Toward Understanding State Representation Learning in MuZero: A Case Study in 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 problem of state representation learning for control from partial and potentially high-dimensional observations. We approach this problem via *cost-driven state representation learning*, where one learns a dynamical model in some latent state space by *predicting (cumulative) costs*. In particular, we establish *finite-sample guarantees* of finding a near-optimal representation function and a near-optimal controller using the learned latent model for infinite-horizon time-invariant Linear Quadratic Gaussian (LQG) control. We study two approaches to cost-driven representation learning, which differ in whether the transition function of the latent model is learned explicitly. The first approach has also been investigated in (Tian et al., 2023), for finite-horizon time-varying LQG. The second approach closely resembles *MuZero*, a recent breakthrough in empirical reinforcement learning, in that it learns latent dynamics *implicitly* by predicting *cumulative costs*. A key technical contribution of this work is to prove persistency of excitation for a new stochastic process that arises from the analysis of quadratic regression in our approach, which may be of independent interest.

1 Introduction

Control with a *learned* latent model has achieved state-of-the-art performance in several reinforcement learning (RL) benchmarks, including board games, Atari games, and visuomotor control (Schrittwieser et al., 2020; Ye et al., 2021; Hafner et al., 2023). To better understand this machinery in RL, we introduce it to a classical optimal control problem, namely Linear Quadratic Gaussian (LQG) control, and study its theoretical, in particular, finite-sample performance. Essential to this approach is the learning of two components: a *state representation* function that maps an observed history to some latent state, and a *latent model* that predicts the transition and cost in the latent state space. The latent model is usually a Markov decision process, using which we obtain a policy in the latent space or execute online planning.

What is the correct *objective* to optimize for learning a good latent model? One popular choice is to learn a function that *reconstructs the observation* from the latent state (Hafner et al., 2019a,b, 2020, 2023). A latent model learned this way is *agnostic to control tasks* and retains all the information about the environment. This class of approaches may achieve satisfactory performance empirically, but are prone to background distraction and control-irrelevant information (Fu et al., 2021). The second class of methods learn an *inverse model* that infers actions from latent states at different time steps (Pathak et al., 2017; Lamb et al., 2022). A latent model learned with this methodology is also task-agnostic but can extract *control-relevant* information. In contrast, *task-relevant* representations can be learned by *predicting costs* in the control task (Oh et al., 2017; Zhang et al., 2020; Schrittwieser et al., 2020). The concept that a good latent state should be able to predict costs is intuitive, as the costs are directly relevant to optimal control. This class of methods is the focus of this work.

The cost-driven state representation learning method of particular interest to us is that of MuZero (Schrittwieser et al., 2020). Announced by DeepMind in 2019, MuZero extends the line of works including AlphaGo (Silver et al., 2016), AlphaGo Zero (Silver et al., 2017), and AlphaZero (Silver et al., 2018) by obviating the knowledge of the game rules. MuZero matches the superhuman performance of AlphaZero in Go, shogi and chess, while outperforming model-free RL algorithms in Atari games. MuZero builds upon the powerful planning procedure of Monte Carlo Tree Search, with the major innovation being *learning a latent model*. The latent model replaces the rule-based simulator during planning, and avoids the burdensome planning in pixel space for Atari games.

MuZero is a milestone algorithm in representation learning for control. Intuitively, the algorithm design makes sense, but its complexity has so far inhibited a formal theoretical study. On the other hand, statistical learning theory for linear dynamical systems and control has evolved rapidly in recent years (Tsiamis et al., 2022); for partially observable linear dynamical systems, much of the work relies on learning *Markov parameters*, lacking a direct connection to the empirical methods used in practice for possibly nonlinear systems. In this work, we aim to bridge the two areas by studying provable MuZero-style latent model learning in LQG control.

The latent model learning of MuZero features three ingredients: 1) stacking frames, i.e., observations, as input to the representation function; 2) predicting costs, “optimal” values, and “optimal” actions from latent states; and 3) implicit learning of latent dynamics by predicting these quantities from latent states at future time steps. These are the defining characteristics of the MuZero-style algorithm that we shall consider. In MuZero, the “optimal” values and actions are found by the powerful online planning procedure. In this work, we simplify the setup by considering data collected using random actions, which are known to suffice for identifying a partially observable linear dynamical system (Oymak and Ozay, 2019). In this setup, the values become those associated with this trivial policy and we do not predict actions since they are simply random noises. Note that although our study of the above ingredients is directly motivated by MuZero, previous empirical works have also explored them. For example, frame stacking has been a widely used technique to handle partial observability (Mnih et al., 2013, 2015); predicting values for learning a latent model has been studied in (Oh et al., 2017), which also learns the latent state transition implicitly.

