadratic Gaussian (LQG) control, one of the most fundamental partially observable control models.

Parallel to the empirical advances of learning for control from pixels, partially observable linear systems has been extensively studied in the context of learning for dynamic control (Oymak and Ozay, 2019; Simchowitz et al., 2020; Lale et al., 2020, 2021; Zheng et al., 2021; Minasyan et al., 2021; Umenberger et al., 2022). In this context, the representation function is more formally referred to as a *filter*, the optimal one being the Kalman filter. Most existing *model-based* learning approaches for LQG control focus on the linear time-invariant (LTI) case, and are based on the idea of *learning Markov parameters* (Ljung, 1998), the mapping from control inputs to observations. Hence, they need to reconstruct observations by definition. Motivated by the empirical successes in control from pixels, we take a different, cost-driven route, in hope of avoiding reconstructing observations or control inputs, which we refer to as *direct latent model learning*.

We focus on finite-horizon time-varying LQG control and address the following question:

Can direct latent model learning provably solve LQG control?

This work answers the question in the affirmative. Below is an overview of the main results. Additional discussion of related work is deferred to Appendix A.

1.1 Overview of main results

Motivated by empirical works on state representation learning for control (Schrittwieser et al., 2020; Zhang et al., 2020) and approximate information states (Subramanian et al., 2020; Yang et al., 2022), we propose a direct model learning method (Algorithm 1), without reconstructing observations or using an inverse model (Mhammedi et al., 2020; Frandsen et al., 2022; Lamb et al., 2022), that has the guarantee informally stated in Theorem 1. In Theorem 1 below, the dependence on dimensions and other system parameters are polynomial.

Theorem 1. (Informal) *Given an unknown time-varying LQG control problem with horizon T, under standard assumptions including stability, controllability (within in ℓ steps) and cost observability, there exists a direct latent model learning algorithm that returns, from n collected trajectories, a state representation function and a controller such that for the LQG control problem*

- *at the first ℓ steps, the state representation function is $\mathcal{O}(\ell^{1/2}n^{-1/4})$-optimal and the controller is $\mathcal{O}(\mathcal{O}(1)^\ell \ell n^{-1/4})$-optimal;*
- *at the next $T - \ell$ steps, the state representation function is $\mathcal{O}(T^{3/2}n^{-1/2})$-optimal and the controller is $\mathcal{O}(T^4 n^{-1})$-optimal.*

Our method parameterizes the state representation function and the latent system (transition and cost functions) separately. Usually, in empirical works, the state representation and transition functions are jointly learned, and they are, in fact, composed in transition prediction. An interesting finding is that in LQG, the scalar cost is sufficiently informative such that using cumulative cost supervision alone can recover the state representation function. Hence, the representation function and the latent system can be learned sequentially: our method first learns the representation function by predicting *cumulative scalar cost* (Algorithm 2), and then fits the transition and cost functions by minimizing the *transition and cost prediction* errors in the latent space (Algorithm 3). The learned latent model then enables planning that leads to a near-optimal controller.

Challenges & our techniques. Overall, to establish finite-sample guarantees, a major technical challenge is to deal with the *quadratic regression* problem in cost prediction, arising from the inherent quadratic form of the cost function in LQG. Directly solving the problem for the representation function involves *quartic* optimization; instead, we propose to solve a quadratic regression problem, followed by low-rank approximate factorization. The quadratic regression problem also appears in identifying the cost matrices, which involves concentration for random variables that are fourth powers of Gaussians. We believe these techniques might be of independent interest.

Moreover, the first ℓ-step *latent* states may not be adequately *excited* (having full-rank covariance), which invalidates the use of most system identification techniques. We instead identify only *relevant directions* of the system parameters, and prove that this is sufficient for learning a near-optimal controller by analyzing state covariance mismatch. This fact is reflected in the separation in the statement of Theorem 1; developing finite-sample analysis in this case is technically challenging.

Implications. For practitioners, one takeaway from our work is the benefit of predicting *multi-step cumulative* costs in direct latent model learning. Whereas cost at a single time step may not be revealing enough of the latent state, cumulative cost across multiple steps can be. This idea has been previously used by MuZero (Schrittwieser et al., 2020) in state representation learning for control, and our work can be viewed as a formal understanding of it in the LQG setting.

Notation. Random vectors are denoted by lowercase letters; sometimes they also denote their realized values. Uppercase letters denote matrices, some of which can be random. 0 can denote the scalar zero, zero vector or zero matrix; 1 denotes either the scalar one or a vector consisting of all ones; I denotes an identity matrix. The dimension, when emphasized, is specified in subscripts, e.g., $0_{d_x \times d_x}, 1_{d_x}, I_{d_x}$. Let $\mathbb{I}_S$ denote the indicator function for set S and $\mathbb{I}_S(A)$ apply to matrix A elementwise. Let $a \wedge b$ denote the minimum between scalars a and b. Given vector $v \in \mathbb{R}^d$, $\|v\|$ denotes its ℓ_2-norm. For some positive semidefinite P, we define $\|v\|_P := (v^\top P v)^{1/2}$. Given symmetric matrices P and Q, $P \succeq Q$ or $Q \preceq P$ means $P - Q$ is positive semidefinite. Semicolon ";" denotes stacking vectors or matrices vertically. For a collection of d-dimensional vectors $(v_t)_{t=i}^j$, let $v_{i:j} := [v_i; v_{i+1}; \ldots; v_j] \in \mathbb{R}^{d(j-i+1)}$ denote the concatenation along the column. For random variable η, let $\|\eta\|_{\psi_\beta}$ denote its β-subweibull norm, a special case of Orlicz norms (Zhang and Wei, 2022), with $\beta = 1, 2$ corresponding to subexponential and subgaussian norms. $\sigma_i(A), \sigma_{\min}(A), \sigma_{\min}^+(A), \sigma_{\max}(A)$ denote its ith largest, minimum, minimum positive, maximum singular values, respectively. $\|A\|_2, \|A\|_F, \|A\|_*$ denote the operator (induced by vector 2-norms), Frobenius, nuclear norms of matrix A, respectively. $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product between matrices. The Kronecker, symmetric Kronecker and Hadamard products between matrices are denoted by "$\otimes$", "$\otimes_s$" and "$\odot$", respectively. $\text{vec}(\cdot)$ and $\text{svec}(\cdot)$ denote flattening a matrix and a symmetric matrix by stacking their columns; $\text{svec}(\cdot)$ does not repeat the off-diagonal elements, but scales them by $\sqrt{2}$ (Schacke, 2004). We adopt the standard use of $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot)$, where the hidden constants are dimension-free but can depend on system parameters.

2 Problem setup

We study partially observable linear time-varying (LTV) dynamical system

$$x_{t+1} = A_t^* x_t + B_t^* u_t + w_t, \qquad y_t = C_t^* x_t + v_t, \qquad t = 0, 1, \ldots, T-1, \tag{2.1}$$

and $y_T = C_T^* x_T + v_T$. For all $t \geq 0$, we have the notation of state $x_t \in \mathbb{R}^{d_x}$, observation $y_t \in \mathbb{R}^{d_y}$, and control $u_t \in \mathbb{R}^{d_u}$. $(w_t)_{t=0}^{T-1}$ are i.i.d. process noises, sampled from $\mathcal{N}(0, \Sigma_{w_t})$, $(v_t)_{t=0}^T$ are i.i.d. observation noises, sampled from $\mathcal{N}(0, \Sigma_{v_t})$. Let initial state x_0 be sampled from $\mathcal{N}(0, \Sigma_0)$.

Let $\Phi_{t,t_0} = A_{t-1}^* A_{t-2}^* \cdots A_{t_0}^*$ for $t > t_0$ and $\Phi_{t,t} = I$. Then $x_t = \Phi_{t,t_0} x_{t_0} + \sum_{\tau=t_0}^{t-1} \Phi_{t,\tau+1} w_\tau$ under zero control input. To ensure the state and the cumulative noise do not grow with time, we make the following uniform exponential stability assumption.

Assumption 1 (Uniform exponential stability)**.** *The system is uniformly exponentially stable. That is, there exists $\alpha > 0, \rho \in (0, 1)$ such that for any $0 \leq t_0 < t \leq T$, $\|\Phi_{t,t_0}\|_2 \leq \alpha \rho^{t-t_0}$.*

Assumption 1 is standard in controlling LTV systems (Zhou and Zhao, 2017; Minasyan et al., 2021), satisfied by a stable LTI system. It essentially says that zero control is a stabilizing policy, and can be relaxed to a given stabilizing policy. Potentially, it can even be relaxed to uniform exponential stabilizability, by using our method for one more step at a time and finding a stabilizing policy incrementally.

Define the ℓ-step controllability matrix

$$\Phi_{t,\ell}^c := [B_t^*, A_t^* B_{t-1}^*, \ldots, A_t^* A_{t-1}^* \cdots A_{t-\ell+2}^* B_{t-\ell+1}^*] \in \mathbb{R}^{d_x \times \ell d_u}$$

for $\ell - 1 \leq t \leq T - 1$, which reduces to the standard controllability matrix $[B, \ldots, A^{\ell-1}B]$ in the LTI setting. We make the following controllability assumption.

Assumption 2 (Controllability). *For all $\ell - 1 \leq t \leq T - 1$, $\mathrm{rank}(\Phi_{t,\ell}^c) = d_x$, $\sigma_{\min}(\Phi_{t,\ell}^c) \geq \nu > 0$.*

Under zero noise, $x_{t+\ell} = \Phi_{t+\ell,t} x_t + \Phi_{t+\ell-1,\ell}^c [u_{t+\ell-1}; \ldots; u_t]$, so Assumption 2 ensures that from any state x, there exist control inputs that drive the state to 0 in ℓ steps, and ν ensures that the equation leading to them is well conditioned. We do not assume controllability for $0 \leq t < \ell - 1$, since we do not want to impose the constraint that $d_u > d_x$. This turns out to present a significant challenge for latent model learning, as seen from the separation of the results before and after the ℓ-steps in Theorem 1.

The quadratic cost functions are given by

$$c_t(x, u) = \|x\|_{Q_t^*}^2 + \|u\|_{R_t^*}^2, \quad 0 \leq t \leq T - 1, \quad c_T(x) = \|x\|_{Q_T^*}^2,$$

for positive semidefinite matrices $(Q_t^*)_{t=0}^T$ and positive definite matrices $(R_t^*)_{t=0}^{T-1}$. Sometimes the cost is defined as a function on observation y. Since the quadratic form $y^\top Q_t^* y = x^\top (C_t^*)^\top Q_t^* C_t^* x$, our analysis still applies if the assumptions on $(Q_t^*)_{t=0}^T$ hold for $((C_t^*)^\top Q_t^* C_t^*)_{t=0}^T$ instead.

(A, C) and $(A, Q^{1/2})$ observabilities are standard assumptions in controlling LTI systems. To differentiate from the former, we call the latter cost observability, since it implies the states are observable through costs. Whereas Markov parameter based approaches need to assume (A, C) observability to identify the system, our cost driven approach does not. Robust control sometimes assumes $(A, Q^{1/2})$ observability with vector cost $Q^{1/2}x$. Here we deal with the more difficult problem of having only the scalar cost. Nevertheless, the notion of cost observability is still important for our approach, formally defined as follows.

Assumption 3 (Cost observability). *For all $0 \leq t \leq \ell - 1$, $Q_t^* \succcurlyeq \mu^2 I$. For all $\ell \leq t \leq T$, there exists $m > 0$ such that the cost observability Gramian (Kailath, 1980)*

$$\sum_{\tau=t}^{t+k-1} \Phi_{\tau,t}^\top Q_\tau^* \Phi_{\tau,t} = Q_t^* + A_t^* Q_{t+1}^* A_t^* + \ldots + (A_{t+k-2}^* \cdots A_t^*)^\top Q_{t+k-1}^* A_{t+k-2}^* \cdots A_t^* \succcurlyeq \mu^2 I,$$

where $k = m \wedge (T - t + 1)$.

This assumption ensures that without noises, if we start with a nonzero state, the cumulative cost becomes positive in m steps. The special requirement for $0 \leq t \leq \ell - 1$ results from the difficulty in lacking controllability. The following is a regularity assumption.

Assumption 4. $(\sigma_{\min}(\Sigma_{v_t}))_{t=0}^T$ *are uniformly lower bounded by* $\sigma_v > 0$. *The operator norms of all matrices in the problem definition are uniformly upper bounded, including* $(A_t^*, B_t^*, R_t^*, \Sigma_{w_t})_{t=0}^{T-1}$, $(C_t^*, Q_t^*, \Sigma_{v_t})_{t=0}^T$. *In other words, they are all* $\mathcal{O}(1)$.

Let $h_t := [y_{0:t}; u_{0:(t-1)}] \in \mathbb{R}^{(t+1)d_y + td_u}$ denote the available history before deciding control u_t. A policy $\pi = (\pi_t : h_t \mapsto u_t)_{t=0}^{T-1}$ determines at time t a control input u_t based on history h_t. With a slight abuse of notation, let $c_t := c_t(x_t, u_t)$ for $0 \le t \le T - 1$ and $c_T := c_T(x_T)$ denote the cost at each time step. Then, $J^\pi := \mathbb{E}^\pi[\sum_{t=0}^T c_t]$ is the expected cumulative cost under policy π, where the expectation is taken over the randomness in the process noises, observation noises and controls. The objective of LQG control is to find a policy π such that J^π is minimized.

If the system parameters $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (C_t^*, Q_t^*)_{t=0}^T)$ are known, the optimal control is obtained by combining the Kalman filter

$$z_0^* = L_0^* y_0, \quad z_{t+1}^* = A_t^* z_t^* + B_t^* u_t + L_{t+1}^*(y_{t+1} - C_{t+1}^*(A_t^* z_t^* + B_t^* u_t)), \quad 0 \le t \le T - 1,$$

with the optimal feedback control gains of the linear quadratic regulator (LQR) $(K_t^*)_{t=0}^{T-1}$, where $(L_t^*)_{t=0}^T$ are the Kalman gains; this is known as the *separation principle*. The Kalman gains and optimal feedback control gains are given by

$$L_t^* = S_t^*(C_t^*)^\top (C_t^* S_t^*(C_t^*)^\top + \Sigma_v)^{-1}, \quad K_t^* = -((B_t^*)^\top P_{t+1}^* B_t^* + R_t)^{-1}(B_t^*)^\top P_{t+1}^* A_t^*,$$

where S_t^* and P_t^* are determined by their corresponding Riccati difference equations (RDEs):

$$S_{t+1}^* = A_t^*(S_t^* - S_t^*(C_t^*)^\top (C_t^* S_t^*(C_t^*)^\top + \Sigma_v)^{-1} C_t^* S_t^*)(A_t^*)^\top + \Sigma_w, \quad S_0^* = \Sigma_0, \tag{2.2}$$

$$P_t^* = (A_t^*)^\top (P_{t+1}^* - P_{t+1}^* B_t^*((B_t^*)^\top P_{t+1}^* B_t^* + R_t^*)^{-1}(B_t^*)^\top P_{t+1}^*)A_t^* + Q_t^*, \quad P_T^* = Q_T^*. \tag{2.3}$$

We consider data-driven control in an unknown LQG system (2.1) with unknown cost matrices $(Q_t^*)_{t=0}^T$. For simplicity, we assume $(R_t^*)_{t=0}^T$ is known, though our approaches can be readily generalized to the case without knowing them; it suffices to identify them in (3.3).

2.1 Latent model of LQG

Under the Kalman filter, the observation prediction error $i_{t+1} := y_{t+1} - C_{t+1}^*(A_t^* z_t^* + B_t^* u_t)$ is called an *innovation*. It is known that i_t is independent of history h_t and $(i_t)_{t=0}^T$ are independent (Bertsekas, 2012). Now we are ready to present the following proposition that represents the system in terms of the state estimates by the Kalman filter, which we shall refer to as the *latent model*.

Proposition 1. *Let* $(z_t^*)_{t=0}^T$ *be state estimates given by the Kalman filter. Then,*

$$z_{t+1}^* = A_t^* z_t^* + B_t^* u_t + L_{t+1}^* i_{t+1},$$

where $L_{t+1}^* i_{t+1}$ *is independent of* z_t^* *and* u_t, *i.e., the state estimates follow the same linear dynamics with noises* $L_{t+1}^* i_{t+1}$. *The cost at step t can be reformulated as functions of the state estimates by*

$$c_t = \|z_t^*\|_{Q_t^*}^2 + \|u_t\|_{R_t^*}^2 + b_t + \gamma_t + \eta_t,$$

6

where $b_t > 0$, and $\gamma_t = \|z_t^* - x_t\|_{Q_t^*}^2 - b_t$, $\eta_t = \langle z_t^*, x_t - z_t^* \rangle_{Q_t^*}$ are both zero-mean subexponential random variables. Under Assumptions 1 and 4, $b_t = \mathcal{O}(1)$ and $\|\gamma_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$; moreover, if control $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $0 \le t \le T$, then $\|\eta_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$.

Proposition 1 states that 1) the dynamics of the state estimates produced by the Kalman filter remains the same as the original system up to noises, determined by $(A_t^*, B_t^*)_{t=0}^{T-1}$; 2) the costs are still determined by $(Q_t^*)_{t=0}^T$ and $(R_t^*)_{t=0}^{T-1}$, up to constants and noises. Hence, a latent model can be parameterized by $((A_t, B_t)_{t=0}^{T-1}, (Q_t)_{t=0}^T)$ (recall that we assume $(R_t^*)_{t=0}^T$ is known for convenience). Note that observation matrices $(C_t^*)_{t=0}^T$ are *not* involved.

Now let us take a closer look at the state representation function. The Kalman filter can be written as $z_{t+1}^* = \overline{A}_t^* z_t^* + \overline{B}_t^* u_t + L_{t+1}^* y_{t+1}$, where $\overline{A}_t^* = (I - L_{t+1}^* C_{t+1}^*) A_t^*$ and $\overline{B}_t^* = (I - L_{t+1}^* C_{t+1}^*) B_t^*$. For $0 \le t \le T$, unrolling the recursion gives

$$
\begin{aligned}
z_t^* &= \overline{A}_{t-1}^* z_{t-1}^* + \overline{B}_{t-1}^* u_{t-1} + L_t^* y_t \\
&= [\overline{A}_{t-1}^* \overline{A}_{t-2}^* \cdots \overline{A}_0^* L_0^*, \dots, L_t^*][y_0; \dots; y_t] + [\overline{A}_{t-1}^* \overline{A}_{t-2}^* \cdots \overline{A}_1^* \overline{B}_0^*, \dots, \overline{B}_{t-1}^*][u_0; \dots; u_{t-1}] \\
&=: M_t^* [y_{0:t}; u_{0:(t-1)}],
\end{aligned}
$$

where $M_t^* \in \mathbb{R}^{d_x \times ((t+1)d_y + t d_u)}$. This means the optimal state representation function is linear in the history of observations and controls. A state representation function can then be parameterized by matrices $(M_t)_{t=0}^T$, and the latent state at step t is given by $z_t = M_t h_t$.

Overall, a policy π is a combination of state representation function $(M_t)_{t=0}^{T-1}$ (M_T is not needed) and feedback gain $(K_t)_{t=0}^{T-1}$ in the latent model; in this case, we write $\pi = (M_t, K_t)_{t=0}^{T-1}$. This contrasts with the disturbance-based parameterization (Youla et al., 1976; Wang et al., 2019; Sadraddini and Tedrake, 2020; Simchowitz et al., 2020; Lale et al., 2020).

3 Methodology: direct latent model learning

State representation learning involves history data that contains samples of three variables: observation, control input, and cost. Each of these can potentially be used as a *supervision* signal, and be used to define a type of state representation learning algorithms. We summarize our categorization as follows.

- Predicting observations defines the class of *observation-reconstruction based* methods, including methods based on Markov parameters (mapping from controls to observations) in linear systems (Lale et al., 2021; Zheng et al., 2021) and methods that learn a mapping from states to observations in more complex systems (Ha and Schmidhuber, 2018; Hafner et al., 2019b,a). This type of method tends to recover all state components.

- Predicting controls defines the class of *inverse model* methods, where the control is predicted from states across different time steps (Mhammedi et al., 2020; Frandsen et al., 2022; Lamb et al., 2022). This type of method tends to recover the controllable state components.

- Predicting (cumulative) costs defines the class of *cost-driven latent model learning* methods (Zhang et al., 2020; Schrittwieser et al., 2020; Yang et al., 2022). This type of method tends to recover the state components relevant to the cost.