In (Tian et al., 2023), we have considered provable cost-driven state representation learning in LQG for the *finite-horizon time-varying* setting. This work builds upon it and complements it in two ways: 1) we extend their algorithm to the *time-invariant* setting with a *stationary* representation function and latent model, which is closer to what has been deployed in practice; 2) we present and analyze a new, MuZero-style latent model learning algorithm. Both 1) and 2) introduce new technical challenges to be addressed. We summarize our contributions as follows.

- We show that two cost-driven state representation learning methods provably solve infinite-horizon time-invariant LQG control by establishing finite-sample guarantees. Both methods only need a single trajectory; one resembles the method in (Tian et al., 2023), and the other resembles MuZero.
- By analyzing the MuZero-style algorithm, we notice the potential issue of *coordinate misalignment*: Costs can be invariant to orthogonal transformations of the latent states, and implicit dynamics learning by predicting *one-step* transition may not recover the latent state coordinates consistently. This insight suggests the need to predict *multi-step* latent transition or other coordinate alignment procedures in the MuZero-style implicit dynamics learning approaches.
- Technically, we overcome the difficulty of having *correlated* data in a single trajectory for latent model learning, as we are dealing with the time-invariant setting and need to aggregate samples across time steps in contrast to (Tian et al., 2023). To do so, we prove a new result about the persistency of excitation for a stochastic process that arises from the analysis of the quadratic regression subroutine in both of our methods.

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. Let $a \wedge b$ denote the minimum between scalars a and b. Let 0 (resp. 1) denote either the scalar or a matrix consisting of all zeros (resp. all ones); let I denote an identity matrix. The dimension, when emphasized, is specified in subscripts, e.g., $0_{d_x \times d_x}, 1_{d_x}, I_{d_x}$. The dimension, when emphasized, is specified in subscripts, e.g., $1_d, I_d$. Given vector $v \in \mathbb{R}^d$, let $\|v\|$ denote its ℓ_2 norm and $\|v\|_P := (v^\top P v)^{1/2}$ for positive semidefinite $P \in \mathbb{R}^{d \times d}$. Given symmetric matrices P and Q, $P \succ Q$ or $Q \prec P$ means $P - Q$ is positive definite, and $P \succcurlyeq Q$ or $Q \preccurlyeq P$ 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 $v_{i:j} := [v_i; v_{i+1}; \dots; v_j] \in \mathbb{R}^{d(j-i+1)}$ and $v_{j:i} := [v_j; v_{j-1}; \dots; v_i] \in \mathbb{R}^{d(j-i+1)}$ denote the concatenation along the column. For random variable x, let $\|x\|_{\psi_\theta}$ denote its θ-sub-Weibull norm for $\theta > 0$, a special case of Orlicz norms (Zhang and Wei, 2022), with $\theta = 1, 2$ corresponding to subexponential and sub-Gaussian norms. For random vector x, y, let $\mathrm{Cov}(x, y)$ denote the covariance matrix between x and y; with a slight abuse of notation, define $\mathrm{Cov}(x) := \mathrm{Cov}(x, x)$. For matrix A, let $\sigma_{\min}(A)$, $\|A\|_2$, $\|A\|_F$, and $\|A\|_*$ denote its minimum eigenvalue, minimum singular value, operator norm (induced by vector ℓ_2 norms), Frobenius norm, and nuclear norm, respectively. $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product between matrices. For square matrix A, let $\lambda_{\min}(A)$ be its minimum eigenvalue and $\rho(A)$ be its spectral radius. Define $\alpha(A) := \sup_{k \geq 0} \|A^k\|_2 \rho(A)^{-k}$. Let $\mathrm{svec}(\cdot)$ denote the operator of flattening a symmetric matrix by stacking its columns; it 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 may depend on system parameters.

2 Problem setup

A partially observable linear time-invariant (LTI) dynamical system is described by

$$x_{t+1} = A^* x_t + B^* u_t + w_t, \quad y_t = C^* x_t + v_t, \tag{2.1}$$

with state $x_t \in \mathbb{R}^{d_x}$, observation $y_t \in \mathbb{R}^{d_y}$, and control $u_t \in \mathbb{R}^{d_u}$ for all $t \geq 0$. Process noises $(w_t)_{t \geq 0}$ and observation noises $(v_t)_{t \geq 0}$ are i.i.d. zero-mean Gaussian random vectors with covariance matrices Σ_w and Σ_v, respectively, and the two sequences are mutually independent. Let initial state x_0 be sampled from $\mathcal{N}(0, \Sigma_0)$. The quadratic cost function is given by

$$c(x, u) = \|x\|_{Q^*}^2 + \|u\|_{R^*}^2, \tag{2.2}$$

where $Q^* \succcurlyeq 0$ and $R^* \succ 0$.