Algorithm 1 Direct latent model learning for LQG systems

1: **Input:** sample size n, input noise magnitude σ_u, singular value threshold $\theta = \Theta(n^{-1/4})$
2: Collect n trajectories using $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, for $0 \le t \le T-1$, to obtain data in the form of

$$\mathcal{D}_{\text{raw}} = (y_0^{(i)}, u_0^{(i)}, c_0^{(i)}, \ldots, y_{T-1}^{(i)}, u_{T-1}^{(i)}, c_{T-1}^{(i)}, y_T^{(i)}, c_T^{(i)})_{i=1}^n$$

3: Run CoReL($\mathcal{D}_{\text{raw}}, \theta$) (Algorithm 2) to obtain state representation function estimate $(\widehat{M}_t)_{t=0}^T$ and latent state estimates $(z_t^{(i)})_{t=0, i=1}^{T, n}$, so that the data are converted to

$$\mathcal{D}_{\text{state}} = (z_0^{(i)}, u_0^{(i)}, c_0^{(i)}, \ldots, z_{T-1}^{(i)}, u_{T-1}^{(i)}, c_{T-1}^{(i)}, z_T^{(i)}, c_T^{(i)})_{i=1}^n$$

4: Run SysID($\mathcal{D}_{\text{state}}$) (Algorithm 3) to obtain system parameter estimates $((\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$
   Find feedback gains $(\widehat{K}_t)_{t=0}^{T-1}$ from $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$ by RDE (2.3)
5: **Return:** policy $\hat{\pi} = (\widehat{M}_t, \widehat{K}_t)_{t=0}^{T-1}$

Our method falls into the cost-driven category, which is more direct than the other two types, in the sense that the cost is directly relevant to planning with a dynamic model, whereas the observation reconstruction functions and inverse models are not. Another reason why we call our method *direct latent model learning* is that compared with Markov parameter-based approaches for linear systems, our approach directly parameterizes the state representation function, without exploiting the structure of the Kalman filter, making our approach closer to empirical practice that was designed for general RL settings.

(Subramanian et al., 2020) proposes to optimize a simple combination of cost and transition prediction errors to learn a latent model. That is, we parameterize a state representation function by matrices $(M_t)_{t=0}^T$ and a latent model by matrices $((A_t, B_t)_{t=0}^{T-1}, (Q_t)_{t=0}^T)$ and then solve

$$\min_{(M_t, Q_t, b_t)_{t=0}^T, (A_t, B_t)_{t=0}^{T-1}} \sum_{t=0}^T \sum_{i=1}^n l_t^{(i)}, \tag{3.1}$$

where $(b_t)_{t=0}^T$ are additional scalar parameters to account for noises, and the loss at step t for trajectory i is defined by

$$l_t^{(i)} = \left(\|M_t h_t^{(i)}\|_{Q_t}^2 + \|u_t^{(i)}\|_{R_t^*}^2 + b_t - c_t^{(i)} \right)^2 + \left(M_{t+1} h_{t+1}^{(i)} - A_t M_t h_t^{(i)} - B_t u_t^{(i)} \right)^2, \tag{3.2}$$

for $0 \le t \le T-1$ and $l_T^{(i)} = \left(\|M_T h_T^{(i)}\|_{Q_T}^2 + b_T - c_T^{(i)} \right)^2$. The optimization problem (3.1) is nonconvex; even if we find the global minimum solution, it is unclear how to establish finite-sample guarantees for it. A main finding of this work is that for LQG, we can solve the cost and transition loss optimization problems sequentially, with the caveat of using cumulative costs.

Our method is summarized in Algorithm 1. It has three steps: cost-driven state representation function learning (CoReL, Algorithm 2), latent system identification (SysID, Algorithm 3), and planning by RDE (2.3). This three-step approach is very similar to World Models (Ha and Schmidhuber, 2018) used in empirical RL, except that in the first step, instead of using an autoencoder to learn the state representation function, we use cost values to supervise the representation learning. Most empirical state representation learning methods (Hafner et al.,

Algorithm 2 CoReL: cost driven state representation learning

1: **Input:** raw data $\mathcal{D}_{\text{raw}}$, singular value threshold $\theta = \Theta((\ell(d_y + d_u))^{1/2}d_x^{3/4}n^{-1/4})$
2: Estimate the state representation function and cost constants by solving

$$(\widehat{N}_t, \widehat{b}_t)_{t=0}^T \in \underset{(N_t=N_t^\top, b_t)_{t=0}^T}{\text{argmin}} \sum_{t=0}^T \sum_{i=1}^n (\big\| [y_{0:t}^{(i)}; u_{0:(t-1)}^{(i)}] \big\|_{N_t}^2 + \sum_{\tau=t}^{t+k-1} \|u_\tau^{(i)}\|_{R_\tau^*}^2 + b_t - \bar{c}_t^{(i)})^2, \qquad (3.3)$$

   where $k = 1$ for $0 \le t \le l-1$ and $k = m \wedge (T-t+1)$ for $\ell \le t \le T$
3: Find $\widetilde{M}_t \in \mathbb{R}^{d_x \times ((t+1)d_y + td_u)}$ such that $\widetilde{M}_t^\top \widetilde{M}_t$ is an approximation of $\widehat{N}_t$
4: For all $0 \le t \le \ell - 1$, set $\widehat{M}_t = \text{TruncSV}(\widetilde{M}_t, \theta)$; for all $\ell \le t \le T$, set $\widehat{M}_t = \widetilde{M}_t$
5: Compute $\widehat{z}_t^{(i)} = \widehat{M}_t[y_{0:t}^{(i)}; u_{0:t}^{(i)}]$ for all $t = 0, \ldots, T$ and $i = 1, \ldots, n$
6: **Return:** state representation estimate $(\widehat{M}_t)_{t=0}^T$ and latent state estimates $(\widehat{z}_t^{(i)})_{t=0,i=1}^{T,n}$

2019b,a; Schrittwieser et al., 2020) use cost supervision as one loss term; the special structure of LQG allows us to use it alone and have theoretical guarantees.

CoReL (Algorithm 2) is the core of our algorithm. Once the state representation function $(\widehat{M}_t)_{t=0}^T$ is obtained, SysId (Algorithm 3) identifies the latent system using ordinary linear and quadratic regression, followed by planning using RDE (2.3) to obtain controller $(\widehat{K}_t)_{t=0}^{T-1}$ from $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$. SysId consists of the standard regression procedures; the full algorithmic detail is deferred to Appendix E. Below we explain the cost-driven state representation learning algorithm (CoReL, Algorithm 2) in detail.

3.1 Learning the state representation function

The state representation function is learned via CoReL (Algorithm 2). Given the raw data consisting of n trajectories, CoReL first solves the regression problem (3.3) to recover the symmetric matrix $\widehat{N}_t$. The target $\bar{c}_t$ of regression (3.3) is defined by

$$\bar{c}_t := c_t + c_{t+1} + \ldots + c_{t+k-1},$$

where $k = 1$ for $0 \le t \le \ell - 1$ and $k = m \wedge (T - t + 1)$ for $\ell \le t \le T$. The superscript in $\bar{c}_t^{(i)}$ denotes the observed $\bar{c}_t$ in the ith trajectory. The quadratic regression has a closed-form solution, by converting it to linear regression using $\|v\|_P^2 = \langle vv^\top, P \rangle_F = \langle \text{svec}(vv^\top), \text{svec}(P) \rangle$.

Why cumulative cost? The state representation function is parameterized by $(M_t)_{t=0}^T$ and the latent state at step t is given by $z_t = M_t h_t$. The single-step cost prediction (neglecting control cost $\|u_t\|_{R_t^*}^2$ and constant b_t) is given by $\|z_t\|_{Q_t}^2 = h_t^\top M_t^\top Q_t M_t h_t$. The regression recovers $(M_t^*)^\top Q_t^* M_t^*$ as a whole, from which we can recover $(Q_t^*)^{1/2} M_t^*$ up to an orthonormal transform. If Q_t^* is positive definite and known, then we can further recover M_t^* from it. However, if Q_t^* does not have full rank, information about M_t^* is partially lost, and there is no way to fully recover M_t^* even if Q_t^* is known. To see why multi-step cumulative cost helps, define $\overline{Q}_t^* := \sum_{\tau=t}^{t+k-1} \Phi_{\tau,t}^\top Q_\tau^* \Phi_{\tau,t}$ for the same k above. Under zero control and zero noise, starting

from x_t at step t, the k-step cumulative cost is precisely $\|x_t\|^2_{\overline{Q}^*_t}$. Under the cost observability assumption (Assumption 3), $(\overline{Q}^*_t)^T_{t=0}$ are positive definite.

The normalized parameterization. Still, since $\overline{Q}^*_t$ is unknown, even if we recover $(M^*_t)^\top \overline{Q}^*_t M^*_t$ as a whole, it is not viable to extract M^*_t and $\overline{Q}^*_t$. Such ambiguity is unavoidable; in fact, for every $\overline{Q}^*_t$ we choose, there is an equivalent parameterization of the system such that the system response is exactly the same. In partially observable LTI systems, it is well known that the system parameters can only be recovered up to a similarity transform (Oymak and Ozay, 2019). Since every parameterization is correct, we simply choose $\overline{Q}^*_t = I$, which we refer to as the *normalized parameterization*. Concretely, let us define $x'_t = (\overline{Q}^*_t)^{1/2} x_t$. Then, the new parameterization is given by

$$x'_{t+1} = A^{*\prime}_t x'_t + B^{*\prime}_t u_t + w'_t, \quad y_t = C^{*\prime}_t x'_t + v_t, \quad c'_t(x', u) = \|x'\|^2_{Q^{*\prime}_t} + \|u\|^2_{R^*_t},$$

and $c'_T(x') = \|x'\|^2_{(Q^*_T)'}$, where for all $t \geq 0$,

$$A^{*\prime}_t = (\overline{Q}^*_{t+1})^{1/2} A^*_t (\overline{Q}^*_t)^{-1/2}, \quad B^{*\prime}_t = (\overline{Q}^*_{t+1})^{1/2} B^*_t, \quad C^{*\prime}_t = C^*_t (\overline{Q}^*_t)^{-1/2},$$
$$w'_t = (\overline{Q}^*_{t+1})^{1/2} w_t, \quad (Q^*_t)' = (\overline{Q}^*_t)^{-1/2} Q^*_t (\overline{Q}^*_t)^{-1/2}.$$

It is easy to verify that under the normalized parameterization the system satisfies Assumptions 1, 2, 3, and 4, up to a change of some constants in the bounds. Without loss of generality, we assume system (2.1) is in the normalized parameterization; otherwise the recovered state representation function and latent system are with respect to the normalized parameterization.

Low-rank approximate factorization. Regression (3.3) has a closed-form solution; solving it gives $(\widehat{N}_t, \widehat{b}_t)^T_{t=0}$. Constants $(\widehat{b}_t)^T_{t=0}$ account for the state estimation error, and are not part of the state representation function; $d_h \times d_h$ symmetric matrices $(\widehat{N}_t)^T_{t=0}$ are estimates of $(M^*_t)^\top M^*_t$ under the normalized parameterization, where $d_h = (t+1)d_y + t d_u$. M^*_t can only be recovered up to an orthonormal transform, since for any orthogonal $S \in \mathbb{R}^{d_x \times d_x}$, $(SM^*_t)^\top SM^*_t = (M^*_t)^\top M^*_t$.

We want to recover $\widetilde{M}_t$ from $\widehat{N}_t$ such that $\widehat{N}_t = \widetilde{M}_t^\top \widetilde{M}_t$. Let $U \Lambda U^\top = \widehat{N}_t$ be its eigenvalue decomposition. Let $\Sigma := \max(\Lambda, 0)$ be the positive semidefinite diagonal matrix containing nonnegative eigenvalues, where "max" applies elementwise. If $d_h \leq d_x$, we can construct $\widetilde{M}_t = [\Sigma^{1/2} U^\top; 0_{(d_x - d_h) \times d_h}]$ by padding zeros. If $d_h > d_x$, however, $\mathrm{rank}(\widehat{N}_t)$ may exceed d_x. Assume that the diagonal elements of Σ are in descending order. Let Σ_{d_x} be the left-top $d_x \times d_x$ block of Σ and U_{d_x} be the left d_x columns of U. By the Eckart-Young-Mirsky theorem, $\widetilde{M}_t = \Sigma_{d_x}^{1/2} U_{d_x}^\top$ is the best approximation among $d_x \times d_h$ matrices in term of the Frobenius norm.

Why singular value truncation in the first ℓ steps? The latent states are used to identify the latent system dynamics, so whether they are sufficiently excited, namely having full-rank covariance, makes a big difference: if not, the system matrices can only be identified partially. Proposition 2 below confirms that the optimal latent state $z^*_t = M^*_t h_t$ indeed have full-rank covariance for $t \geq \ell$.

Proposition 2. *If system* (2.1) *satisfies Assumptions* 2 *(controllability) and* 4 *(regularity), then under control* $(u_t)_{t=0}^{T-1}$, *where* $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, $\sigma_{\min}(\text{Cov}(z_t^*)) = \Omega(\nu^2)$, M_t^* *has rank* d_x *and* $\sigma_{\min}(M_t^*) = \Omega(\nu t^{-1/2})$ *for all* $\ell \le t \le T$.

Proposition 2 implies that for all $\ell \le t \le T$, N_t^* has rank d_x, so if d_x is not provided, this gives a way to discover it. For $\ell \le t \le T$, Proposition 2 guarantees that as long as $\widetilde{M}_t$ is close enough to M_t^*, it also has full rank, and so does $\text{Cov}(\widetilde{M}_t h_t)$. Hence, we simply take the final estimate $\widehat{M}_t = \widetilde{M}_t$. Without further assumptions, however, there is no such guarantee for $(\text{Cov}(z_t^*))_{t=0}^{\ell-1}$ and $(M_t^*)_{t=0}^{\ell-1}$. We make the following minimal assumption to ensure that the minimum positive singular value $(\sigma_{\min}^+(\text{Cov}(z_t^*)))_{t=0}^{\ell-1}$ are uniformly lower bounded.

Assumption 5. *For* $0 \le t \le \ell - 1$, $\sigma_{\min}^+(M_t^*) \ge \beta > 0$.

Still, for $0 \le t \le \ell - 1$, Assumption 5 does not guarantee the rank of $\text{Cov}(\widetilde{M}_t h_t)$, not even its minimum positive singular value; that is why we introduce TRUNCSV that truncates the singular values of $\widetilde{M}_t$ by a threshold $\theta > 0$. Concretely, we take $\widehat{M}_t = (\mathbb{I}_{[\theta, +\infty)}(\Sigma_{d_x}^{1/2}) \odot \Sigma_{d_x}^{1/2}) U_{d_x}^\top$. Then, $\widehat{M}_t$ has the same singular values as $\widetilde{M}_t$ except that those below θ are zeroed. We take $\theta = \Theta((\ell(d_y + d_u))^{1/2} d_x^{3/4} n^{-1/4})$ to ensure a sufficient lower bound on the minimum positive singular value of $\widehat{M}_t$ while not increasing the statistical errors.

4 Theoretical guarantees

Theorem 2 below offers finite-sample guarantees for our approach. Overall, it confirms direct latent model learning (Algorithm 1) as a viable path to solving LQG control.

Theorem 2. *Given an unknown LQG system* (2.1), *under Assumptions* 1, 2, 3, 4 *and* 5, *if we run Algorithm* 1 *with* $n \ge \text{poly}(T, d_x, d_y, d_u, \log(1/p))$, *then with probability at least* $1 - p$, *state representation function* $(\widehat{M}_t)_{t=0}^T$ *is* $\text{poly}(\ell, d_x, d_y, d_u) n^{-1/4}$ *optimal in the first* ℓ *steps, and* $\text{poly}(\nu^{-1}, T, d_x, d_y, d_u) n^{-1/2}$ *optimal in the next* $(T - \ell)$ *steps. Also, the learned controller* $(\widehat{K}_t)_{t=0}^{T-1}$ *is* $\text{poly}(\ell, \beta^{-1}, m, d_x, d_y, d_u) c^\ell n^{-1/4}$ *optimal for some dimension-free constant* $c > 0$ *depending on system parameters in the first* ℓ *steps, and* $\text{poly}(T, \nu^{-1}, m, d_x, d_y, d_u, \log(1/p)) n^{-1}$ *optimal in the last* $(T - \ell)$ *steps.*

From Theorem 2, we observe a separation of the convergence rates before and after time step ℓ, resulting from the loss of the full-rankness of $(\text{Cov}(z_t^*))_{t=0}^{\ell-1}$ and $(M_t^*)_{t=0}^{\ell-1}$. In more detail, the proof sketch goes as follows. Quadratic regression guarantees that $\widehat{N}_t$ converges to N_t^* at a rate of $n^{-1/2}$ for all $0 \le t \le T$. Before step ℓ, $\widehat{M}_t$ suffers a square root decay of the rate to $n^{-1/4}$ because M_t^* may not have rank d_x. Since $(\widehat{z}_t)_{t=0}^{\ell-1}$ may not have full-rank covariances, $(A_t^*)_{t=0}^{\ell-1}$ are only recovered partially. As a result, $(\widehat{K}_t)_{t=0}^{\ell-1}$ may not stabilize $(A_t^*, B_t^*)_{t=0}^{\ell-1}$, causing the exponential dependence on ℓ. This means if n is not big enough, this controller may be inferior to zero control, since the system $(A_t^*, B_t^*)_{t=0}^{\ell-1}$ is uniformly exponential stable (Assumption 1) and zero control has suboptimality gap linear in ℓ. After step ℓ, $\widehat{M}_t$ retains the $n^{-1/2}$ convergence rate, and so do $(\widehat{A}_t, \widehat{B}_t)$; the certainty equivalent controller then has an order of n^{-1} suboptimality gap for LQ control (Mania et al., 2019). A full proof is deferred to Appendix E.

Theorem 2 states the guarantees for the state representation function $(\widehat{M}_t)_{t=0}^{T}$ and the controller $(\widehat{K}_t)_{t=0}^{T-1}$ separately. One may wonder the suboptimality gap of $\hat{\pi} = (\widehat{M}_t, \widehat{K}_t)_{t=0}^{T-1}$ in combination; after all, this is the output policy. The new challenge is that a suboptimal controller is applied to a suboptimal state estimation. An exact analysis requires more effort, but a reasonable conjecture is that $(\widehat{M}_t, \widehat{K}_t)_{t=0}^{T-1}$ has the same order of suboptimality gap as $(\widehat{K}_t)_{t=0}^{T-1}$: before step ℓ, the extra suboptimality gap resulted from $(\widehat{M}_t)_{t=0}^{\ell-1}$ can be analyzed by considering perturbation $\widehat{K}_t(\widehat{M}_t - M_t^*)h_t$ on controls; after step ℓ, similar to the analysis of the LQG suboptimality gap in (Mania et al., 2019), the overall suboptimality gap can be analyzed by a Taylor expansion of the value function at $(M_t^*, K_t^*)_{t=\ell}^{T-1}$, with $(\widehat{K}_t\widehat{M}_t - K_t^*M_t^*)_{t=\ell}^{T-1}$ being perturbations.

5 Concluding remarks

We examined the direct latent model learning methods in time-varying LQG control. With a finite-sample analysis, we showed that a direct, cost-driven state representation learning algorithm effectively solves LQG. In the analysis, we revealed the importance of using multi-step cumulative costs as the supervision signal, and a separation of the convergence rates before and after step ℓ, the controllability index, due to early-stage insufficient excitement of the system. A major limitation of our method is the use of history-based state representation functions; recovering the recursive Kalman filter would be ideal, whose sample complexity, we find, has an exponential dependence on the horizon with our current techniques, which we discuss in detail in Appendix B.

This work has opened up many opportunities for future research. An immediate question is how our direct latent model learning approach performs in the infinite-horizon LTI setting. Moreover, one may wonder about the extent to which direct latent model learning generalizes to nonlinear observations or systems. Investigating the connection with reduced-order control is also an interesting question, which may reveal the unique advantage of direct latent model learning. Finally, one argument for favoring latent-model-based over model-free methods is their ability to generalize across different tasks; direct latent model learning may offer a perspective to formalize this intuition.

Acknowledgement

References

Dimitri Bertsekas. *Dynamic Programming and Optimal Control: Volume I*, volume 1. Athena Scientific, 2012.

Rajendra Bhatia and Fuad Kittaneh. On the singular values of a product of operators. *SIAM Journal on Matrix Analysis and Applications*, 11(2):272–277, 1990.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Fei Deng, Ingook Jang, and Sungjin Ahn. Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations. *arXiv preprint arXiv:2110.14565*, 2021.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 512–519. IEEE, 2016.

Abraham Frandsen, Rong Ge, and Holden Lee. Extracting latent state representations with linear dynamics from rich observations. In *International Conference on Machine Learning*, pages 6705–6725. PMLR, 2022.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *International Conference on Machine Learning*, pages 3480–3491. PMLR, 2021.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019b.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Botao Hao, Yasin Abbasi Yadkori, Zheng Wen, and Guang Cheng. Bootstrapping upper confidence bound. *Advances in Neural Information Processing Systems*, 32, 2019.

Ali Jadbabaie, Horia Mania, Devavrat Shah, and Suvrit Sra. Time varying regression with hidden linear dynamics. *arXiv preprint arXiv:2112.14862*, 2021.

Thomas Kailath. *Linear Systems*, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

Nicholas Komaroff. On bounds for the solution of the Riccati equation for discrete-time control systems. In *Control and Dynamic Systems*, volume 78, pages 275–311. Elsevier, 1996.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

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.

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

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Didolkar, Dipendra Misra, Dylan Foster, Lekan Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of controllable latent states with multi-step inverse models. *arXiv preprint arXiv:2207.08229*, 2022.

Lennart Ljung. System identification. In *Signal Analysis and Prediction*, pages 163–173. Springer, 1998.

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

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. *Advances in Neural Information Processing Systems*, 33:14532–14543, 2020.

Edgar Minasyan, Paula Gradu, Max Simchowitz, and Elad Hazan. Online control of unknown time-varying dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Masashi Okada and Tadahiro Taniguchi. Dreaming: Model-based reinforcement learning by latent imagination without reconstruction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4209–4215. IEEE, 2021.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Sadra Sadraddini and Russ Tedrake. Robust output feedback control with guaranteed constraint satisfaction. In *Proceedings of the 23rd International Conference on Hybrid Systems: Computation and Control*, pages 1–10, 2020.

Kathrin Schacke. On the Kronecker product. *Master's Thesis, University of Waterloo*, 2004.

Peter Hans Schoenemann. *A solution of the orthogonal Procrustes problem with applications to orthogonal and oblique rotation*. University of Illinois at Urbana-Champaign, 1964.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering Atari, Go, Chess and Shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

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

Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *arXiv preprint arXiv:2010.08843*, 2020.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. In *Advances in Neural Information Processing Systems*, 2022.

Jack Umenberger, Max Simchowitz, Juan Carlos Perdomo, Kaiqing Zhang, and Russ Tedrake. Globally convergent policy search for output estimation. In *Advances in Neural Information Processing Systems*, 2022.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University press, 2018.

Martin J Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Tongzhou Wang, Simon S Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised MDPs: Learning world models better than the world itself. *arXiv preprint arXiv:2206.15477*, 2022.

Yuh-Shyang Wang, Nikolai Matni, and John C Doyle. A system-level approach to controller synthesis. *IEEE Transactions on Automatic Control*, 64(10):4079–4093, 2019.

Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. *Advances in Neural Information Processing Systems*, 28, 2015.

Lujie Yang, Kaiqing Zhang, Alexandre Amice, Yunzhu Li, and Russ Tedrake. Discrete approximate information states in partially observable environments. In *2022 American Control Conference (ACC)*, pages 1406–1413. IEEE, 2022.

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

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Huiming Zhang and Haoyu Wei. Sharper sub-weibull concentrations. *Mathematics*, 10(13):2252, 2022.

Kaiqing Zhang, Xiangyuan Zhang, Bin Hu, and Tamer Basar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949–2964, 2021.

Marvin Zhang, Sharad Vikram, Laura Smith, Pieter Abbeel, Matthew Johnson, and Sergey Levine. Solar: Deep structured representations for model-based reinforcement learning. In *International Conference on Machine Learning*, pages 7444–7453. PMLR, 2019.

Qinghua Zhang and Liangquan Zhang. Boundedness of the Kalman filter revisited. *IFAC-PapersOnLine*, 54(7):334–338, 2021.

Yang Zheng, Luca Furieri, Maryam Kamgarpour, and Na Li. Sample complexity of linear quadratic Gaussian (LQG) control for output feedback systems. In *Learning for Dynamics and Control*, pages 559–570. PMLR, 2021.

Bin Zhou and Tianrui Zhao. On asymptotic stability of discrete-time linear time-varying systems. *IEEE Transactions on Automatic Control*, 62(8):4274–4281, 2017.

A   Related work

Empirical methods for control from observations.   To control from high-dimensional observations, a natural idea is to compress the observation to a latent space using autoencoders, and then to learn a controller on this latent-space model (Watter et al., 2015; Finn et al., 2016). This methodology has two potential problems: 1) such representations are not geared towards control, containing irrelevant information; 2) controlling from observations usually involves partial observations, in which case the latent states are not necessarily Markovian.

Recurrent neural networks model long-term dependencies, and can be used to overcome the second issue (Ha and Schmidhuber, 2018; Hafner et al., 2019b,a). In partially observable MDPs, the belief states are known to be Markovian, whose update function precisely takes a recurrent form. Hence, the hidden states in RNNs can be seen as an approximation of the belief states. World Models (Ha and Schmidhuber, 2018) still use autoencoders to learn embeddings, which are then used to train an RNN in the next stage. The embedding of the autoencoder and the hidden state of the RNN are used together as the latent state; the RNN is both a component of the state representation function and the latent dynamics function. PlaNet (Hafner et al., 2019b) replaces the autoencoder in World Models by another RNN, and learns the state representation function and the latent dynamics jointly using a combined loss function. Dreamer (Hafner et al., 2019a) simplifies PlaNet by using only one RNN (the representation function), and replaces the latent online planning component in PlaNet by actor-critic. However, since these three methods reconstruct the observation from the latent state, they still suffer the first problem of not learning control-purposed state representations.

(Hafner et al., 2019a) reports failure results about using *reward* as the only supervision for the representation function, suggesting the hardness of learning a latent model with only the scalar reward information. Nevertheless, this has been shown to be possible. MuZero (Schrittwieser et al., 2020) does this by adding value function obtained from Monte Carlo Tree Search as supervision, and Deep Bisimulation for Control (DBC) (Zhang et al., 2020) does this by using a loss function based on a bisimulation metric.

On the other hand, MuZero and DBC parameterize the state representation function as a feedforward neural network, not accounting for the potential issue of partial observability. MuZero uses frame stacking for Atari games indeed, which is directly applicable to DBC as well; it is effective if stacking the frames makes the system fully observable. That is, the underlying Markovian states can be recovered from the stacked frames (observations). For LQG, even if the system is output observable and has no noise, stacking the observations does not suffice for recovering the current state; the control inputs also need to be stacked. Without output observability, the finite history is not necessarily Markovian; in the presence of noise, the finite-memory state estimator is not optimal, given that the optimal state estimator (Kalman filter) needs the full history by its recurrent form. Our algorithm aims to recover the optimal state estimator, so stack the full history of observations and controls; this also exempts our algorithm from assuming output observability.

In all these empirical methods, the state representation function is directly parameterized by feedforward or recurrent neural networks. The motivation of this work is to examine whether

this direct parameterization provably solves LQG, a fundamental partially observable system.

Theoretical works on controlling unknown linear systems from observations. Subramanian et al. (2020) propose two conditions for a latent model to be useful for control, namely the cost and transition prediction errors are uniformly small for all possible trajectories. Our finding is that for LQG, it suffices to have a latent model with small errors along the trajectories induced by the optimal control. The conditions that $\|Q - Q^*\|_2$, $\|A - A^*\|_2$, $\|B - B^*\|_2$ are small lead to cost and transition prediction errors that scale as the norm of the state, which is guaranteed to be bounded under the optimal controller.

(Simchowitz et al., 2020) considers online control and designs disturbance-based controller from Markov parameters. Their main consideration is robust control in the face of nonstochastic noise, while ours is stochastic noise. (Lale et al., 2020) considers online control with stochastic noise and strongly convex costs. However, both works attempt to design controller based on disturbances, instead of latent states, and they require the reconstruction of observations.

(Lale et al., 2021) considers online control based on latent models estimated from Markov parameters, assuming known positive semidefinite Q^* and positive definite R^*, relying on the analysis of the Ho-Kalman algorithm (Oymak and Ozay, 2019). (Zheng et al., 2021) is the closest to our setup, based on frequency domain identification of the Markov parameters, from which system parameters are identified. Markov parameters based approaches only apply to infinite-horizon LTI systems: it is unclear how to apply them to LTV systems; they may not even apply to finite-horizon LTI systems because they require a truncation on the order of $-\log(N)/\log(\rho(A^*))$, not necessarily satisfied by the given horizon. Even for infinite-horizon LTI systems, Markov parameter-based approach critically relies on the assumption of linear observations, limiting their practical applications, and requires the *reconstruction of observations*. In contrast, our method is closer to direct model learning methods widely used in practice, and has the potential to deal with nonlinear observation models. More recently, Uehara et al. (2022) studies statistically efficient RL in partially observable dynamical systems with function approximation, including observable LQG as an example. The algorithm in Uehara et al. (2022) is a model-free actor-critic method based on a finite-memory policy parameterization. However, it does not learn any latent model, and uses the optimistic principle that can be computationally intractable. Moreover, the sampling oracle (not directly sampling from the roll-out trajectories) and the observability assumption (the observation matrix is full column-rank) are both stronger than ours.

B   Discussion on learning the Kalman filter

We know that the Kalman filter takes the form of

$$z_0^* = L_0^* y_0, \quad z_{t+1}^* = A_t^* z_t^* + B_t^* u_t + L_{t+1}^*(y_{t+1} - C_{t+1}^*(A_t^* z_t^* + B_t^* u_t)), \quad 0 \le t \le T-1,$$

which can be parameterized by matrices $((A_t, B_t)_{t=0}^{T-1}, (C_t, L_t)_{t=0}^{T})$ in an unknown system. If we parameterize it this way, then since we will estimate $(C_t^*)_{t=0}^{T}$, we can essentially reconstruct

the observation, even if the algorithm may not explicitly do so, which defeats our purpose of avoiding it. Moreover, in this way, the representation function directly gives us the transition function $(A_t, B_t)_{t=0}^{T-1}$ of the latent model. This property is unrealistically strong in more complex scenarios: just imagine that once you know how to obtain latent states from history, this property says now you also know how the latent states evolve! Let z_t denote the latent state. If we seriously consider this approach, the loss term for the transition prediction error is

$$\sum_{i=1}^{n} \| A_t z_t^{(i)} + B_t u_t^{(i)} + L_{t+1}(y_{t+1}^{(i)} - C_{t+1}(A_t z_t^{(i)} + B_t u_t^{(i)})) - (A_t z_t^{(i)} + B_t u_t^{(i)}) \|^2$$
$$= \sum_{i=1}^{n} \| L_{t+1}(y_{t+1}^{(i)} - C_{t+1}(A_t z_t^{(i)} + B_t u_t^{(i)})) \|^2.$$

Then it is clear that we are minimizing the reconstruction error weighted by L_{t+1}.

As mentioned in Section 2.1, the Kalman filter can be alternatively written as $z_{t+1}^* = \overline{A}_t^* z_t^* + \overline{B}_t^* u_t + L_{t+1}^* y_{t+1}$, where $\overline{A}_t^* = (I - L_{t+1}^* C_{t+1}^*) A_t^*$ and $\overline{B}_t^* = (I - L_{t+1}^* C_{t+1}^*) B_t^*$. Hence, we can alternatively parameterize the Kalman filter by $((\overline{A}_t, \overline{B}_t)_{t=0}^{T-1}, (L_t)_{t=0}^{T})$, and separately parameterize the latent model by $((A_t, B_t)_{t=0}^{T-1}, (Q_t)_{t=0}^{T})$; this parameterization avoids the reconstruction issue.

Similar to CoReL (Algorithm 2), we can use the following procedure to learn $((\overline{A}_t^*, \overline{B}_t^*)_{t=0}^{T-1}, (L_t^*)_{t=0}^{T})$ incrementally: at step t, solve

$$\widehat{N}_t, \widehat{b}_t \in \underset{N_t = N_t^\top, b_t}{\arg\min} \sum_{i=1}^{n} (\| [\widehat{z}_{t-1}^{(i)}; u_{t-1}^{(i)}; y_t^{(i)}] \|_{N_t}^2 + \| u_t^{(i)} \|_{R_t^*}^2 + b_t - c_t^{(i)})^2$$

and then factorize $\widehat{N}_t$ to obtain $[\widehat{\overline{A}}_{t-1}, \widehat{\overline{B}}_{t-1}, \widehat{L}_t]$, which gives $\widehat{z}_t^{(i)} = [\widehat{\overline{A}}_{t-1}, \widehat{\overline{B}}_{t-1}, \widehat{L}_t][\widehat{z}_{t-1}^{(i)}; u_{t-1}^{(i)}; y_t^{(i)}]$ for all $i = 1, \ldots, n$, used for the next iteration. However, due to error compounding, existing techniques fail to provide sample complexity guarantees polynomially in horizon T.

For $0 \leq t \leq \ell$, since $\mathrm{Cov}(z_t^*)$ does not have full rank, we still need to use the $(M_t^*)_{t=0}^{\ell}$ parameterization. For $\ell + 1 \leq t \leq T$, to obtain guarantees on $[\widehat{\overline{A}}_{t-1}, \widehat{\overline{B}}_{t-1}, \widehat{L}_t]$ recovered from $\widehat{N}_t$, we need the lemmas on matrix factorization discussed in Appendix D.2. Using either Lemma 7 or Lemma 8 has problems. Lemma 7 has dependence on $\sigma_{\min}^{-1}([\overline{A}_t^*, \overline{B}_t^*, L_{t+1}^*])$, which may well exceed one. Compounding with time then brings exponential dependence on horizon T. One may wonder if Lemma 8 helps, since it does not depend on minimum singular values. The problem with Lemma 8 is that the rate suffers from a square root decrease, which incurs double exponential dependence on horizon T.

Of course, once we obtain $(\widehat{z}_t^{(i)})_{t=\ell}^{T}$ from $(\widehat{M}_t)_{t=\ell}^{T}$, we can fit the Kalman filter afterwards; still, directly recovering Kalman filter by cost-driven state representation learning remains challenging.

C Propositions and auxiliary results

In this section, we provide proofs of Propositions 1 and 2, and introduce Proposition 3. We also present auxiliary results needed in the proofs of the propositions and later analysis.

C.1 Proof of Proposition 1

Proof. By the property of the Kalman filter, $z_t^* = \mathbb{E}[x_t \mid y_{0:t}, u_{0:(t-1)}]$ is a function of the past history $(y_{0:t}, u_{0:(t-1)})$. u_t is a function of the past history $(y_{0:t}, u_{0:(t-1)})$ and some independent random variables. Since i_{t+1} is independent of the past history $(y_{0:t}, u_{0:(t-1)})$, it is independent of z_t^* and u_t. For the cost function,

$$c_t = \|z_t^*\|_{Q_t^*}^2 + \|u_t\|_{R_t^*}^2 + \|z_t^* - x_t\|_{Q_t^*}^2 + 2\langle z_t^*, x_t - z_t^* \rangle_{Q_t^*}.$$

Let $b_t = \mathbb{E}[\|x_t - z_t^*\|_{Q_t^*}^2]$ be a constant that depends on system parameters $(A_t^*, B_t^*, \Sigma_{w_t})_{t=0}^{T-1}$, $(C_t^*, \Sigma_{v_t})_{t=0}^T$ and Σ_0. Then, random variable $\gamma_t := \|z_t^* - x_t\|_{Q_t^*}^2 - b_t$ has zero mean. Since $(x_t - z_t^*)$ is Gaussian, its squared norm is subexponential. Since z_t^* and $(x_t - z_t^*)$ are independent zero-mean Gaussian random vectors (Bertsekas, 2012), their inner product η_t is a zero-mean subexponential random variable.

If the system is uniformly exponential stable (Assumption 1) and the system parameters are regular 4, then $(S_t^*)_{t=0}^T$ given by RDE (2.2) has a bounded operator norm determined by system parameters $(A_t^*, B_t^*, C_t^*, \Sigma_{w_t})_{t=0}^{T-1}$, $(\Sigma_{v_t})_{t=0}^T$ and Σ_0 (Zhang and Zhang, 2021). Since $S_t^* = \mathrm{Cov}(x_t - z_t^*)$, $\|\gamma_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$ by Lemma 1. By Assumption 1, if we apply zero control to the system, then $\|\mathrm{Cov}(z_t^*)\|_2 = \mathcal{O}(1)$. By Lemma 1, $\eta_t = \langle z_t^*, x_t - z_t^* \rangle$ satisfies $\|\eta_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$. □

C.2 Proof of Proposition 2

Proof. For $\ell \leq t \leq T$, unrolling the Kalman filter gives

$$
\begin{aligned}
z_t^* &= A_{t-1}^* z_{t-1}^* + B_{t-1}^* u_{t-1} + L_t^* i_t \\
&= A_{t-1}^* (A_{t-2}^* z_{t-2}^* + B_{t-2}^* u_{t-2} + L_{t-1}^* n_{t-1}) + B_{t-1}^* u_{t-1} + L_t^* i_t \\
&= [B_{t-1}^*, \ldots, A_{t-1}^* A_{t-2}^* \cdots A_{t-\ell+1}^* B_{t-\ell}^*][u_{t-1}; \ldots; u_{t-\ell}] + A_{t-1}^* A_{t-2}^* \cdots A_{t-\ell}^* z_{t-\ell}^* \\
&\quad + [L_t^*, \ldots, A_{t-1}^* A_{t-2}^* \cdots A_{t-\ell+1}^* L_{t-\ell+1}^*][i_t; \ldots; i_{t-\ell+1}],
\end{aligned}
$$

where $(u_\tau)_{\tau=t-\ell}^{t-1}$, $z_{t-\ell}^*$ and $(i_\tau)_{\tau=t-\ell+1}^t$ are independent. The matrix multiplied by $[u_{t-1}; \ldots; u_{t-\ell}]$ is precisely the controllability matrix $\Phi_{t-1,\ell}^c$. Then

$$
\begin{aligned}
\mathrm{Cov}(z_t^*) = \mathbb{E}[z_t^*(z_t^*)^\top] &\succcurlyeq \Phi_{t-1,\ell}^c \mathbb{E}[[u_{t-1}; \ldots; u_{t-\ell}][u_{t-1}; \ldots; u_{t-\ell}]^\top](\Phi_{t-1,\ell}^c)^\top \\
&= \sigma_u^2 \Phi_{t-1,\ell}^c (\Phi_{t-1,\ell}^c)^\top.
\end{aligned}
$$

By the controllability assumption (Assumption 2), $\mathrm{Cov}(z_t^*)$ has full rank and

$$\sigma_{\min}(\mathrm{Cov}(z_t^*)) \geq \sigma_u^2 \nu^2.$$

On the other hand, since $z_t^* = M_t^* h_t$,

$$\mathrm{Cov}(z_t^*) = \mathbb{E}[M_t^* h_t h_t^\top (M_t^*)^\top] \preccurlyeq \sigma_{\max}(\mathbb{E}[h_t h_t^\top]) M_t^* (M_t^*)^\top.$$

Since $h_t = [y_{0:t}; u_{0:(t-1)}]$ and $(\mathrm{Cov}(y_t))_{t=0}^T$, $(\mathrm{Cov}(u_t))_{t=0}^{T-1}$ have $\mathcal{O}(1)$ operator norms, by Lemma 2, $\mathrm{Cov}(h_t) = \mathbb{E}[h_t h_t^\top] = \mathcal{O}(t)$. Hence,

$$0 < \sigma_u^2 \nu^2 \leq \sigma_{\min}(\mathrm{Cov}(z_t^*)) = \mathcal{O}(t)\sigma_{d_x}^2(M_t^*).$$

This implies that $\mathrm{rank}(M_t^*) = d_x$ and $\sigma_{\min}(M_t^*) = \Omega(\nu t^{-1/2})$. □

C.3 Proposition 3

The following proposition does not appear in the main body, but is important for analyzing CoReL (Algorithm 2) in Section E.

Proposition 3. *Let $z_t^{*\prime} = \widehat{x}_t'$ be the state estimates by the Kalman filter under the normalized parameterization. If we apply $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for all $0 \le t \le T-1$, then for $0 \le t \le T$,*

$$\overline{c}_t := c_t + c_{t+1} + \ldots + c_{t+k-1} = \|z_t^{*\prime}\|^2 + \sum_{\tau=t}^{t+k-1} \|u_\tau\|_{R_\tau^*}^2 + b_t' + e_t',$$

where $k = 1$ for $0 \le t \le \ell - 1$ and $k = m \wedge (T - t + 1)$ for $\ell \le t \le T$, $b_t' = \mathcal{O}(k)$, and e_t' is a zero-mean subexponential random variable with $\|e_t'\|_{\psi_1} = \mathcal{O}(k d_x^{1/2})$.