A policy/controller π determines an action/control input u_t at time step t based on the history $[y_{0:t}; u_{0:(t-1)}]$ up to this time step. For $t \geq 0$, let $c_t := c(x_t, u_t)$ denote the cost at time step t. Given a policy π, let

$$J^\pi := \limsup_{T \to \infty} \mathbb{E}\left[\frac{1}{T} \sum\nolimits_{t=0}^{T-1} c_t\right] \tag{2.3}$$

denote the time-averaged expected cost. The objective of LQG control is to find a policy π such that J^π is minimized.

In the fully observable setting, known as the linear quadratic regulator (LQR) problem, we have $y_t = x_t$. A linear controller with feedback gain $K \in \mathbb{R}^{d_u \times d_x}$ determines action $u_t = K x_t$ at time step t. Let $J^K(A^*, B^*, Q^*, R^*)$ denote the time-averaged expected cost (2.3) in the LQR problem (A^*, B^*, Q^*, R^*) under feedback gain K and define $J^*(A^*, B^*, Q^*, R^*) := \min_K J^K(A^*, B^*, Q^*, R^*)$.

We make the following standard assumptions.

Assumption 1. *System dynamics (2.1) and cost (2.2) satisfy:*

1. *The system is stable, that is, $\rho(A^*) < 1$.*
2. *(A^*, B^*) is ν-controllable for some $\nu > 0$, that is, the controllability matrix*

$$\Phi_c(A^*, B^*) := [B^*, A^* B^*, \ldots, (A^*)^{d_x - 1} B^*]$$

has rank d_x and $\sigma_{\min}(\Phi_c(A^*, B^*)) \geq \nu$.

3. *(A^*, C^*) is ω-observable for some $\omega > 0$, that is, the observability matrix*

$$\Phi_o(A^*, C^*) := [C^*; C^*A^*; \dots; C^*(A^*)^{d_x - 1}]$$

   has rank d_x and $\sigma_{\min}(\Phi_o(A^*, C^*)) \geq \omega$.

4. *$(A^*, \Sigma_w^{1/2})$ is κ-controllable for some $\kappa > 0$.*

5. *$(A^*, (Q^*)^{1/2})$ is μ-observable for some $\mu > 0$.*

6. *$\Sigma_v \succcurlyeq \sigma_v^2 I$ for some $\sigma_v > 0$; this can always be achieved by inserting Gaussian noises with full-rank covariance matrices to the observations.*

7. *$R^* \succcurlyeq r^2 I$ for some $r > 0$.*

8. *The operator norms of A^*, B^*, C^*, Q^*, R^*, Σ_w, Σ_v, Σ_0 and $\alpha(A^*), \alpha(\overline{A}^*)$ are $\mathcal{O}(1)$, where we recall that for a square matrix A, $\alpha(A) := \sup_{k \geq 0} \|A^k\|_2 \rho(A)^{-k}$; the singular value lower bounds ν, ω, ν, κ, σ_v, r and spectral radii $\rho(A^*), \rho(\overline{A}^*)$ are $\Omega(1)$, where $\overline{A}^*$ is defined in §2.1.*

If the system parameters $(A^*, B^*, C^*, Q^*, R^*, \Sigma_w, \Sigma_v)$ are known, the optimal policy is obtained by combining the Kalman filter

$$z_{t+1}^* = A^* z_t^* + B^* u_t + L^*(y_{t+1} - C^*(A^* z_t^* + B^* u_t)) \tag{2.4}$$

with the optimal feedback gain K^* of the linear quadratic regulator such that $u_t = K^* z_t^*$, where L^* is the Kalman gain, and at the initial time step, we can set, e.g., $z_0^* = L^* y_0$. This fact is known as the *separation principle*, and the Kalman gain and optimal feedback gain are given by

$$L^* = S^*(C^*)^\top (C^* S^* (C^*)^\top + \Sigma_v)^{-1}, \tag{2.5}$$

$$K^* = -((B^*)^\top P^* B^* + R)^{-1} (B^*)^\top P^* A^*, \tag{2.6}$$

where S^* and P^* are determined by their respective discrete-time algebraic Riccati equations (DAREs):

$$S^* = A^* \big(S^* - S^*(C^*)^\top (C^* S^* (C^*)^\top + \Sigma_v)^{-1} C^* S^*\big) (A^*)^\top + \Sigma_w, \tag{2.7}$$

$$P^* = (A^*)^\top \big(P^* - P^* B^* ((B^*)^\top P^* B^* + R^*)^{-1} (B^*)^\top P^*\big) A^* + Q^*. \tag{2.8}$$