Proof. By Proposition 1, $z_{t+1}^{*\prime} = A_t^{*\prime} z_t^{*\prime} + B_t^{*\prime} u_t + L_{t+1}^{*\prime} i_{t+1}'$, where $L_{t+1}^{*\prime}, i_{t+1}'$ are the Kalman gain and the innovation under the normalized parameterization, respectively. Under Assumptions 1 and 4, $(i_t')_{t=0}^T$ are Gaussian random vectors whose covariances have $\mathcal{O}(1)$ operator norms, and $(L_t^{*\prime})_{t=0}^T$ have $\mathcal{O}(1)$ operator norms (Zhang and Zhang, 2021). Hence, The covariance of $L_{t+1}^{*\prime} i_{t+1}'$ has $\mathcal{O}(1)$ operator norm. Since $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, $j_t := B_t^{*\prime} u_t + i_t'$ can be viewed as a Gaussian noise vector whose covariance has $\mathcal{O}(1)$ operator norm. By Proposition 1,

$$c_t = \|z_t^{*\prime}\|_{Q_t^{*\prime}}^2 + \|u_t\|_{R_t^*}^2 + b_t + e_t,$$

where $e_t := \gamma_t + \eta_t$ is subexponential with $\|e_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$. Let $\Phi_{t,t_0}' = A_{t-1}^{*\prime} A_{t-2}^{*\prime} \cdots A_{t_0}^{*\prime}$ for $t > t_0$ and $\Phi_{t,t}' = I$. Then, for $\tau \ge t$,

$$z_\tau^{*\prime} = \Phi_{\tau,t}' z_t^{*\prime} + \sum_{s=t}^{\tau-1} \Phi_{\tau,s}' j_s := \Phi_{\tau,t}' z_t^{*\prime} + j_{\tau,t}',$$

where $j_{t,t}' = 0$ and for $\tau > t$, $j_{\tau,t}'$ is a Gaussian random vector with bounded covariance due to uniform exponential stability (Assumption 1). Therefore,

$$\begin{aligned}
\overline{c}_t &= \sum_{\tau=t}^{t+k-1} c_\tau \\
&= \sum_{\tau=t}^{t+k-1} \left(\|\Phi_{\tau,t}' z_t^{*\prime} + j_{\tau,t}'\|_{Q_\tau^{*\prime}}^2 + \|u_\tau\|_{R_\tau^*}^2 + b_\tau + e_\tau \right) \\
&= (z_t^{*\prime})^\top \left(\sum_{\tau=t}^{t+k-1} (\Phi_{\tau,t}')^\top Q_t^{*\prime} \Phi_{\tau,t}' \right) z_t^{*\prime} + \sum_{\tau=t}^{t+k-1} \|u_\tau\|_{R_\tau^*}^2 \\
&\quad + \sum_{\tau=t}^{t+k-1} \left(\|j_{\tau,t}'\|_{Q_\tau^{*\prime}}^2 + (j_{\tau,t}')^\top Q_\tau^{*\prime} \Phi_{\tau,t}' z_t^{*\prime} + b_\tau + e_\tau \right) \\
&= \|z_t^{*\prime}\|^2 + \sum_{\tau=t}^{t+k-1} \|u_\tau\|_{R_\tau^*}^2 + b_t' + e_t',
\end{aligned}$$

where $\sum_{\tau=t}^{t+k-1} (\Phi_{\tau,t}')^\top Q_t^{*\prime} \Phi_{\tau,t}' = I$ is due to the normalized parameterization, $b_t' := \sum_{\tau=t}^{t+k-1} (b_\tau + \mathbb{E}[\|j_\tau'\|_{Q_\tau^{*\prime}}^2]) = \mathcal{O}(k)$, and

$$e_t' := \sum_{\tau=t}^{t+k-1} \left(\|j_\tau'\|_{Q_\tau^{*\prime}}^2 - \mathbb{E}[\|j_\tau'\|_{Q_\tau^{*\prime}}^2] + (j_\tau')^\top Q_\tau^{*\prime} \Phi_{\tau,t}' z_t^{*\prime} + e_\tau \right)$$

has zero mean and is subexponential with $\|e_t'\|_{\psi_1} = \mathcal{O}(k d_x^{1/2})$. $\qquad\square$

C.4 Auxiliary results

Lemma 1. *Let $x \sim \mathcal{N}(0, \Sigma_x)$ and $y \sim \mathcal{N}(0, \Sigma_y)$ be d-dimensional Gaussian random vectors. Let Q be a $d \times d$ positive semidefinite matrix. Then, there exists an absolute constant $c > 0$ such that*

$$\|\langle x, y \rangle_Q\|_{\psi_1} \leq c\sqrt{d}\|Q\|_2\sqrt{\|\Sigma_x\|_2\|\Sigma_y\|_2}.$$

Proof. Since $|\langle x, y \rangle_Q| = |x^\top Q y| \leq \|x\|\|Q\|_2\|y\|$,

$$\|\langle x, y \rangle_Q\|_{\psi_1} = \||\langle x, y \rangle_Q|\|_{\psi_1} \leq \|\|x\|\|Q\|_2\|y\|\|_{\psi_1} = \|Q\|_2 \cdot \|\|x\|\|y\|\|_{\psi_1}.$$

For $x \sim \mathcal{N}(0, \Sigma_x)$, we know that $\|x\|$ is subgaussian. Actually, by writing $x = \Sigma_x^{1/2}g$ for $g \sim \mathcal{N}(0, I)$, we have

$$\|\|x\|\|_{\psi_2} = \|\|\Sigma_x^{1/2}g\|\|_{\psi_2} \leq \|\|\Sigma_x^{1/2}\|_2\|g\|\|_{\psi_2} = \|\Sigma_x^{1/2}\|_2\|\|g\|\|_{\psi_2}.$$

The distribution of $\|g\|_2$ is known as χ distribution, and we know that $\|\|g\|\|_{\psi_2} = c'd^{1/4}$ for an absolute constant $c' > 0$. Hence, $\|\|x\|\|_{\psi_2} \leq c'd^{1/4}\|\Sigma_x\|_2^{1/2}$. Similarly, $\|\|y\|\|_{\psi_2} \leq c'd^{1/4}\|\Sigma_y\|_2^{1/2}$. Since $\|\|x\|\|y\|\|_{\psi_1} \leq \|\|x\|\|_{\psi_2}\|\|y\|\|_{\psi_2}$ (see, e.g., (Vershynin, 2018, Lemma 2.7.7)), we have

$$\|\langle x, y \rangle_Q\|_{\psi_1} \leq (c')^2\sqrt{d}\|Q\|_2\sqrt{\|\Sigma_x\|_2\|\Sigma_y\|_2}.$$

Taking $c = (c')^2$ concludes the proof. $\qquad\square$

Lemma 2. *Let x, y be random vectors of dimensions d_x, d_y, respectively, defined on the same probability space. Then, $\|\mathrm{Cov}([x; y])\|_2 \leq \|\mathrm{Cov}(x)\|_2 + \|\mathrm{Cov}(y)\|_2$.*

Proof. Let $\mathrm{Cov}([x; y]) = DD^\top$ be a factorization of the positive semidefinite matrix $\mathrm{Cov}([x; y])$, where $D \in \mathbb{R}^{(d_x+d_y)\times(d_x+d_y)}$. Let D_x and D_y be the matrices consisting of the first d_x rows and the last d_y rows of D, respectively. Then,

$$\mathrm{Cov}([x; y]) = DD^\top = [D_x; D_y][D_x^\top, D_y^\top] = \begin{bmatrix} D_xD_x^\top & D_xD_y^\top \\ D_yD_x^\top & D_yD_y^\top \end{bmatrix}.$$

Hence, $\mathrm{Cov}(x) = D_xD_x^\top$ and $\mathrm{Cov}(y) = D_yD_y^\top$. The proof is completed by noticing that

$$\begin{aligned} \|\mathrm{Cov}([x; y])\|_2 = \|D^\top D\|_2 &= \|[D_x^\top, D_y^\top][D_x; D_y]\|_2 \\ &= \|D_x^\top D_x + D_y^\top D_y\|_2 \\ &\leq \|D_x^\top D_x\|_2 + \|D_y^\top D_y^\top\|_2 = \|\mathrm{Cov}(x)\|_2 + \|\mathrm{Cov}(y)\|_2. \end{aligned}$$

$\qquad\square$

Lemma 3. *Let x and y be random vectors defined on the same probability space. Then, $\|\mathrm{Cov}(x, y)\|_2 \leq \|\mathrm{Cov}(x)^{1/2}\|_2\|\mathrm{Cov}(y)^{1/2}\|_2$.*

Proof. Let d_x, d_y denote the dimensions of x, y, respectively. For any vectors $v \in \mathbb{R}^{d_x}$, $w \in \mathbb{R}^{d_y}$, by the Cauchy-Schwarz inequality,

$$
\begin{aligned}
(v^\top \mathrm{Cov}(x,y)w)^2 &= (\mathbb{E}[v^\top xy^\top w])^2 \\
&\leq \mathbb{E}[(v^\top x)^2]\mathbb{E}[(w^\top y)^2] \\
&= \mathbb{E}[v^\top xx^\top v]\mathbb{E}[w^\top yy^\top w] \\
&= (v^\top \mathrm{Cov}(x)v)(w^\top \mathrm{Cov}(y)w).
\end{aligned}
$$

Taking the maximum over v, w on both sides subject to $\|v\|, \|w\| \leq 1$ gives

$$
\|\mathrm{Cov}(x,y)\|_2^2 \leq \|\mathrm{Cov}(x)\|_2 \|\mathrm{Cov}(y)\|_2,
$$

which completes the proof. $\qquad\square$

D Lemmas in the analysis

D.1 Quadratic regression bound

Recall that $\mathrm{svec}(\cdot)$ denotes the flattening of a symmetric matrix that does not repeat the off-diagonal elements twice, but scale them by $\sqrt{2}$, so that the Euclidean norm of the resulting vector equals the Frobenius norm of the symmetric matrix.

As noted in Section 3.1, the quadratic regression can be converted to linear regression using $\|h\|_P^2 = \langle hh^\top, P\rangle_F = \langle \mathrm{svec}(hh^\top), \mathrm{svec}(P)\rangle$. To analyze this linear regression, we need the following lemma from (Jadbabaie et al., 2021).

Lemma 4 ((Jadbabaie et al., 2021, Proposition 1)). *Let $(h_0^{(i)})_{i=1}^n$ be n independent observations of the r-dimensional random vector $h_0 \sim \mathcal{N}(0, I_r)$. Let $f_0^{(i)} := \mathrm{svec}(h_0^{(i)}(h_0^{(i)})^\top)$. There exists an absolute constant $c > 0$, such that as long as $n \geq cr^4 \log(cr^2/p)$, with probability at least $1 - p$,*

$$
\sigma_{\min}\left(\sum_{i=1}^n f_0^{(i)}(f_0^{(i)})^\top \right) \geq n.
$$

Lemma 4 lower bounds the minimum singular value of a matrix whose elements are fourth powers of elements in standard Gaussian random vectors. The next lemma allows us to extend the lower bound to general Gaussian random vectors.

Lemma 5. *Let $D = PL$, where P is a $d \times d$ permutation matrix and L is a $d \times r$ lower triangular matrix with positive diagonal elements. For any $h_0 \in \mathbb{R}^r$, define $h := Dh_0 \in \mathbb{R}^d$. Let $f_0 := \mathrm{svec}(h_0 h_0^\top)$ and $f := \mathrm{svec}(hh^\top)$. Then $f = \Phi f_0$, where $\Phi = U_d (D \otimes D) U_r^\top$ is a $\frac{d(d+1)}{2} \times \frac{r(r+1)}{2}$ matrix for some semi-orthonormal matrices U_d, U_r, and $\otimes$ denotes the Kronecker product. Moreover, $\mathrm{rank}(\Phi) = \frac{r(r+1)}{2}$.*

Proof. Recall that svec and vec can be related with a rectangular semi-orthonormal matrix (Schacke, 2004). Specifically, one can construct a $\frac{d(d+1)}{2} \times d^2$ matrix U_d that satisfies $U_d U_d^\top = I_{\frac{d(d+1)}{2}}$, such that for any $d \times d$ symmetric matrix A,

$$
\mathrm{svec}(A) = U_d \mathrm{vec}(A), \quad \mathrm{vec}(A) = U_d^\top \mathrm{svec}(A).
$$

23

Therefore,

$$f = \text{svec}(hh^\top) = \text{svec}(Dh_0 h_0^\top D^\top)$$
$$= U_d \text{vec}(Dh_0 h_0^\top D^\top)$$
$$= U_d(D \otimes D)\text{vec}(h_0 h_0^\top)$$
$$= U_d(D \otimes D)U_r^\top \text{svec}(h_0 h_0^\top) = \Phi f_0,$$

where U_d and U_r are the $\frac{d(d+1)}{2} \times d^2$ and $\frac{r(r+1)}{2} \times r^2$ matrices that satisfy $U_d U_d^\top = I_{\frac{d(d+1)}{2}}$ and $U_r U_r^\top = I_{\frac{r(r+1)}{2}}$, and $\Phi := U_d(D \otimes D)U_r^\top$. This proves the first part of the lemma.

By the properties of the Kronecker product (Schacke, 2004), we know that $(D \otimes D)$ is positive definite, and in fact, $\sigma_{\min}(D \otimes D) = \sigma_{\min}(D)^2$. However, it is unclear whether Φ has full column rank from its expression. To see why Φ has full column rank, let us expand Φ. Let $(l_{ii})_{i=1}^r$ be the diagonal elements of L. The permutation matrix P rearranges the rows of L and does not affect the singularity of Φ. By rearranging the rows and columns of Φ, we find that the $\frac{d(d+1)}{2} \times \frac{r(r+1)}{2}$ matrix Φ can be made lower triangular, with diagonal elements being $l_{ii}l_{jj}$ for all $1 \leq i \leq j \leq r$. Hence, M has full column rank. That is, $\text{rank}(\Phi) = \frac{r(r+1)}{2}$. $\qquad\square$

Note that if $d = r$, that is, $\Sigma := DD^\top$ has full rank, we can alternatively write the Φ in Lemma 5 as $\Phi := \Sigma^{1/2} \otimes_s \Sigma^{1/2}$. Then, by the properties of the symmetric Kronecker product (Schacke, 2004),

$$\sigma_{\min}(\Phi) = \sigma_{\min}(\Sigma^{1/2})^2 = \sigma_{\min}(\Sigma) > 0.$$

The following lemma is the main result in Section D.1.

Lemma 6 (Quadratic regression). *Define random variable $c := (h^*)^\top N^* h^* + b^* + e$, where $h^* \sim \mathcal{N}(0, \Sigma_*)$ is a d-dimensional Gaussian random vector, $N^* \in \mathbb{R}^{d \times d}$ is a positive semidefinite matrix, $b^* \in \mathbb{R}$ is a constant and e is a zero-mean subexponential random variable with $\|e\|_{\psi_1} \leq E$. Assume that $\|N^*\|_2$ and $\|\Sigma_*\|_2$ are $\mathcal{O}(1)$. Let $\sigma_{\min}(\Sigma_*^{1/2}) \geq \beta > 0$. Define $h := h^* + \delta$ where the Gaussian noise vector δ can be correlated with h^* and its covariance matrix Σ_δ satisfies $\|\Sigma_\delta^{1/2}\|_2 \leq \epsilon \leq \beta/2$. Suppose we get n independent observations $c^{(i)}$ of c and $h^{(i)}$ of h. Consider the regression problem*

$$\widehat{N}, \widehat{b} \in \underset{N=N^\top, b}{\arg\min} \sum_{i=1}^n (c^{(i)} - \|h^{(i)}\|_N^2 - b)^2. \tag{D.1}$$

There exists an absolute constant $c > 0$, such that as long as $n \geq cd^4 \log(cd^2/p)\log(1/p)$, with probability at least $1 - p$,

$$\|\widehat{N} - N^*\|_F = \mathcal{O}(\beta^{-2}d(\epsilon + En^{-1/2})), \quad |\widehat{b} - b^*| = \mathcal{O}(d\epsilon + En^{-1/2}).$$

Proof. Let $\Sigma = \text{Cov}(h)$ and $r = \text{rank}(\Sigma)$. By the Cholesky decomposition, there exist a $d \times d$ permutation matrix P and a $d \times r$ lower triangular matrix L with positive diagonal elements, such that

$$\Sigma = P[L, 0_{d \times (d-r)}][L^\top; 0_{(d-r) \times d}]P^\top = PLL^\top P^\top = DD^\top,$$

24

where we define $D := PL$. We can parameterize h by Dh_0, where $h_0 \sim \mathcal{N}(0, I_r)$ is an r-dimensional standard Gaussian random vector. Correspondingly, an independent observation $h^{(i)}$ can be expressed as $Gh_0^{(i)}$, where $h_0^{(i)}$ is an independent observation of h_0. It follows that $hh^\top = Dh_0 h_0^\top D^\top$ and $h^{(i)}(h^{(i)})^\top = Dh_0^{(i)}(h_0^{(i)})^\top D^\top$. Let $f_0 := \text{svec}(h_0 h_0^\top)$. By Lemma 5,

$$f = \text{svec}(hh^\top) = \text{svec}(Dh_0 h_0^\top D^\top) = \Phi f_0,$$

where Φ is a $\frac{d(d+1)}{2} \times \frac{r(r+1)}{2}$ matrix that has full column rank. Define $f_0^{(i)} := \text{svec}(h_0^{(i)}(h_0^{(i)})^\top)$ and $F_0 := [f_0^{(1)}, \ldots, f_0^{(n)}]^\top$ be an $n \times \frac{r(r+1)}{2}$ matrix whose ith row is $(f_0^{(i)})^\top$. Define $f^{(i)}, F, (f^*)^{(i)}, F^*$ similarly. Then $F = F_0 \Phi^\top$. The regression D.1 can be written as

$$\text{svec}(\widehat{N}), \widehat{b} \in \underset{\text{svec}(N), b}{\arg\min} \sum_{i=1}^n \left(c^{(i)} - (f^{(i)})^\top \text{svec}(N) - b \right)^2 \qquad (D.2)$$

Solving linear regression (D.2) for $\text{svec}(N)$ gives

$$F^\top F \text{svec}(\widehat{N}) = \sum_{i=1}^n f^{(i)}(c^{(i)} - b).$$

Substituting $c^{(i)} = ((f^*)^{(i)})^\top \text{svec}(N^*) + b^* + e^{(i)}$ into the above equation yields

$$F^\top F \text{svec}(\widehat{N}) = F^\top F^* \text{svec}(N^*) + (b^* - \widehat{b}) \sum_{i=1}^n f^{(i)} + \sum_{i=1}^n e^{(i)} f^{(i)}$$
$$= F^\top F^* \text{svec}(N^*) + (b^* - \widehat{b}) F^\top 1_n + F^\top \varepsilon,$$

where ε denotes the vector whose ith element is $e^{(i)}$. Substituting $F = F_0 \Phi^\top$ into the above equation yields

$$\Phi F_0^\top F_0 \Phi^\top \text{svec}(\widehat{N}) = \Phi F_0^\top F^* \text{svec}(N^*) + \Phi F_0^\top 1_n (b^* - \widehat{b}) + \Phi F_0^\top \varepsilon.$$

Since Φ has full column rank, this gives

$$F_0^\top F_0 \Phi^\top \text{svec}(\widehat{N}) = F_0^\top F^* \text{svec}(N^*) + F_0^\top 1_n (b^* - \widehat{b}) + F_0^\top \varepsilon. \qquad (D.3)$$

Solving the regression (D.2) for b and substituting $c^{(i)} = ((f^*)^{(i)})^\top \text{svec}(N^*) + b^* + e^{(i)}$ into the resulting equation yield

$$\widehat{b} = b^* + \frac{1}{n}(1_n^\top F^* \text{svec}(N^*) - 1_n^\top F \text{svec}(\widehat{N}) + 1_n^\top \varepsilon). \qquad (D.4)$$

By substituting (D.4) into (D.3) and rearranging the terms, we have

$$F_0^\top (I_n - \frac{1_n 1_n^\top}{n}) F_0 \Phi^\top \text{svec}(\widehat{N} - N^*) = F_0^\top (I_n - \frac{1_n 1_n^\top}{n})(F^* - F)\text{svec}(N^*) + F_0^\top (I_n - \frac{1_n 1_n^\top}{n})\varepsilon.$$

Note that $J_n := I_n - 1_n 1_n^\top / n$ is a positive definite matrix with $n-1$ eigenvalues being ones and the other being $1 - 1/n$, so the eigenvalues of $F_0^\top J_n F_0$ do not differ much from those of $F_0^\top F_0$. In particular, since $F_0^\top J_n F_0 \succcurlyeq \sigma_{\min}(J_n) F_0^\top F_0$, $\sigma_{\min}(F_0^\top J_n F_0) \geq (1 - 1/n)\sigma_{\min}(F_0^\top F_0)$. Let $\widetilde{F}_0 := J_n^{1/2} F_0$, $\widetilde{F} := J_n^{1/2} F$ and $\widetilde{F}^* := J_n^{1/2} F^*$. Then we have

$$\|\Phi^\top \text{svec}(\widehat{N} - N^*)\| = \|\widetilde{F}_0^\dagger (\widetilde{F}^* - \widetilde{F})\text{svec}(N^*) + \widetilde{F}_0^\dagger J_n^{1/2} \varepsilon\|$$
$$\leq \underbrace{\|\widetilde{F}_0^\dagger (\widetilde{F}^* - \widetilde{F})\text{svec}(N^*)\|}_{(a)} + \underbrace{\|\widetilde{F}_0^\dagger J_n^{1/2} \varepsilon\|}_{(b)}, \qquad (D.5)$$

25

where we have assumed $F_0^\top F_0$ (and hence $\widetilde{F}_0^\top \widetilde{F}_0$) is invertible, which holds with probability at least $1 - p$ if $n \geq cd^4 \log(cd^2/p)$ for an absolute constant $c > 0$ by Lemma 4.