Assumptions 1.2 to 1.7 guarantee the existence and uniqueness of positive definite solutions S^* and P^*; Assumption 1.8 further guarantees that their operator norms are $\mathcal{O}(1)$ and minimum singular values are $\Omega(1)$. The assumption on $\alpha(A^*), \alpha(\overline{A}^*), \rho(A^*), \rho(\overline{A}^*)$ provides guarantees for state estimation from a finite history and has also been made in the literature (Mania et al., 2019; Oymak and Ozay, 2019). If $\rho(A^*)$ or $\rho(\overline{A}^*)$ equals zero, then $(A^*)^{d_x}$ or $(\overline{A}^*)^{d_x}$ is a zero matrix by the Cayley-Hamilton theorem, so using history length $H \geq d_x$ completely eliminates the truncation errors. Thus, Assumption 1.8 does not lose generality.

We consider the data-driven control setting, where the LQG model $(A^*, B^*, C^*, Q^*, \Sigma_w, \Sigma_v)$ is unknown. For simplicity, we assume R^* is known, though our approaches can be readily extended to the case where it is unknown by learning it from predicting costs.

2.1 Latent model of infinite-horizon time-invariant LQG

The stationary Kalman filter (2.4) asymptotically produces the optimal *state estimation* in the sense of minimum mean squared errors. With a finite horizon, however, the optimal state estimator is time-varying, given by

$$z^*_{t+1} = A^* z^*_t + B^* u_t + L^*_{t+1}(y_{t+1} - C^*(A^* z^*_t + B^* u_t)), \tag{2.9}$$

where L^*_t is the time-varying Kalman gain, converging to L^* as $t \to \infty$. This convergence is equivalent to that of error covariance matrix $\mathbb{E}[(x_t - z^*_t)(x_t - z^*_t)^\top]$, which is exponentially fast (Komaroff, 1994). Hence, for simplicity, we assume this error covariance matrix is stationary at the initial time step by the choice of z^*_0 so that $L^*_t = L^*$ for $t \geq 1$; this assumption has also been adopted in the literature (Lale et al., 2020, 2021; Jadbabaie et al., 2021). The *innovation* term $i_{t+1} := y_{t+1} - C^*(A^* z^*_t + B^* u_t)$ is independent of the history $(y_0, u_0, y_1, \ldots, u_{t-1}, y_t)$ and $(i_t)_{t\geq 1}$ are mutually independent. The following proposition taken from Proposition 1 in (Tian et al., 2023) represents the system in terms of the state estimates obtained by the Kalman filter, which we refer to as the *latent model*.

Proposition 1. *Let $(z^*_t)_{t\geq 1}$ be state estimates given by the Kalman filter. Then, for $t \geq 0$,*

$$z^*_{t+1} = A^* z^*_t + B^* u_t + L^* i_{t+1},$$

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

$$c_t = \|z^*_t\|^2_{Q^*} + \|u_t\|^2_{R^*} + b^* + \gamma_t + \eta_t,$$

where $b^* = \mathbb{E}[\|x_t - z^*_t\|^2_{Q^*}] > 0$, and $\gamma_t = \|x_t - z^*_t\|^2_{Q^*} - b^*$, $\eta_t = \langle z^*_t, x_t - z^*_t \rangle_{Q^*}$ are both zero-mean subexponential random variables independent of . Moreover, $b^* = \mathcal{O}(1)$ and $\|\gamma_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$; if control $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $t \geq 0$, then we have $\|\eta_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$.

Proposition 1 shows that the dynamics of the state estimates computed by the time-varying Kalman filter is the same as the original system up to noises; the costs are also the same, up to constants and noises. Hence, a latent model can be parameterized by (A, B, Q, R^*), with the constant b^* and noises neglected due to their irrelevance to planning. A stationary latent policy is a linear controller $u_t = K z_t$ on latent state z_t, parameterized by feedback gain $K \in \mathbb{R}^{d_u \times d_x}$.