Term (a) is upper bounded by

$$\sigma_{\min}(\widetilde{F}_0)^{-1}\|(\widetilde{F}^* - \widetilde{F})\mathrm{svec}(N^*)\| = \mathcal{O}(\sigma_{\min}(F_0)^{-1})\|J_n^{1/2}(F^* - F)\mathrm{svec}(N^*)\|$$
$$= \mathcal{O}(n^{-1/2})\|(F^* - F)\mathrm{svec}(N^*)\|,$$

with probability at least $1 - p$, due to Lemma 4. Using arguments similar to those in (Mhammedi et al., 2020, Section B.2.13), we have

$$\|(F^* - F)\mathrm{svec}(N^*)\|^2 = \sum_{i=1}^n \left\langle \mathrm{svec}((h^*)^{(i)}((h^*)^{(i)})^\top) - \mathrm{svec}(h^{(i)}(h^{(i)})^\top), \mathrm{svec}(N^*) \right\rangle$$

$$\overset{(i)}{=} \sum_{i=1}^n \left\langle (h^*)^{(i)}((h^*)^{(i)})^\top - h^{(i)}(h^{(i)})^\top, N^* \right\rangle_F$$

$$\overset{(ii)}{\leq} \|N^*\|_2^2 \sum_{i=1}^n \|(h^*)^{(i)}((h^*)^{(i)})^\top - h^{(i)}(h^{(i)})^\top\|_*^2$$

$$\overset{(iii)}{\leq} 2\|N^*\|_2^2 \sum_{i=1}^n \|(h^*)^{(i)}((h^*)^{(i)})^\top - h^{(i)}(h^{(i)})^\top\|_F^2 = 2\|N^*\|_2^2\|F^* - F\|_F^2,$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius product between matrices in (i), $\|\cdot\|_*$ denotes the nuclear norm in (ii), and (iii) follows from the fact that the matrix $(h^*)^{(i)}((h^*)^{(i)})^\top - h^{(i)}(h^{(i)})^\top$ has at most rank two. Now we want to show that $\|F^* - F\|_F^2$ is on the order of n. To see this, first note that since $h = h^* + \delta$, we have $\Sigma = \Sigma_* + \Sigma_\delta + 2\mathrm{Cov}(h^*, \delta)$, and

$$\|\Sigma\|_2 = \|\Sigma_*\|_2 + \|\Sigma_\delta\|_2 + 2\|\mathrm{Cov}(h^*, \delta)\|_2$$

$$\overset{(i)}{\leq} \|\Sigma_*\|_2 + \|\Sigma_\delta\|_2 + 2(\|\Sigma_*\|_2\|\Sigma_\delta\|_2)^{1/2}$$

$$\leq 2(\|\Sigma_*\|_2 + \|\Sigma_\delta\|_2) = \mathcal{O}(\|\Sigma_*\|_2),$$

where (i) is due to Lemma 3. Then,

$$\mathbb{E}[\|h^*(h^*)^\top - hh^\top\|_F^2] = \mathbb{E}[\|h^*(h^* - h)^\top + (h^* - h)h^\top\|_F^2]$$

$$\leq \mathbb{E}[2(\|h^*\|^2 + \|h\|^2)\|\delta\|^2]$$

$$\overset{(i)}{=} \mathcal{O}(1)(\mathbb{E}[\|h^*\|^4] \cdot \mathbb{E}[\|\delta\|^4])^{1/2},$$

where (i) is due to the Cauchy-Schwarz inequality and $\|\Sigma\|_2 = \mathcal{O}(\|\Sigma_*\|_2)$. Since h^* and δ are both Gaussian random vectors, we know that $\mathbb{E}[\|h^*\|^4] = \mathcal{O}(d^2\|\Sigma_*\|_2^2)$ and $\mathbb{E}[\|\delta\|^4] = \mathcal{O}(d^2\|\Sigma_\delta\|_2^2)$. Then

$$\mathbb{E}[\|h^*(h^*)^\top - hh^\top\|_F^2] = \mathcal{O}(d^2\|\Sigma_*\|_2\|\Sigma_\delta\|_2).$$

Each element in $h^*(h^*)^\top - hh^\top = ((h^* + h)(h^* - h)^\top) = -(h^* + h)\delta^\top$ is the product of two Gaussian random variables and hence, is subexponential. Although $\|h^*(h^*)^\top - hh^\top\|_F^2$, as a sum of the squares of subexponential variables, is not subexponential, it is indeed 1/2-subweibull (Kuchibhotla and Chakrabortty, 2018; Hao et al., 2019; Zhang and Wei, 2022), and

has similar concentration inequalities to those of subexponential distributions. Its subweibull norm is given by

$$\left\| \|h^*(h^*)^\top - hh^\top\|_F^2 \right\|_{\phi_{1/2}} = \mathcal{O}(d^2 \|\Sigma_*\|_2 \|\Sigma_\delta\|_2),$$

since each element in $h^*(h^*)^\top - hh^\top$ has a subexponential norm bounded by $\mathcal{O}(\|\Sigma_*^{1/2}\| \|\Sigma_\delta^{1/2}\|)$. By (Hao et al., 2019, Theorem 3.1),

$$
\begin{aligned}
\|F^* - F\|_F^2 &= \sum_{i=1}^n \|(h^*)^{(i)}((h^*)^{(i)})^\top - h^{(i)}(h^{(i)})^\top\|_F^2 \\
&= \mathcal{O}(1) d^2 \|\Sigma_*\|_2 \|\Sigma_\delta\|_2 (n + \sqrt{n}\log(1/p)) = \mathcal{O}(d^2 \|\Sigma_*\|_2 \|\Sigma_\delta\|_2 n).
\end{aligned}
$$

Hence, we obtain that the term (a) in (D.5) satisfies

$$(a) = \mathcal{O}(d \|N^*\|_2 \|\Sigma_*^{1/2}\|_2 \|\Sigma_\delta^{1/2}\|_2) = \mathcal{O}(d \|N^*\|_2 \|\Sigma_*^{1/2}\|_2 \epsilon).$$

Now we consider term (b) in (D.5):

$$(b) = \|\widetilde{F}_0^\dagger J_n^{1/2} \varepsilon\| = \|(\widetilde{F}_0^\top \widetilde{F}_0)^{-1} \widetilde{F}_0^\top J_n^{1/2} \varepsilon\| \le \sigma_{\min}(\widetilde{F}_0^\top \widetilde{F}_0)^{-1} \|F_0^\top J_n \varepsilon\| = \mathcal{O}(n^{-1}) \|F_0^\top \varepsilon'\|,$$

where $\varepsilon' := J_n \varepsilon = (I_n - 1_n 1_n^\top / n) \varepsilon$. Since ε is a vector consisting of i.i.d. subexponential random variables with subponential norm bounded by E, the ith element $(e')^{(i)}$ in ε' is also subexponential, with the subexponential norm bounded by

$$\sqrt{(1 - \frac{1}{n})^2 + \frac{n-1}{n^2}} E = \sqrt{\frac{n-1}{n}} E \le E.$$

Note that $\|F_0^\top \varepsilon'\| = \|\sum_{i=1}^n f_0^{(i)} (e')^{(i)}\|$. Consider the jth component in the summation $\sum_{i=1}^n [f_0^{(i)}]_j (e')^{(i)}$. Recall that $f_0 = \text{svec}(h_0 h_0^\top)$, so $[f_0]_j$ is either the square of a standard Gaussian random variable or $\sqrt{2}$ times the product of two independent standard Gaussian random variables. Hence, $[f_0]_j$ is subexponential with $\|[f_0]_j\|_{\psi_1} = \mathcal{O}(1)$. As a result, the product $[f_0]_j e'$ is $\frac{1}{2}$-subweibull, with the subweibull norm being $\mathcal{O}(E)$. By (Hao et al., 2019, Theorem 3.1),

$$\sum_{i=1}^n [f_0^{(i)}]_j (e')^{(i)} = \mathcal{O}(En^{1/2}).$$

Hence, the norm of the $d(d+1)/2$-dimensional vector $F_0^\top \varepsilon'$ is $\mathcal{O}(dEn^{1/2})$. As a result, term (b) is $\mathcal{O}(dEn^{-1/2})$. Combining the bounds on (a) and (b), we have

$$\|\Phi^\top \text{svec}(\widehat{N} - N^*)\| = \mathcal{O}(1)(d\|N^*\|_2 \|\Sigma_*^{1/2}\|_2 \epsilon + dEn^{-1/2}) = \mathcal{O}(d\epsilon + dEn^{-1/2}). \tag{D.6}$$

By Lemma 5, $\Phi = U_d(D \otimes D)U_r$, where U_d, U_r are semi-orthonormal matrices that connect vec and svec. Hence,

$$
\begin{aligned}
\Phi^\top \text{svec}(\widehat{N} - N^*) &= U_r(D \otimes D)^\top U_d^\top \text{svec}(\widehat{N} - N^*) \\
&= U_r(D^\top \otimes D^\top)\text{vec}(\widehat{N} - N^*) \\
&= U_r \text{vec}(D^\top(\widehat{N} - N^*)D) \\
&= \text{svec}(D^\top(\widehat{N} - N^*)D).
\end{aligned}
$$

It follows that

$$\|\Phi^\top \mathrm{svec}(\widehat{N} - N^*)\| = \|\mathrm{svec}(D^\top(\widehat{N} - N^*)D)\|$$
$$= \|D^\top(\widehat{N} - N^*)D\|_F$$
$$= (\mathrm{tr}(D^\top(\widehat{N} - N^*)DD^\top(\widehat{N} - N^*)D))^{1/2}$$
$$= (\mathrm{tr}((\widehat{N} - N^*)DD^\top(\widehat{N} - N^*)DD^\top))^{1/2}$$
$$= (\mathrm{tr}((\widehat{N} - N^*)\Sigma(\widehat{N} - N^*)\Sigma))^{1/2}$$
$$= \|\Sigma^{1/2}(\widehat{N} - N^*)\Sigma^{1/2}\|_F.$$

Hence, the bound we prove in (D.6) also applies to $\|\Sigma^{1/2}(\widehat{N} - N^*)\Sigma^{1/2}\|_F$. Note that our assumption that $\|\Sigma_\delta\|_2 \leq \sigma_{\min}(\Sigma_*)/2$ ensures $\sigma_{\min}(\Sigma) = \Omega(\sigma_{\min}(\Sigma_*))$. Since $\|\Sigma^{1/2}(\widehat{N} - N^*)\Sigma^{1/2}\|_F \geq \sigma_{\min}(\Sigma)\|\widehat{N} - N^*\|_F$, we have

$$\|\widehat{N} - N^*\|_F = \mathcal{O}(\sigma_{\min}(\Sigma_*)^{-1})(d\|N^*\|_2\|\Sigma_*^{1/2}\|_2\epsilon + dEn^{-1/2}) = \mathcal{O}(\beta^{-2}d(\epsilon + En^{-1/2})).$$

From (D.4), we have

$$|\widehat{b} - b^*| = \underbrace{\left|\frac{1}{n}1_n^\top(F^* - F)\mathrm{svec}(N^*)\right|}_{(c)} + \underbrace{\left|\frac{1}{n}1_n^\top F_0\Phi^\top\mathrm{svec}(N^* - \widehat{N})\right|}_{(d)} + \underbrace{\left|\frac{1}{n}1_n^\top\varepsilon\right|}_{(e)}.$$

We have proved that $\|(F^* - F)\mathrm{svec}(N^*)\|_2 = \mathcal{O}(d\|N^*\|_2\|\Sigma_*^{1/2}\|_2\epsilon n^{1/2})$. Since $\|1_n/n\| = n^{-1/2}$,

$$(c) = \left|\frac{1}{n}1_n^\top(F^* - F)\mathrm{svec}(N^*)\right| = \mathcal{O}(d\epsilon).$$

On the other hand,

$$\|1_n^\top F_0\|_2 = \left\|\sum_{i=1}^n f_0^{(i)}\right\|_2 = \mathcal{O}(dn^{1/2}),$$

since the jth element $[f_0]_j$ of f_0 is subexponential with $\|[f_0]_j\|_{\psi_1} = O(1)$. Hence,

$$(d) = \left|\frac{1}{n}1_n^\top F_0\Phi^\top\mathrm{svec}(N^* - \widehat{N})\right| = \mathcal{O}(dn^{-1/2})\|\Phi^\top\mathrm{svec}(\widehat{N} - N^*)\|_2 = \mathcal{O}(d^2\epsilon n^{-1/2} + d^2 En^{-1}).$$

Combining with $(e) = n^{-1}|1_n^\top\varepsilon| = \mathcal{O}(En^{-1/2})$, we have

$$|\widehat{b} - b^*| = \mathcal{O}(d\|N^*\|_2\|\Sigma_*^{1/2}\|_2\epsilon + En^{-1/2}) = \mathcal{O}(d\epsilon + En^{-1/2}),$$

where the $\mathcal{O}(d^2\epsilon n^{-1/2})$ and $\mathcal{O}(d^2 En^{-1})$ terms are absorbed into $\mathcal{O}(d\epsilon)$ and $\mathcal{O}(En^{-1/2})$, respectively, by our choice of n. $\qquad\square$

D.2 Matrix factorization bound

Given two $m \times n$ matrices A, B, we are interested in bounding $\min_{S^\top S=I}\|SA - B\|_F$ using $\|A^\top A - B^\top B\|_F$. The minimum problem is known as the orthogonal Procrustes problem, solved by Schoenemann (1964); Schönemann (1966). Specifically, the minimum is attained at $S = UV^\top$, where $U\Sigma V^\top = BA^\top$ is its singular value decomposition.

If $m \leq n$ and $\mathrm{rank}(A) = n$, then the following lemma from (Tu et al., 2016) establishes that the distance between A and B is of the same order of $\|A^\top A - B^\top B\|_F$.

28

Lemma 7 ((Tu et al., 2016, Lemma 5.4)). *For $m \times n$ matrices A, B, let $\sigma_m(A)$ denote its mth largest singular value. Then*

$$\min_{S^\top S = I} \|SA - B\|_F \leq (2(\sqrt{2} - 1))^{-1/2} \sigma_m(A)^{-1} \|A^\top A - B^\top B\|_F.$$

If $\sigma_{\min}(A)$ equals zero, the above bound becomes vacuous. In general, the following lemma shows that the distance is of the order of the square root of the $\|A^\top A - B^\top B\|_F$, with a multiplicative $\sqrt{d}$ factor, where $d = \min(2m, n)$.

Lemma 8. *For $m \times n$ matrices A, B, $\min_{S^\top S = I} \|SA - B\|_F^2 \leq \sqrt{d} \|A^\top A - B^\top B\|_F$, where $d = \min(2m, n)$.*

Proof. Let $U\Sigma V^\top = BA^\top$ be its singular value decomposition. By substituting the solution $UV^\top$ of the orthogonal Procrustes problem, the square of the attained minimum equals

$$\|UV^\top A - B\|_F^2 = \left\langle UV^\top A - B, UV^\top A - B \right\rangle_F$$
$$= \|A\|_F^2 + \|B\|_F^2 - 2\left\langle UV^\top A, B \right\rangle_F$$
$$\overset{(i)}{=} \|A\|_F^2 + \|B\|_F^2 - 2\mathrm{tr}(\Sigma)$$
$$= \|A^\top A\|_* + \|B^\top B\|_* - 2\|BA^\top\|_*,$$

where (i) is due to the property of U, V.

To establish the relationship between $\|A^\top A\|_* + \|B^\top B\|_* - 2\|BA^\top\|_*$ and $\|A^\top A - B^\top B\|_F$, we need to operate in the space of singular values. For $m \times n$ matrix M, let $(\sigma_1(M), \ldots, \sigma_{d'}(M))$ be its singular values in descending order, where $d' = m \wedge n$.

In terms of singular values,

$$\|A^\top A\|_* + \|B^\top B\|_* - 2\|BA^\top\|_* \overset{(i)}{=} \|A^\top A + B^\top B\|_* - 2\|BA^\top\|_*$$
$$\overset{(ii)}{=} \sum_{i=1}^{d} \sigma_i(A^\top A + B^\top B) - 2\sum_{i=1}^{d'} \sigma_i(BA^\top),$$

where (i) holds since $A^\top A$ and $B^\top B$ are positive semidefinite matrices, and in (ii) $d :=$ $\min(2m, n)$ since $\mathrm{rank}(A^\top A + B^\top B) \leq n$ and $\mathrm{rank}(A^\top A + B^\top B) \leq \mathrm{rank}(A^\top A) + \mathrm{rank}(B^\top B) \leq 2m$. If $x \geq y > 0$, then $x - y \leq \sqrt{x^2 - y^2}$. For all $1 \leq i \leq d$, $2\sigma_i(BA^\top) \leq \sigma_i(A^\top A + B^\top B)$ (Bhatia and Kittaneh, 1990). Take $\sigma_i(A^\top A + B^\top B)$ as x and $2\sigma_i(BA^\top)$ as y; it follows that

$$\sigma_i(A^\top A + B^\top B) - 2\sigma_i(BA^\top) \leq \sqrt{\sigma_i^2(A^\top A + B^\top B) - 4\sigma_i^2(BA^\top)}.$$

Let $\sigma_i(BA^\top) := 0$ for $d' < i \le d$. Combining the above yields

$$
\|A^\top A\|_* + \|B^\top B\|_* - 2\|BA^\top\|_*
$$

$$
\le \sum_{i=1}^{d} \sqrt{\sigma_i^2(A^\top A + B^\top B) - 4\sigma_i^2(BA^\top)}
$$

$$
\overset{(i)}{\le} \sqrt{d}\sqrt{\sum_{i=1}^{d} \sigma_i^2(A^\top A + B^\top B) - 4\sum_{i=1}^{d'} \sigma_i^2(BA^\top)}
$$

$$
\overset{(ii)}{\le} \sqrt{d}\sqrt{\|A^\top A + B^\top B\|_F^2 - 4\left\langle A^\top A, B^\top B\right\rangle_F}
$$

$$
\le \sqrt{d}\|A^\top A - B^\top B\|_F,
$$

where (i) is due to the Cauchy-Schwarz inequality, and (ii) uses

$$
\sum_{i=1}^{d'} \sigma_i^2(BA^\top) = \|BA^\top\|_F^2 = \mathrm{tr}(AB^\top BA^\top) = \mathrm{tr}(A^\top AB^\top B) = \left\langle A^\top A, B^\top B\right\rangle_F.
$$

This completes the proof. □

D.3 Linear regression bound

A standard assumption in analyzing linear regression $y = A^* x + e$ is that $\mathrm{Cov}(x)$ has *full rank*. However, as discussed in Section 3.1, we need to handle rank deficient $\mathrm{Cov}(x)$ in the first ℓ steps of system identification. The following lemma lower bounds the minimum positive singular value of the data matrix in the potentially rank deficient case.

Lemma 9 (Mimimum positive singular value)**.** *Let d-dimensional random vector $x \sim \mathcal{N}(0, \Sigma)$ and $(x^{(i)})_{i=1}^{n}$ be n independent observations of x. Let X be the matrix whose ith row is $(x^{(i)})^\top$. Then, as long as $n \ge 8d + 16\log(1/p)$, with probability at least $1 - p$, $\sigma_{\min}^+(X) \ge \sigma_{\min}^+(\Sigma^{1/2})n^{1/2}/2$.*

Proof. Let r be the rank of Σ and D be an $d \times r$ matrix so that $\Sigma = DD^\top$. We can parameterize x by Dg, where $g \sim \mathcal{N}(0, I_r)$ is an r-dimensional standard Gaussian random vector. Then sample $x^{(i)}$ can be expressed as $Dg^{(i)}$ where $g^{(i)}$ is an independent observation of g, and matrix X can be expressed as $GD^\top$, where G is a matrix of size $n \times r$, whose ith row is $(g^{(i)})^\top$. Hence, $X^\top X = DG^\top GD^\top$. It follows that $\mathrm{rank}(X^\top X) \le \mathrm{rank}(DD^\top)$. On the other hand, $X^\top X \succcurlyeq \sigma_{\min}^2(G)DD^\top$. By (Wainwright, 2019, Theorem 6.1), with probability at least $1 - p$,

$$
\sigma_{\min}(G) \ge n^{1/2} - d^{1/2} - (2\log(1/p))^{1/2}.
$$

Since $n \ge 8d + 16\log(1/p)$, we have $\sigma_{\min}(G) \ge n^{1/2}/2 > 0$. This implies that with the same probability, $\mathrm{rank}(X^\top X) \ge \mathrm{rank}(DD^\top)$. Combining it with the inequality in the other direction, we have $\mathrm{rank}(X^\top X) = \mathrm{rank}(DD^\top) = r$. Therefore, with the same probability,

$$
\sigma_{\min}^+(X) = (\sigma_r(X^\top X))^{1/2} \ge (\sigma_r(DD^\top)n)^{1/2}/2 = \sigma_{\min}^+(\Sigma^{1/2})n^{1/2}/2,
$$

which completes the proof. □

The following lemma is the main result in Section D.3.

Lemma 10 (Noisy rank deficient linear regression). *Define random vector $y^* := A^*x^* + e$, where $x^* \sim \mathcal{N}(0, \Sigma_*)$ and $e \sim \mathcal{N}(0, \Sigma_e)$ are d_1 and d_2 dimensional Gaussian random vectors, respectively. Define $x := x^* + \delta_x$ and $y := y^* + \delta_y$ where the Gaussian noise vectors δ_x, δ_y can be correlated with x^*, y^*, respectively, and satisfy $\|\mathrm{Cov}(\delta_x)^{1/2}\|_2 \leq \epsilon_x$, $\|\mathrm{Cov}(\delta_y)^{1/2}\|_2 \leq \epsilon_y$. Assume that $\|A^*\|_2$, $\|\Sigma_*\|_2$ and $\|\Sigma_e\|_2$ are $\mathcal{O}(1)$. Let $\sigma_{\min}^+(\Sigma_*^{1/2}) \geq \beta > 0$ and $\sigma_{\min}^+(\mathrm{Cov}(x)^{1/2}) \geq \theta > 0$. Assume that $\theta = \Omega(\epsilon_x)$, $\theta = \Omega(\epsilon_y)$ for absolute constants and θ has at least $n^{-1/4}$ dependence on n. Suppose we get n independent observations $x^{(i)}, y^{(i)}$ of x and y. Consider the minimum Frobenius norm solution*

$$\widehat{A} \in \underset{A}{\mathrm{argmin}} \sum_{i=1}^n (y^{(i)} - Ax^{(i)})^2.$$

Then, there exists an absolute constant $c > 0$, such that if $n \geq c(d_1 + d_2 + \log(1/p))$, with probability at least $1 - p$,

$$\|(\widehat{A} - A^*)\Sigma_*^{1/2}\|_2 = \mathcal{O}(\beta^{-1}\epsilon_x + \epsilon_y + n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2}).$$

Proof. Let $r = \mathrm{rank}(\Sigma_*)$ and $\Sigma_* = DD^\top$ where $D \in \mathbb{R}^{d_1 \times r}$. We can view x^* as generated from an r-dimensional standard Gaussian $g \sim \mathcal{N}(0, I_r)$, by $x^* = Dg$; $x^{(i)}$ can then be viewed as $Dg^{(i)}$, where $(g^{(i)})_{i=1}^n$ are independent observations of g. Let X denote the matrix whose ith row is $(x^{(i)})^\top$; $X^*, Y, G, E, \Delta_x, \Delta_y$ are defined similarly.

To solve the regression problem, we set its gradient to be zero and substitute $Y = A^*X^* + E + \Delta_y$ to obtain

$$\widehat{A}X^\top X = A^*(X^*)^\top X + E^\top X + \Delta_y^\top X.$$

Substituting X by $GD^\top + \Delta_x$ gives

$$\widehat{A}DG^\top GD^\top + \widehat{A}(DG^\top \Delta_x + \Delta_x^\top GD^\top + \Delta_x^\top \Delta_x) = A^*DG^\top GD^\top + A^*DG^\top \Delta_x$$
$$+ (E^\top + \Delta_y^\top)(GD^\top + \Delta_x).$$

By rearranging the terms, we have

$$(\widehat{A} - A^*)DG^\top GD^\top = A^*DG^\top \Delta_x - \widehat{A}(DG^\top \Delta_x + \Delta_x^\top GD^\top + \Delta_x^\top \Delta_x) + (E^\top + \Delta_y^\top)(GD^\top + \Delta_x).$$

Hence,

$$\|(\widehat{A} - A^*)D\|_2$$
$$= \left\|\left(A^*DG^\top \Delta_x - \widehat{A}(DG^\top \Delta_x + \Delta_x^\top GD^\top + \Delta_x^\top \Delta_x) + (E^\top + \Delta_y^\top)(GD^\top + \Delta_x)\right)(D^\top)^\dagger(G^\top G)^{-1}\right\|_2.$$

We claim that as long as $n \geq 16(d_1 + d_2 + \log(1/p))$, with probability at least $1 - 4p$, $\|\widehat{A}\|_2 = \mathcal{O}(\|A^*\|_2)$ (Claim 1). Then, since $\|D^\dagger\|_2 = (\sigma_{\min}^+(\Sigma_*))^{-1} \leq \beta^{-1}$,

$$\|(\widehat{A} - A^*)D\|_2 = \mathcal{O}(\beta^{-1}\|A^*\|_2)(\|D\|_2\|G\|_2\|\Delta_x\|_2 + \|\Delta_x\|_2^2)\|(G^\top G)^{-1}\|_2)$$
$$+ \mathcal{O}(1)(\|G^\dagger E\|_2 + \|G^\dagger \Delta_y\|_2 + \beta^{-1}(\|E\|_2 + \|\Delta_y\|_2)\|\Delta_x\|_2\|(G^\top G)^{-1}\|_2).$$

31

By (Wainwright, 2019, Theorem 6.1), the Gaussian ensemble Δ_x satisfies that with probability at least $1 - p$,

$$\|\Delta_x\|_2 \leq (n\|\mathrm{Cov}(\delta_x)\|_2)^{1/2} + (\mathrm{tr}(\mathrm{Cov}(\delta_x)))^{1/2} + (2\|\mathrm{Cov}(\delta_x)\|_2 \log(1/p))^{1/2}$$
$$\leq \|\mathrm{Cov}(\delta_x)\|_2^{1/2}(n^{1/2} + d_1^{1/2} + (2\log(1/p))^{1/2}).$$

Since $n \geq 8d_1 + 16\log(1/p)$, we have $\|\Delta_x\|_2 \leq 2\epsilon_x n^{1/2}$. Similarly,

$$\|\Delta_y\|_2 = \mathcal{O}(\epsilon_y n^{1/2}), \quad \|E\|_2 = \mathcal{O}((\|\Sigma_e\|_2 n)^{1/2}), \quad \|G\|_2 = \mathcal{O}(n^{1/2}), \quad \sigma_{\min}(G) = \Omega(n^{1/2}).$$

It follows that $\|(G^\top G)^{-1}\|_2 = \mathcal{O}(n^{-1})$ and $\|G^\dagger\|_2 = \mathcal{O}(n^{-1/2})$. Hence,

$$\|(\widehat{A} - A^*)D\|_2 = \mathcal{O}(\beta^{-1}\|A^*\|_2)(\|D\|_2\epsilon_x + \epsilon_x^2)$$
$$+ \mathcal{O}(1)(\|G^\dagger E\|_2 + \epsilon_y + \beta^{-1}(\|\Sigma_e^{1/2}\|_2 + \epsilon_y)\epsilon_x)$$
$$= \mathcal{O}(\beta^{-1}\|A^*\|_2\epsilon_x + \epsilon_y + \|G^\dagger E\|_2),$$

where we consider ϵ_x and ϵ_y as quantities much smaller than one such that terms like $\epsilon_x^2, \epsilon_x\epsilon_y$ are absorbed into ϵ_x, ϵ_y. It remains to control $\|G^\dagger E\|_2$. (Mhammedi et al., 2020, Section B.2.11) proves via a covering number argument that with probability at least $1 - p$,

$$\|G^\dagger E\|_2 = \mathcal{O}(1)\sigma_{\min}(G)^{-1}\|\Sigma_e^{1/2}\|_2(d_1 + d_2 + \log(1/p))^{1/2} = \mathcal{O}(n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2}).$$

Overall, we obtain that

$$\|(\widehat{A} - A^*)\Sigma_*^{1/2}\|_2 = \|(\widehat{A} - A^*)D\|_2 = \mathcal{O}(\beta^{-1}\|A^*\|_2\epsilon_x + \epsilon_y + n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2}),$$

which completes the proof. $\qquad\qquad\square$

Claim 1. *Under the conditions in Lemma 10, as long as $n \geq 16(d_1 + d_2 + \log(1/p))$, with probability at least $1 - 4p$, $\|\widehat{A}\|_2 = \mathcal{O}(\|A^*\|_2)$.*

Proof. A minimum norm solution is given by the following closed-form solution of $\widehat{A}$ using pseudoinverse (Moore-Penrose inverse):

$$\widehat{A} = (A^*(X^*)^\top + E^\top + \Delta_y^\top)X(X^\top X)^\dagger$$
$$= A^*(X^\dagger X^*)^\top + (X^\dagger E)^\top + (X^\dagger \Delta_y)^\top$$
$$= A^*(X^\dagger X)^\top - A^*(X^\dagger \Delta_x)^\top + (X^\dagger E)^\top + (X^\dagger \Delta_y)^\top.$$

Then

$$\|\widehat{A}\|_2 \leq \|A^*\|_2 + (\|A^*\|_2\|\Delta_x\|_2 + \|\Delta_y\|_2)(\sigma_{\min}^+(X))^{-1} + \|X^\dagger E\|_2,$$

where we note that $\|X^\dagger\|_2 = \sigma_{\min}^+(X)^{-1}$ when $X \neq 0$. Let $\theta = \sigma_{\min}^+(\mathrm{Cov}(x)^{1/2})$. By Lemma 9, as long as $n \geq 16(d_1 + d_2 + \log(1/p))$, with probability at least $1 - p$, $(\sigma_{\min}^+(X))^{-1} \leq 2\theta^{-1}n^{-1/2}$. We have shown in the proof of Lemma 10 that with probability at least $1 - 2p$,

$$\|\Delta_x\|_2 \leq 2\epsilon_x n^{1/2}, \quad \|\Delta_y\|_2 \leq 2\epsilon_y n^{1/2}.$$

Similar to the proof of Lemma 10, by (Mhammedi et al., 2020, Section B.2.11), with probability at least $1 - p$,

$$\|X^\dagger E\|_2 = \mathcal{O}(1)\sigma_{\min}^+(X)^{-1}\|\Sigma_e^{1/2}\|_2(d_1 + d_2 + \log(1/p))^{1/2} = \mathcal{O}(\theta^{-1}n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2}).$$

Combining the bounds above, we obtain

$$\|\widehat{A}\|_2 \leq \|A^*\|_2 + (\|A^*\|_2 2\epsilon_x n^{1/2} + 2\epsilon_y n^{1/2})2\theta^{-1}n^{-1/2} + \mathcal{O}(\theta^{-1}n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2})$$
$$= \|A^*\|_2(1 + 4\theta^{-1}\epsilon_x) + 4\theta^{-1}\epsilon_y + \mathcal{O}(\theta^{-1}n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2}).$$

Hence, as long as $\theta = \Omega(\epsilon_x)$, $\theta = \Omega(\epsilon_y)$ for absolute constants and θ has at least $n^{-1/4}$ dependence on n, $\|\widehat{A}\|_2$ is bounded by $c\|A^*\|_2$ for an absolute constant $c > 0$. □

In Lemma 10, if Σ_* has full rank and $\sigma_{\min}(\Sigma_*) \geq \beta > 0$ and $\epsilon_x, \epsilon_y \leq \beta/2$, then $\sigma_{\min}(\text{Cov}(x)) = \Omega(\beta)$, and

$$\|\widehat{A} - A^*\|_2 = \mathcal{O}(\beta^{-1}(\beta^{-1}\epsilon_x + \epsilon_y + n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2})).$$

The following lemma shows that we can strengthen the result by removing the β^{-1} factor before ϵ_x.

Lemma 11 (Noisy linear regression). *Define random variable $y^* = A^*x^* + e$, where $x^* \sim \mathcal{N}(0, \Sigma_*)$ and $e \sim \mathcal{N}(0, \Sigma_e)$ are d_1 and d_2 dimensional random vectors. Assume that $\|A^*\|_2$, $\|\Sigma_*\|_2$ and $\|\Sigma_e\|_2$ are $\mathcal{O}(1)$, and $\sigma_{\min}(\Sigma_*^{1/2}) \geq \beta > 0$. Define $x := x^* + \delta_x$ and $y := y^* + \delta_y$ where the Gaussian noise vectors δ_x, δ_y can be correlated with x^*, y^*, respectively, and satisfy $\|\text{Cov}(\delta_x)^{1/2}\|_2 \leq \epsilon_x$, $\|\text{Cov}(\delta_y)^{1/2}\|_2 \leq \epsilon_y$, and $\epsilon_x, \epsilon_y \leq \beta/2$. Suppose we get n independent observations $x^{(i)}, y^{(i)}$ of x and y. Consider the minimum Frobenius norm solution*

$$\widehat{A} \in \underset{A}{\arg\min} \sum_{i=1}^n (y^{(i)} - Ax^{(i)})^2.$$

Then, there exists an absolute constant $c > 0$, such that if $n \geq c(d_1 + d_2 + \log(1/p))$, with probability at least $1 - p$,

$$\|\widehat{A} - A^*\|_2 = \mathcal{O}(\beta^{-1}(\epsilon_x + \epsilon_y + n^{-1/2}(d_1 + d_2 + \log(1/p))^{1/2})).$$

Proof. Following the proof of Claim 1, we have

$$\|\widehat{A} - A^*\|_2 \leq \|A^*\|\|X^\dagger\Delta_x\|_2 + \|X^\dagger E\|_2 + \|X^\dagger\Delta_y\|_2.$$

Combining with the bounds on $\|X^\dagger\Delta_x\|_2$, $\|X^\dagger\Delta_y\|_2$ and $\|X^\dagger E\|_2$ concludes the proof. □

D.4 Certainty equivalent linear quadratic control

As shown in Lemma 10, if the input of linear regression does not have full-rank covariance, then the parameters can only be identified in certain directions. The following lemma studies the performance of the certainty equivalent optimal controller in this case.

Lemma 12 (Rank deficient LQ control). *Consider the finite-horizon time-varying linear dynamical system $x_{t+1} = A_t^* x_t + B_t^* u_t + w_t$ with unknown $(A_t^*, B_t^*)_{t=0}^{T-1}$.*

- *Assume that $(A_t^*)_{t=0}^{T-1}$ is uniformly exponential stable (Assumption 1).*
- *Assume that $x_0 \sim \mathcal{N}(0, \Sigma_0)$, $w_t \sim \mathcal{N}(0, \Sigma_{w_t})$ for all $0 \leq t \leq T - 1$, are independent, where Σ_0 and $(\Sigma_{w_t})_{t=0}^{T-1}$ do not necessarily have full rank.*
- *Instead of x_t, we observe $x_t' = x_t + \delta_{x_t}$, where the Gaussian noise vector δ_{x_t} can be correlated with x_t and satisfy $\|\mathrm{Cov}(\delta_{x_t})^{1/2}\| \leq \varepsilon$ for all $0 \leq t \leq T$.*
- *Let $\Sigma_t := \mathrm{Cov}(x_t)$ and $\Sigma_t' := \mathrm{Cov}(x_t')$ under control $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for all $0 \leq t \leq T - 1$. Assume $\sigma_{\min}^+((\Sigma_t)^{1/2}) \geq \beta > 0$ and $\sigma_{\min}^+((\Sigma_t')^{1/2}) \geq \theta > 0$ for all $0 \leq t \leq T$, and that $\theta = \Omega(\varepsilon)$ with at least $n^{-1/4}$ dependence on n.*
- *Assume $(Q_t^*)_{t=0}^T, (\widehat{Q}_t)_{t=0}^T$ are positive definite with $\mathcal{O}(1)$ operator norms, and $\|\widehat{Q}_t - Q_t^*\|_2 \leq \varepsilon$ for all $0 \leq t \leq T$.*

Collect n trajectories using $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for some $\sigma_u > 0$ and $0 \leq t \leq T - 1$. Identify $(\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1}$ by SysID (Algorithm 3), which uses ordinary linear regression and takes a minimum Frobenius norm solution. Let $(\widehat{K}_t)_{t=0}^{T-1}$ be the optimal controller in system $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$. Then, there exists an absolute constant $a > 0$, such that if $n \geq a(d_x + d_u + \log(1/p))$, with probability at least $1 - p$, $(\widehat{K}_t)_{t=0}^{T-1}$ is ϵ-optimal in system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^T)$ for

$$\epsilon = \mathcal{O}(c^T((1 + \beta^{-1})d_x \varepsilon + d_x(d_x + d_u + \log(1/p))^{1/2} n^{-1/2})),$$

where dimension-free constant $c > 0$ depends on the system parameters.

Proof. First of all, we establish that $(K_t^*)_{t=0}^{T-1}, (\widehat{K}_t)_{t=0}^{T-1}$ are bounded. Since $(Q_t^*)_{t=0}^T, (\widehat{Q}_t)_{t=0}^T$ are positive definite, the system is fully cost observable. By (Zhang and Zhang, 2021), the optimal feedback gains $(K_t^*)_{t=0}^{T-1}$ have operator norms bounded by dimension-free constants depending on the system parameters. Note that (Zhang and Zhang, 2021) studies the boundedness of the Kalman filter, which is dual to our optimal control problem, but their results carry over. One can also see the boundedness from the compact formulation (Zhang et al., 2021), as described in the proof of Lemma 13, using known bounds for RDE solutions (Komaroff, 1996). The same argument applies to $(\widehat{K}_t)_{t=0}^{T-1}$ if we can show $(\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1}$ have $\mathcal{O}(1)$ operator norms. SysID (Algorithm 3) identifies $[\widehat{A}_t, \widehat{B}_t]$ by ordinary linear regression for all $0 \leq t \leq T - 1$. By Claim 1, with a union bound for all $0 \leq t \leq T - 1$, as long as $n \geq 16(d_x + d_u + \log(T/p))$, with probability at least $1 - 4p$, $\|[\widehat{A}_t, \widehat{B}_t]\|_2 = \mathcal{O}(\|[A_t^*, B_t^*]\|_2) = \mathcal{O}(1)$. Hence, with the same probability, $(\widehat{K}_t)_{t=0}^{T-1}$ have operator norms bounded by dimension-free constants depending on the system parameters. In the following, we assume $\|K_t^*\|_2, \|\widehat{K}_t\|_2 \leq \kappa$ for all $0 \leq t \leq T - 1$, where κ is a dimension-free constant that depends on system parameters.

Since Σ_t can be rank deficient, we cannot guarantee $\|\widehat{A}_t - A_t^*\|_2$ is small. Instead, for all $0 \leq t \leq T - 1$, by Lemma 10,

$$([\widehat{A}_t, \widehat{B}_t] - [A_t^*, B_t^*])\mathrm{diag}(\Sigma_t^{1/2}, \sigma_u I) = \mathcal{O}(\delta),$$

where $\delta := (1 + \beta^{-1})\varepsilon + (d_x + d_u + \log(1/p))^{1/2}n^{-1/2}$, and $\epsilon_x, \epsilon_y, d_1, d_2$ in Lemma [10] correspond to $\varepsilon, \varepsilon, d_x + d_u, d_x$ here, respectively. This implies that $\|(\widehat{A}_t - A_t^*)(\Sigma_t)^{1/2}\|_2 = \mathcal{O}(\delta)$ and $\|\widehat{B}_t - B_t^*\|_2 = \mathcal{O}(\delta)$ for all $0 \le t \le T - 1$.

Let Ξ_t denote the covariance of x_t under state feedback controller $(K_t)_{t=0}^{T-1}$, where $\|K_t\| \le \kappa$ for all $0 \le t \le T - 1$. We claim that there exists $b_t > 0$, such that $\Xi_t \preccurlyeq b_t \Sigma_t$. We prove it by induction. At step 0, clearly, $\Xi_0 = \Sigma_0$. Suppose $\Xi_t \preccurlyeq b_t \Sigma_t$. For step $t + 1$,

$$\Sigma_{t+1} = A_t^* \Sigma_t (A_t^*)^\top + \sigma_u^2 B_t^* (B_t^*)^\top + \Sigma_{w_t},$$

$$\Xi_{t+1} = (A_t^* + B_t^* K_t)\Xi_t(A_t^* + B_t^* K_t)^\top + \Sigma_{w_t} \preccurlyeq 2A_t^* \Xi_t (A_t^*)^\top + 2B_t^* K_t \Xi_t K_t^\top (B_t^*)^\top + \Sigma_{w_t}.$$

Hence, for $b \ge 1$,

$$b\Sigma_{t+1} - \Xi_{t+1} \succcurlyeq (b - 2b_t)A_t^* \Sigma_t (A_t^*)^\top + b\sigma_u^2 B_t^* (B_t^*)^\top - 2B_t^* K_t \Xi_t K_t^\top (B_t^*)^\top.$$

To ensure $b_{t+1}\Sigma_{t+1} \succcurlyeq \Xi_{t+1}$, it suffices to take $b_{t+1} = \max\{2, 2\sigma_u^{-2}\kappa^2\|\Sigma_t\|_2\}b_t$. Hence, $b_t \le a_0^t$, where constant $a_0 > 0$ is dimension-free and depends on system parameters. By the definition of the operator norm, we have $\|(\widehat{A}_t - A_t^*)\Xi_t^{1/2}\|_2 \le \|(\widehat{A}_t - A_t^*)(b_t\Sigma_t)^{1/2}\|_2 = \mathcal{O}(b_t^{1/2}\delta)$ for all $0 \le t \le T - 1$.

Let $\widehat{\Xi}_t$ denote the covariance of x_t under state feedback controller $(K_t)_{t=0}^{T-1}$ in the system $(\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1}$. Then, by definition,

$$\begin{aligned}
&\left\|\Xi_{t+1} - \widehat{\Xi}_{t+1}\right\|_2 \\
&= \left\|(A_t^* + B_t^* K_t)\Xi_t(A_t^* + B_t^* K_t)^\top + \Sigma_{w_t} - \left((\widehat{A}_t + \widehat{B}_t K_t)\widehat{\Xi}_t(\widehat{A}_t + \widehat{B}_t K_t)^\top + \Sigma_{w_t}\right)\right\|_2 \\
&= \left\|(A_t^* + B_t^* K_t)\Xi_t(A_t^* + B_t^* K_t)^\top - (\widehat{A}_t + \widehat{B}_t K_t)\Xi_t(\widehat{A}_t + \widehat{B}_t K_t)^\top \right. \\
&\qquad \left. + (\widehat{A}_t + \widehat{B}_t K_t)(\Xi_t - \widehat{\Xi}_t)(\widehat{A}_t + \widehat{B}_t K_t)^\top\right\|_2 \\
&= \mathcal{O}\left(\kappa b_t^{1/2}\left\|(A_t^* + B_t^* K_t)\Xi_t^{1/2} - (\widehat{A}_t + \widehat{B}_t K_t)\Xi_t^{1/2}\right\|_2\right) + \mathcal{O}(1 + \kappa^2)\|\widehat{\Xi}_t - \Xi_t\|_2,
\end{aligned}$$

where the operator norms of A_t^*, B_t^* are $\mathcal{O}(1)$ and the operator norms of $\widehat{A}_t, \widehat{B}_t$ are $\mathcal{O}(1)$ with probability at least $1 - 4p$. Since

$$\left\|(A_t^* + B_t^* K_t)\Xi_t^{1/2} - (\widehat{A}_t + \widehat{B}_t K_t)\Xi_t^{1/2}\right\|_2 = \left\|(A_t^* - \widehat{A}_t)\Xi_t^{1/2} + (B_t^* - \widehat{B}_t)K_t\Xi_t^{1/2}\right\|_2 = \mathcal{O}(\delta\kappa b_t^{1/2}),$$

we have $\left\|\widehat{\Xi}_{t+1} - \Xi_{t+1}\right\|_2 = \mathcal{O}(\delta\kappa^2 b_t + (1 + \kappa^2)\|\widehat{\Xi}_t - \Xi_t\|_2)$. Combining with $\widehat{\Xi}_0 = \Xi_0 = \Sigma_0$ gives

$$\|\widehat{\Xi}_t - \Xi_t\|_2 = \mathcal{O}(b_t(a_1 + a_1\kappa^2)^t\delta) = \mathcal{O}((a_2 + a_2\kappa^2)^t\delta)$$

for some dimension-free constants $a_1, a_2 > 0$ that depend on the system parameters.

Let $J((K_t)_{t=0}^{T-1})$ denote the expected cumulative cost under state feedback controller $(K_t)_{t=0}^{T-1}$ in system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^T)$ and $\widehat{J}((K_t)_{t=0}^{T-1})$ the corresponding expected cumulative cost in system $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$. Notice that

$$\begin{aligned}
J((K_t)_{t=0}^{T-1}) = \mathbb{E}\left[\sum_{t=0}^T c_t\right] &= \mathbb{E}\left[\sum_{t=0}^T x_t^\top (Q_t^*)' x_t\right] \\
&= \mathbb{E}\left[\sum_{t=0}^T \left\langle (Q_t^*)', x_t x_t^\top \right\rangle_F\right] \\
&= \sum_{t=0}^T \left\langle (Q_t^*)', \Xi_t \right\rangle_F,
\end{aligned}$$

where $(Q_t^*)' := (Q_t^* + K_t^\top R_t^* K_t)$ for $0 \le t \le T-1$ and $(Q_T^*)' := Q_T^*$. Similarly, $\widehat{J}((K_t)_{t=0}^{T-1}) = \sum_{t=0}^{T} \left\langle \widehat{Q}_t', \widehat{\Xi}_t \right\rangle_F$, where $\widehat{Q}_t' = \widehat{Q}_t + K_t^\top R_t^* K_t$ for $0 \le t \le T-1$ and $\widehat{Q}_T' = \widehat{Q}_T$. Hence,

$$
\begin{aligned}
\left| J((K_t)_{t=0}^{T-1}) - \widehat{J}((K_t)_{t=0}^{T-1}) \right| &= \sum_{t=0}^{T} \left\langle (Q_t^*)' - \widehat{Q}_t', \Xi_t \right\rangle_F + \sum_{t=0}^{T} \left\langle \widehat{Q}_t', \Xi_t - \widehat{\Xi}_t \right\rangle_F \\
&= \varepsilon d_x \sum_{t=0}^{T} \mathcal{O}(b_t) + \kappa^2 d_x \sum_{t=0}^{T} \mathcal{O}((a_2 + a_2 \kappa^2)^t \delta) \\
&= \varepsilon d_x \sum_{t=0}^{T} \mathcal{O}(a_0^t) + \delta \kappa^2 d_x \sum_{t=0}^{T} \mathcal{O}((a_2 + a_2 \kappa^2)^t) \\
&= \mathcal{O}(\delta d_x c^T),
\end{aligned}
$$

where $\|\widehat{Q}_t'\|_2 = \mathcal{O}(\kappa^2)$, ε is absorbed into $\delta = (1 + \beta^{-1})\varepsilon + (d_x + d_u + \log(1/p))^{1/2} n^{-1/2}$, and dimension-free constant $c > 0$ depends on the system parameters.

Finally, let $(K_t^*)_{t=0}^{T-1}$ be the optimal feedback gains in system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^{T})$. By the union bound, with probability at least $1 - 8p$,

$$
|J((\widehat{K}_t)_{t=0}^{T-1}) - \widehat{J}((\widehat{K}_t)_{t=0}^{T-1})| = \mathcal{O}(\delta d_x c^T), \quad |J((K_t^*)_{t=0}^{T-1}) - \widehat{J}((K_t^*)_{t=0}^{T-1})| = \mathcal{O}(\delta d_x c^T).
$$

Therefore,

$$
\begin{aligned}
&J((\widehat{K}_t)_{t=0}^{T-1}) - J((K_t^*)_{t=0}^{T-1}) \\
&= J((\widehat{K}_t)_{t=0}^{T-1}) - \widehat{J}((\widehat{K}_t)_{t=0}^{T-1}) + \widehat{J}((\widehat{K}_t)_{t=0}^{T-1}) - \widehat{J}((K_t^*)_{t=0}^{T-1}) + \widehat{J}((K_t^*)_{t=0}^{T-1}) - J((K_t^*)_{t=0}^{T-1}) \\
&= \mathcal{O}(d_x c^T \delta) = \mathcal{O}(c^T((1 + \beta^{-1})d_x \varepsilon + d_x(d_x + d_u + \log(1/p))^{1/2} n^{-1/2})).
\end{aligned}
$$

The proof is completed by rescaling $8p$ to p. $\qquad\square$

If the input of linear regression has full-rank covariance, then by Lemma 11, the system parameters can be fully identified. The certainty equivalent optimal controller in this case has a much better guarantee compared to the rank-deficient case, as shown in the following lemma.

Lemma 13 (LQ control). *Assume the finite-horizon linear time-varying system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^{T})$ is stabilizable. Let $(K_t^*)_{t=0}^{T-1}$ be the optimal controller and $(P_t^*)_{t=0}^{T}$ be the solution to RDE (2.3) of system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^{T})$. Let $(\widehat{K}_t)_{t=0}^{T-1}$ be the optimal controller and $(\widehat{P}_t)_{t=0}^{T}$ be the solution to RDE (2.3) of system $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^{T})$, where $\|\widehat{A}_t - A_t^*\|_2 \le \epsilon$, $\|\widehat{B}_t - B_t^*\|_2 \le \epsilon$ and $\|\widehat{Q}_t - Q_t^*\|_2 \le \epsilon$. Then there exists dimension-free constant $\epsilon_0 > 0$ with ϵ_0^{-1} depending polynomially on system parameters, such that as long as $\epsilon \le \epsilon_0$, $\|\widehat{P}_t - P_t^*\|_2 = \mathcal{O}(\epsilon)$, $\|\widehat{K}_t - K_t^*\|_2 = \mathcal{O}(\epsilon)$ for all $t \ge 0$, and $(\widehat{K}_t)_{t=0}^{T-1}$ is $\mathcal{O}((d_x \wedge d_u)T\epsilon^2)$-optimal in system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^{T})$.*

Proof. Mania et al. (2019) have studied this problem in the infinite-horizon LTI setting; here we extend their result to the finite-horizon LTV setting.

We adopt the following compact formulation of a finite-horizon LTV system, introduced in (Zhang et al., 2021, Section 3), to reduce our setting to the infinite-horizon LTI setting; since noisy and noiseless LQR has the same associated Riccati equation and optimal controller, we consider the

36

noiseless case here:

$$x = [x_0; \ldots; x_T], \quad u = [u_0; \ldots; u_{T-1}], \quad w = [x_0; w_0; \ldots; w_{T-1}]$$

$$A^* = \begin{bmatrix} 0_{d_x \times d_x T} & 0_{d_x \times d_x} \\ \text{diag}(A_0^*, \ldots, A_{T-1}^*) & 0_{d_x T \times d_x} \end{bmatrix}, \quad B^* = \begin{bmatrix} 0_{d_x \times d_u T} \\ \text{diag}(B_0^*, \ldots, B_{T-1}^*) \end{bmatrix},$$

$$Q^* = \text{diag}(Q_0^*, \ldots, Q_T^*), \quad R^* = \text{diag}(R_0^*, \ldots, R_{T-1}^*) \quad K = [\text{diag}(K_0, \ldots, K_{T-1}), 0_{d_u T \times d_x}].$$

The control inputs using state feedback controller $(K_t)_{t=0}^{T-1}$ can be characterized by $u = Kx$. Let $(P_t^K)_{t=0}^T$ be the associated cumulative cost matrix starting from step t. Then

$$P_t^K = (A_t^* + B_t^* K_t)^\top P_{t+1}^K (A_t^* + B_t^* K_t) + Q_t^* + K_t^\top R_t^* K_t,$$

and $P^K := \text{diag}(P_0^K, \ldots, P_T^K)$ is the solution to

$$P^K = (A^* + B^* K)^\top P^K (A^* + B^* K) + Q^* + K^\top R^* K.$$

Similarly, the optimal cumulative cost matrix $P^* := \text{diag}(P_0^*, \ldots, P_T^*)$ produced by the RDE (2.3) in system (A^*, B^*, R^*, Q^*) (that is, system $((A_t^*, B_t^*, R_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^T)$) satisfies

$$P^* = (A^*)^\top (P^* - P^* B^* ((B^*)^\top P^* B^* + R^*)^{-1} (B^*)^\top P^*) A^* + Q^*.$$

Let $\widehat{P} = \text{diag}(\widehat{P}_0, \ldots, \widehat{P}_T)$ be the optimal cumulative cost matrices in system $(\widehat{A}, \widehat{B}, \widehat{Q}, R^*)$ (that is, system $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$) by the RDE (2.3). Define $K^* := [\text{diag}(K_0^*, \ldots, K_T^*), 0_{d_u T \times d_x}]$, where $(K_t^*)_{t=0}^{T-1}$ is the optimal controller in system (A^*, B^*, Q^*, R^*), and define $\widehat{K}$ similarly for system $(\widehat{A}, \widehat{B}, \widehat{Q}, R^*)$. By definition, (A^*, B^*, Q^*) is stabilizable and observable in the sense of LTI systems. Therefore, by (Mania et al., 2019, Propositions 1 and 2), there exists dimension-free constant $\epsilon_0 > 0$ with ϵ_0^{-1} depending polynomially on system parameters such that as long as $\epsilon \leq \epsilon_0$,

$$\|\widehat{P} - P^*\|_2 = \mathcal{O}(\epsilon), \quad \|\widehat{K} - K^*\|_2 = \mathcal{O}(\epsilon),$$

and that $\widehat{K}$ stabilizes system (A^*, B^*). By (Fazel et al., 2018, Lemma 12),

$$J((\widehat{K}_t)_{t=0}^{T-1}) - J^* = \sum_{t=0}^T \text{tr}(\Sigma_t (\widehat{K}_t - K_t^*)^\top (R_t^* + (B_t^*)^\top P_{t+1}^* B_t^*)(\widehat{K}_t - K_t^*))$$
$$= \text{tr}(\Sigma (\widehat{K} - K^*)^\top (R^* + (B^*)^\top P^* B^*)(\widehat{K} - K^*)),$$

where $\Sigma = \text{diag}(\Sigma_0, \ldots, \Sigma_T)$ and Σ_t is $\mathbb{E}[x_t x_t^\top]$ in system (A^*, B^*) under state feedback controller $\widehat{K}$. As a result,

$$J((\widehat{K}_t)_{t=0}^{T-1}) - J^* \leq \|\Sigma\|_2 \|R^* + (B^*)^\top P^* B^*\|_2 \|\widehat{K} - K^*\|_F^2.$$

Since $\widehat{K}$ stabilizes system (A^*, B^*), Σ has bounded norm. Since

$$\|\widehat{K} - K\|_F \leq ((d_x \wedge d_u) T)^{1/2} \|\widehat{K} - K\|_2,$$

we conclude that

$$J((\widehat{K}_t)_{t=0}^{T-1}) - J^* = \mathcal{O}((d_x \wedge d_u) T \epsilon^2).$$

$\square$

E   Proof of Theorem 2

The full SysId algorithm is shown in Algorithm 3.

Algorithm 3 SysId: system identification

1: **Input:** data in the form of $(z_0^{(i)}, u_0^{(i)}, c_0^{(i)}, \dots, z_{T-1}^{(i)}, u_{T-1}^{(i)}, c_{T-1}^{(i)}, z_T^{(i)}, c_T^{(i)})_{i=1}^n$
2: Estimate the system dynamics by   ▷ pick min. Frobenius norm solution by pseudoinverse

$$(\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1} \in \underset{(A_t, B_t)_{t=0}^{T-1}}{\mathrm{argmin}} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \|A_t z_t^{(i)} + B_t u_t^{(i)} - z_{t+1}^{(i)}\|^2 \tag{E.1}$$

3: For all $0 \leq t \leq \ell - 1$ and $t = T$, set $\widehat{Q}_t = I_{d_x}$
4: For all $\ell \leq t \leq T - 1$, obtain $\widetilde{Q}_t$ by

$$\widetilde{Q}_t, \widehat{b}_t \in \underset{Q_t = Q_t^\top, b_t}{\mathrm{argmin}} \sum_{i=1}^{n} (\|z_t^{(i)}\|_{Q_t}^2 + \|u_t^{(i)}\|_{R_t^*}^2 + b_t - c_t^{(i)})^2, \tag{E.2}$$

and set $\widehat{Q}_t = U \max(\Lambda, 0) U^\top$, where $\widetilde{Q}_t = U \Lambda U^\top$ is its eigenvalue decomposition
5: **Return:** system parameters $((\widehat{A}_t, \widehat{B}_t)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$

We state the full version of Theorem 2 below.

Theorem 3. *Given an unknown LQG system* (2.1), *let* $(M_t^{*\prime})_{t=0}^T$ *and* $((A_t^{*\prime}, B_t^{*\prime})_{t=0}^{T-1}, (Q_t^{*\prime})_{t=0}^T)$ *be the optimal state representation function and the true system parameters under the normalized parameterization. Suppose Assumptions 1, 2, 3, 4 and 5 hold and we run Algorithm 1 with* $n \geq \mathrm{poly}(T, d_x, d_y, d_u, \log(1/p))$, *where hidden constants are dimension-free and depend polynomially on system parameters. Then, with probability at least* $1 - p$, *the following guarantees on the state representation function* $(\widehat{M}_t)_{t=0}^T$ *and controller* $(\widehat{K}_t)_{t=0}^{T-1}$ *hold.*

There exists orthonormal matrices $(S_t)_{t=0}^T$, *such that for* $0 \leq t \leq \ell - 1$,

$$\|\widehat{M}_t - S_t M_t^{*\prime}\|_2 = \mathcal{O}(t(d_y + d_u)^{1/2} d_x^{3/4} n^{-1/4}),$$

and controller $(\widehat{K}_t)_{t=0}^{\ell-1}$ *is* ϵ-*optimal in LQ system* $((S_{t+1} A_t^{*\prime} S_t^{-1}, S_{t+1} B_t^{*\prime}, R_t^*)_{t=0}^{\ell-1}, (S_t Q_t^{*\prime} S_t^\top)_{t=0}^{\ell-1})$ *with terminal cost matrix* $P_t^{*\prime}$ *being produced by RDE* (2.3) *using the normalized parameterization, for*

$$\epsilon = \mathcal{O}((1 + \beta^{-1}) d_x^{7/4} (d_y + d_u)^{1/2} \ell c^\ell n^{-1/4});$$

for $\ell \leq t \leq T$,

$$\|\widehat{M}_t - S_t M_t^{*\prime}\|_2 = \mathcal{O}(\nu^{-1} m t^{3/2} (d_y + d_u) d_x^{1/2} n^{-1/2}),$$

and controller $(\widehat{K}_t)_{t=\ell}^{T-1}$ *is* ϵ-*optimal in LQ system* $((S_{t+1} A_t^{*\prime} S_t^{-1}, S_{t+1} B_t^{*\prime}, R_t^*)_{t=\ell}^{T-1}, (S_t Q_t^{*\prime} S_t^\top)_{t=\ell}^T)$, *for*

$$\epsilon = \mathcal{O}((1 + \nu^{-6}) d_x^4 (d_y + d_u + \log(1/p))^2 m^2 T^4 n^{-1}).$$

Proof. Algorithm 1 has three main steps: state representation learning (CoReL, Algorithm 2), latent system identification (SysId, Algorithm 3), and planning by RDE (2.3). Correspondingly, the analysis below is organized around these three steps.

38

Recovery of the state representation function. By Proposition 3, with $u_t = \mathcal{N}(0, \sigma_u^2 I)$ for all $0 \le t \le T-1$ to system (2.1), the k-step cumulative cost starting from step t, where $k = 1$ for $0 \le t \le \ell-1$ and $k = m \wedge (T-t+1)$ for $\ell \le t \le T$, is given under the normalized parameterization by

$$\bar{c}_t := c_t + c_{t+1} + \ldots + c_{t+k-1} = \|z_t^{*\prime}\|^2 + \sum_{\tau=t}^{t+k-1} \|u_\tau\|_{R_\tau^*}^2 + b_t' + e_t',$$

where $b_t' = \mathcal{O}(k)$, and e_t' is a zero-mean subexponential random variable with $\|e_t'\|_{\psi_1} = \mathcal{O}(k d_x^{1/2})$. Then, it is clear that Algorithm 2 recovers latent states $z_t^{*\prime} = M_t^{*\prime} h_t$, where $0 \le t \le T$, by a combination of quadratic regression and low-rank approximate factorization. Below we drop the superscript prime for notational simplicity, but keep in mind that the optimal state representation function $(M_t^*)_{t=0}^T$, the corresponding latent states $(z_t^*)_{t=0}^T$, and the true latent system parameters $((A_t^*, B_t^*)_{t=0}^{T-1}, (Q_t^*)_{t=0}^T)$ are all with respect to the normalized parameterization.

Let $N_t^* := (M_t^*)^\top M_t^*$ for all $0 \le t \le T$. For quadratic regression, Lemma 6 and the union bound over all time steps guarantee that as long as $n \ge cT^4(d_y + d_u)^4 \log(cT^3(d_y + d_u)^2/p) \log(T/p)$ for an absolute constant $c > 0$, with probability at least $1 - p$, for all $0 \le t \le T$,

$$\|\widehat{N}_t - N_t^*\|_F = \mathcal{O}(\sigma_{\min}(\text{Cov}(h_t^*))^{-2} t(d_y + d_u) n^{-1/2}) = \mathcal{O}(k t(d_y + d_u) d_x^{1/2} n^{-1/2}).$$

Now let us bound the distance between $\widehat{M}_t$ and M_t^*. Recall that we use $d_h = (t+1)d_y + t d_u$ as a shorthand. The estimate $\widehat{N}_t$ may not be positive semidefinite. Let $\widehat{N}_t = U\Lambda U^\top$ be its eigenvalue decomposition, with the $d_h \times d_h$ matrix Λ having descending diagonal elements. Let $\Sigma := \max(\Lambda, 0)$. Then $\widetilde{N}_t =: U\Sigma U^\top$ is the projection of $\widehat{N}_t$ onto the positive semidefinite cone (Boyd et al., 2004, Section 8.1.1) with respect to the Frobenius norm. Since $N_t^* \succcurlyeq 0$,

$$\|\widetilde{N}_t - N_t^*\|_F \le \|\widehat{N}_t - N_t^*\|_F.$$

The low-rank factorization is essentially a combination of low-rank approximation and matrix factorization. For $d_h < d_x$, $\widetilde{M}_t =: [\Sigma^{1/2}U^\top; 0_{(d_x-d_h) \times d_h}]$ constructed by padding zeros satisfies $\widetilde{M}_t^\top \widetilde{M}_t = \widetilde{N}_t$. For $d_h \ge d_x$, construct $\widetilde{M}_t =: \Sigma_{d_x}^{1/2} U_{d_x}^\top$, where Σ_{d_x} is the left-top $d_x \times d_x$ block in Σ and U_{d_x} consists of d_x columns of U from the left. By the Eckart-Young-Mirsky theorem, $\widetilde{M}_t^\top \widetilde{M}_t = U_{d_x} \Sigma_{d_x} U_{d_x}^\top$ satisfies

$$\|\widetilde{M}_t^\top \widetilde{M}_t - \widetilde{N}_t\|_F = \min_{N \in \mathbb{R}^{d_h \times d_h}, \text{rank}(N) \le d_x} \|N - \widetilde{N}_t\|_F.$$

Hence,

$$\|\widetilde{M}_t^\top \widetilde{M}_t - N_t^*\|_F \le \|\widetilde{M}_t^\top \widetilde{M}_t - \widetilde{N}_t\|_F + \|\widetilde{N}_t - N_t^*\|_F \le 2\|\widetilde{N}_t - N_t^*\|_F = \mathcal{O}(k t(d_y + d_u) d_x^{1/2} n^{-1/2}).$$

From now on, we consider $0 \le t \le \ell-1$ and $\ell \le t \le T$ separately, since, as we will show, in the latter case we have the additional condition that $\text{rank}(M_t^*) = d_x$.

For $0 \le t \le \ell-1$, $k = 1$. By Lemma 8, there exists a $d_x \times d_x$ orthonormal matrix S_t, such that $\|\widetilde{M}_t - S_t M_t^*\|_2 \le \|\widetilde{M}_t - S_t M_t^*\|_F = \mathcal{O}((t(d_y + d_u))^{1/2} d_x^{3/4} n^{-1/4})$. Recall that $\widehat{M}_t = \text{TRUNCSV}(\widetilde{M}_t, \theta)$; that is, $\widehat{M}_t = (\mathbb{I}_{[\theta, +\infty)}(\Sigma_{d_x}^{1/2}) \odot \Sigma_{d_x}^{1/2}) U_{d_x}^\top$. Then

$$\|\widehat{M}_t - \widetilde{M}_t\|_2 = \|(\mathbb{I}_{[\theta, +\infty)}(\Sigma_{d_x}^{1/2}) \odot \Sigma_{d_x}^{1/2} - \Sigma_{d_x}^{1/2}) U_{d_x}^\top\|_2 \le \|\mathbb{I}_{[\theta, +\infty)}(\Sigma_{d_x}^{1/2}) \odot \Sigma_{d_x}^{1/2} - \Sigma_{d_x}^{1/2}\|_2 \le \theta.$$

Hence, the distance between $\widehat{M}_t$ and M_t^* satisfies

$$\|\widehat{M}_t - S_t M_t^*\|_2 = \|\widehat{M}_t - \widetilde{M}_t + \widetilde{M}_t - S_t M_t^*\|_2 \leq \|\widehat{M}_t - \widetilde{M}_t\|_2 + \|\widetilde{M}_t - S_t M_t^*\|_2$$
$$\leq \theta + \mathcal{O}((t(d_y + d_u))^{1/2} d_x^{3/4} n^{-1/4}).$$

θ should be chosen in such a way that it keeps the error on the same order; that is, $\theta = \mathcal{O}((t(d_y + d_u))^{1/2} d_x^{3/4} n^{-1/4})$. As a result, since $\widehat{z}_t = \widehat{M}_t h_t$ and $z_t^* = M_t^* h_t$,

$$\mathrm{Cov}(\widehat{z}_t - S_t z_t^*) = \mathrm{Cov}((\widehat{M}_t - S_t M_t^*) h_t) = (\widehat{M}_t - S_t M_t^*) \mathbb{E}[h_t h_t^\top](\widehat{M}_t - S_t M_t^*)^\top.$$

Since $h_t = [y_{0:t}; u_{0:(t-1)}]$ and $(\mathrm{Cov}(y_t))_{t=0}^T$, $(\mathrm{Cov}(u_t))_{t=0}^{T-1}$ have $\mathcal{O}(1)$ operator norms, by Lemma 2, $\mathbb{E}[h_t h_t^\top] = \mathrm{Cov}(h_t) = \mathcal{O}(t)$. Hence,

$$\|\mathrm{Cov}(\widehat{z}_t - S_t z_t^*)^{1/2}\|_2 \leq \|\mathbb{E}[h_t h_t^\top]^{1/2}\|_2 \|\widehat{M}_t - S_t M_t^*\|_2 = \mathcal{O}(t(d_y + d_u)^{1/2} d_x^{3/4} n^{-1/4}).$$

On the other hand, since each element in h_t contains independent Gaussian noise,

$$\sigma_{\min}^+(\mathrm{Cov}(\widehat{z}_t)^{1/2}) = (\sigma_{\min}^+(\widehat{M}_t \mathbb{E}[h_t h_t^\top]\widehat{M}_t^\top))^{1/2} \geq \min(\sigma_u, \sigma_v)\sigma_{\min}^+(\widehat{M}_t) = \Omega(\theta).$$

Thus, singular value truncation ensures a lower bound for the minimum positive singular value of $\mathrm{Cov}(\widehat{z}_t)$. As shown in the proof of Lemma 10, this property is important for ensuring the system identification outputs $\widehat{A}_t$ and $\widehat{B}_t$ have bounded norms.

Note that $\mathbb{E}[\|\widehat{z}_t - S_t z_t^*\|] = \mathcal{O}(d_x^{1/2}\|\mathrm{Cov}(\widehat{z}_t - S_t z_t^*)^{1/2}\|_2) = \mathcal{O}(t(d_y + d_u)^{1/2} d_x^{5/4} n^{-1/4})$. In other words, we recover the latent states up to an orthonormal transform.

For $\ell \leq t \leq T$, $k \leq m$. By Proposition 2, $\mathrm{Cov}(z_t^*)$ has full rank, $\sigma_{\min}(\mathrm{Cov}(z_t^*)) = \Omega(\nu^2)$ and $\sigma_{\min}(M_t^*) = \Omega(\nu t^{-1/2})$. Recall that for $\ell \leq t \leq T$, we simply set $\widehat{M}_t = \widetilde{M}_t$. Then, by Lemma 7, there exists a $d_x \times d_x$ orthonormal matrix S_t, such that

$$\|\widehat{M}_t - S_t M_t^*\|_F = \mathcal{O}(\sigma_{\min}^{-1}(M_t^*))\|\widetilde{M}_t^\top \widetilde{M}_t - N_t^*\|_F = \mathcal{O}(\nu^{-1} m t^{3/2}(d_y + d_u) d_x^{1/2} n^{-1/2}),$$

which is also an upper bound on $\|\widehat{M}_t - S_t M_t^*\|_2$. As a result,

$$\|\mathrm{Cov}(\widehat{z}_t - S_t z_t^*)^{1/2}\|_2 \leq \|\mathbb{E}[h_t h_t^\top]^{1/2}\|_2 \|\widehat{M}_t - S_t M_t^*\|_2 = \mathcal{O}(\nu^{-1} m t^2(d_y + d_u) d_x^{1/2} n^{-1/2}).$$

Hence, there exists an absolute constant $c > 0$, such that if $n \geq c\nu^{-4} m^2 T^4 (d_y + d_u)^2 d_x$,

$$\|\mathrm{Cov}(\widehat{z}_t - S_t z_t^*)^{1/2}\|_2 \leq \sigma_{\min}(\mathrm{Cov}(z_t^*)^{1/2})/2,$$

which ensures that $\sigma_{\min}(\mathrm{Cov}(\widehat{z}_t)) = \Omega(\sigma_{\min}(\mathrm{Cov}(z_t^*))) = \Omega(\nu^2)$.

Recovery of the latent dynamics. The latent system $(A_t^*, B_t^*)_{t=0}^{T-1}$ is identified in Algorithm 3, using $(\widehat{z}_t^{(i)})_{i=1,t=0}^{N,T}$ produced by Algorithm 2, by the ordinary linear regression. Recall from Proposition 1 that $z_{t+1}^* = A_t^* z_t^* + B_t^* u_t + L_{t+1} i_{t+1}$. With the transforms on z_t^* and z_{t+1}^*, we have

$$S_{t+1} z_{t+1}^* = (S_{t+1} A_t^* S_t^\top) S_t z_t^* + S_{t+1} B_t^* u_t + S_{t+1} L_{t+1} i_{t+1},$$

and $(z_t^*)^\top Q_t^* z_t^* = (S_t z_t^*)^\top S_t Q_t^* S_t^\top S_t z_t^*$. Under control $u_t \sim \mathcal{N}(0, \sigma_u^2 I_{d_u})$ for $0 \le t \le T - 1$, we know that z_t^* is a zero-mean Gaussian random vector; so is $S_t z_t^*$. Let $\Sigma_t^* = \mathbb{E}[S_t z_t^* (z_t^*)^\top S_t^\top]$ be its covariance.

For $0 \le t \le \ell - 1$, by Lemma 10, there exists an absolute constant $c > 0$, such that as long as $n \ge c(d_x + d_u + \log(1/p))$, with probability at least $1 - p$,

$$\|([\widehat{A}_t, \widehat{B}_t] - [S_{t+1} A_t^* S_t^\top, S_{t+1} B_t^*]) \text{diag}((\Sigma_t^*)^{1/2}, \sigma_u I_{d_u})\|_2$$
$$= \mathcal{O}((1 + \beta^{-1}) \|\text{Cov}([\widehat{z}_t; u_t] - [S_t z_t^*; u_t])^{1/2}\|_2$$
$$+ \|\text{Cov}(\widehat{z}_{t+1} - S_{t+1} z_{t+1}^*)^{1/2}\|_2 + n^{-1/2}(d_x + d_u + \log(1/p))^{1/2}),$$

which implies

$$\|(\widehat{A}_t - S_{t+1} A_t^* S_t^\top)(\Sigma_t^*)^{1/2}\|_2 = \mathcal{O}((1 + \beta^{-1}) t (d_y + d_u)^{1/2} d_x^{3/4} n^{-1/4}),$$
$$\|\widehat{B}_t - S_{t+1} B_t^*\|_2 = \mathcal{O}((1 + \beta^{-1}) t (d_y + d_u)^{1/2} d_x^{3/4} n^{-1/4}).$$

For $\ell \le t \le T - 1$, by Lemma 11, with probability at least $1 - p$,

$$\|[\widehat{A}_t, \widehat{B}_t] - [S_{t+1} A_t^* S_t^\top, S_{t+1} B_t^*]\|_2$$
$$= \mathcal{O}((\nu^{-1} + \sigma_u^{-1})(\|\text{Cov}([\widehat{z}_t; u_t] - [S_t z_t^*; u_t])^{1/2}\|_2$$
$$+ \|\text{Cov}(\widehat{z}_{t+1} - S_{t+1} z_{t+1}^*)^{1/2}\|_2 + n^{-1/2}(d_x + d_u + \log(1/p))^{1/2})),$$

which implies

$$\|\widehat{A}_t - S_{t+1} A_t^* S_t^\top\|_2 = \mathcal{O}\left((1 + \nu^{-1})(mt^2(d_y + d_u)d_x^{1/2} + (\log(1/p))^{1/2})n^{-1/2}\right),$$
$$\|\widehat{B}_t - S_{t+1} B_t^*\|_2 = \mathcal{O}\left((1 + \nu^{-1})(mt^2(d_y + d_u)d_x^{1/2} + (\log(1/p))^{1/2})n^{-1/2}\right).$$

We note that since the observation of u_t is exact, a tighter bound on $\|\widehat{B}_t - S_{t+1} B_t^*\|_2$ is possible.

By Assumption 3, $(Q_t^*)_{t=0}^{\ell-1}$ and Q_T^* are positive definite; they are identity matrices under the normalized parameterization. For $(Q_t^*)_{t=\ell}^{T-1}$, which may not be positive definite, we identify them in SysID (Algorithm 3) by (E.2). We have shown that there exists an absolute constant $c > 0$, such that as long as $n \ge c\nu^{-4} m^2 T^4 (d_y + d_u)^2 d_x$, $\|\text{Cov}(\widehat{z}_t - S_t z_t^*)^{1/2}\|_2 \le \sigma_{\min}(\text{Cov}(z_t^*)^{1/2})/2$. Since orthonormal matrices do not change singular values, $\sigma_{\min}(\text{Cov}(z_t^*)^{1/2}) = \sigma_{\min}(\text{Cov}(S_t z_t^*)^{1/2})$. Hence, the requirement on $\|\Sigma_\delta^{1/2}\|_2$ in Lemma 6 is satisfied, with h^* and δ in Lemma 6 corresponding to $S_t z_t^*$ and $(\widehat{z}_t - S_t z_t^*)$ here. By Lemma 6,

$$\|\widetilde{Q}_t - S_t Q_t^* S_t^\top\|_F = \mathcal{O}(\nu^{-2} d_x((1 + \nu^{-1}) mt^2 (d_y + d_u) d_x^{1/2} n^{-1/2} + m d_x^{1/2} n^{-1/2}))$$
$$= \mathcal{O}((1 + \nu^{-3}) mt^2 (d_y + d_u) d_x^{3/2} n^{-1/2}).$$

By construction, $\widehat{Q}_t$ is the projection of $\widetilde{Q}_t$ onto the positive semidefinite cone with respect to the Frobenius norm (Boyd et al., 2004, Section 8.1.1). Since $S_t Q_t^* S_t^\top \succcurlyeq 0$,

$$\|\widehat{Q}_t - S_t Q_t^* S_t^\top\|_F \le \|\widetilde{Q}_t - S_t Q_t^* S_t^\top\|_F.$$

Certainty equivalent linear quadratic control. The last step of Algorithm 1 is to compute the optimal controller in the estimated system $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$ by RDE (2.3).

RDE (2.3) proceeds backward. For $\ell \leq t \leq T-1$, $\|\widehat{A}_t - S_{t+1}A_t^* S_t^\top\|_2$, $\|\widehat{B}_t - S_{t+1}B_t^*\|_2$ and $\|\widehat{Q}_t - S_t Q_t^* S_t^\top\|_2$ are all bounded by

$$\mathcal{O}((1+v^{-3})((d_y + d_u)d_x^{3/2}mT^2 + \log(1/p)^{1/2})n^{-1/2}).$$

By Lemma 13, for $n \geq \text{poly}(T, d_x, d_y, d_u, \log(1/p))$, where hidden constants are dimension-free and depend polynomially on system parameters, the controller $(\widehat{K}_t)_{t=\ell}^{T-1}$ is ϵ-optimal in system $((S_{t+1}A_t^* S_t^{-1}, S_{t+1}B_t^*, R_t^*)_{t=\ell}^{T-1}, (S_t Q_t^* S_t^\top)_{t=\ell}^T)$, for

$$\epsilon = \mathcal{O}((d_x \wedge d_u)T^2(1+v^{-6})((d_y + d_u)^2 d_x^3 m^2 T^2 + \log(1/p))n^{-1}), \tag{E.3}$$

that is, if

$$n \geq c(1+v^{-6})d_x^4(d_y + d_u + \log(1/p))^2 m^2 T^4 \epsilon^{-1},$$

for a dimension-free constant $c > 0$ that depends on system parameters.

For $0 \leq t \leq \ell - 1$, $(\widehat{K}_t)_{t=0}^{\ell-1}$ and $(K_t^*)_{t=0}^{\ell-1}$ are optimal controllers in the ℓ-step systems with terminal costs given by $\widehat{P}_\ell$ and P_ℓ^* respectively, where $\widehat{P}_\ell$ and P_ℓ^* are the solutions to RDE (2.3) in systems $((\widehat{A}_t, \widehat{B}_t, R_t^*)_{t=0}^{T-1}, (\widehat{Q}_t)_{t=0}^T)$ and $((S_{t+1}A_t^* S_t^{-1}, S_{t+1}B_t^*, R_t^*)_{t=0}^{T-1}, (S_t Q_t^* S_t^\top)_{t=0}^T)$ at step ℓ, respectively. By Lemma 13,

$$\|\widehat{P}_\ell - P_\ell^*\|_2 = \mathcal{O}((1+v^{-3})((d_y + d_u)d_x^{3/2}mT^2 + (\log(1/p))^{1/2})n^{-1/2}),$$

which is dominated by the $n^{-1/4}$ rate of $\|(\widehat{A}_t - S_{t+1}A_t^* S_t^\top)(\Sigma_t^*)^{1/2}\|_2$ and $\|\widehat{B}_t - S_{t+1}B_t^*\|_2$ for $0 \leq t \leq \ell - 1$. Then, by Lemma 12, $(\widehat{K}_t)_{t=0}^{\ell-1}$ is ϵ-optimal in system $((S_{t+1}A_t^* S_t^{-1}, S_{t+1}B_t^*, R_t^*)_{t=0}^{\ell-1}, (S_t Q_t^* S_t^\top)_{t=0}^{\ell-1})$ with terminal cost matrix P_t^*, for

$$\begin{aligned}
\epsilon &= \mathcal{O}(c^\ell((1+\beta^{-1})d_x((1+\beta^{-1})\ell(d_y + d_u)^{1/2}d_x^{3/4}n^{-1/4}) + d_x(d_x + d_u + \log(1/p))^{1/2}n^{-1/2})) \\
&= \mathcal{O}((1+\beta^{-1})d_x^{7/4}\ell(d_y + d_u)^{1/2}c^\ell n^{-1/4}), \tag{E.4}
\end{aligned}$$

where dimension-free constant $c > 0$ depends on the system parameters; that is, if

$$n \geq a_0(1+\beta^{-4})d_x^7(d_y + d_u)^2 \ell^4 a^\ell \epsilon^{-4},$$

for some dimension-free constants $a_0, a > 0$ that depend on system parameters. The bound (E.4) is much worse than (E.3), because for $0 \leq t \leq \ell - 1$, z_t^* does not have full-rank covariance, and $S_{t+1}A_t^* S_t^\top$ is only recovered partially. Even with large enough data, linear regression has no guarantee for $\|\widehat{A}_t - S_{t+1}A_t^* S_t^\top\|_2$ to be small; we do not know the controllability of $(\widehat{A}_t, \widehat{B}_t)_{t=0}^{\ell-1}$, not even its stabilizability. $\qquad\square$