The latent model enables us to find a good latent policy. To learn such a latent model and to deploy a latent policy in the original partially observable system, we need a representation function. Let $\overline{A}^* := (I - L^* C^*) A^*$ and $\overline{B}^* := (I - L^* C^*) B^*$. Then, the Kalman filter can be written as $z^*_{t+1} = \overline{A}^* z^*_t + \overline{B}^* u_t + L^* y_{t+1}$. For $t \geq 0$, unrolling the recursion gives

$$\begin{aligned}
z^*_t &= \overline{A}^*(\overline{A}^* z^*_{t-2} + \overline{B}^* u_{t-2} + L^* y_{t-1}) + \overline{B}^* u_{t-1} + L^* y_t \\
&= [(\overline{A}^*)^{t-1} L^*, \ldots, L^*] y_{1:t} + [(\overline{A}^*)^{t-1} \overline{B}^*, \ldots, \overline{B}^*] u_{0:(t-1)} + (\overline{A}^*)^t z^*_0 \\
&=: M^*_t [y_{1:t}; u_{0:(t-1)}; z^*_0],
\end{aligned}$$

where $M_t^* \in \mathbb{R}^{d_x \times (td_y + td_u + d_x)}$. This means the representation function can be parameterized as linear mappings for full histories (with y_0 replaced by z_0^*). Despite the simplicity, the input dimension of the function grows linearly in time, making it intractable to estimate the state using the full history for large t; nor it is necessary, since the impact of old data decreases exponentially. Under Assumption 1, $\rho(\overline{A}^*) < 1$ (Bertsekas, 2012, Appendix E.4). With an H-step truncated history, the state estimate can be written as

$$\begin{aligned} z_t^* &= [(\overline{A}^*)^{H-1} L^*, \ldots, L^*] y_{(t-H+1):t} + [(\overline{A}^*)^{H-1} \overline{B}^*, \ldots, \overline{B}^*] u_{(t-H):(t-1)} + \delta_t \\ &=: M^* [y_{(t-H+1):t}; u_{(t-H):(t-1)}] + \delta_t, \end{aligned} \tag{2.10}$$

where $\delta_t = (\overline{A}^*)^H z_{t-H}^*$, whose impact decays exponentially in H and can be neglected for sufficiently large H, since z_{t-H}^* converges to a stationary distribution and its norm is bounded with high probability. Hence, the representation function that we aim to recover is $M^* \in \mathbb{R}^{d_x \times H(d_y + d_u)}$, which takes as input the H-step history $h_t = [y_{(t-H+1):t}; u_{(t-H):(t-1)}]$. Henceforth, we let $d_h := H(d_y + d_u)$. Then, a representation function is parameterized by matrix $M \in \mathbb{R}^{d_x \times d_h}$.

Overall, a policy is a combination of a state representation function M and a feedback gain K in the latent model, denoted by $\pi = (M, K)$. Learning to solve LQG control in this framework can thus be achieved by: 1) learning state representation function M; 2) extracting latent model (A, B, Q, R^*); and 3) finding the optimal K by planning in the latent model. Next, we introduce our approach following this pipeline.

3 Method

In practice, latent model learning methods collect trajectories by interacting with the system online using some policy; the trajectories are used to improve the learned latent model, which in turn improves the policy. In LQG control, it is known that the simple setup allows us to learn a good latent model from a single trajectory, collected using zero-mean Gaussian inputs; see e.g., (Oymak and Ozay, 2019). This is also how we assume the data are collected. We note that our results also apply to data from multiple independent trajectories using inputs from the same zero-mean Gaussian distribution.

In our cost-driven state representation learning approach, state representations are learned by predicting costs. To learn the transition function in the latent model, two approaches have been explored in the literature. The first approach explicitly minimizes the transition prediction error (Subramanian et al., 2020; Hafner et al., 2019a). Algorithmically, the overall loss is a combination of cost prediction and transition prediction errors. The second approach, as MuZero in Schrittwieser et al. (2020) takes, learns the transition dynamics *implicitly*, by minimizing *cost prediction errors* at *future states* generated from the transition function (Oh et al., 2017; Schrittwieser et al., 2020). Algorithmically, the overall loss aggregates the cost prediction errors *across multiple time steps*. In both approaches, the coupling of different terms in the loss makes finite-sample analysis difficult. As observed in (Tian et al., 2023), the structure of LQG allows us to learn the representation function independently of learning the transition function. This allows us to formulate both approaches under the same cost-driven state representation learning framework (c.f. Algorithm 1).

Algorithm 1 Cost-driven state representation learning

```
1: **Input:** length $T$, history length $H$, noise magnitude $\sigma_u$
2: Collect a trajectories of length $T+H$ using $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, for $t \geq 0$, to obtain

       $\mathcal{D}_{\text{raw}} = (y_0, u_0, c_0, y_1, u_1, c_1, \ldots, y_{T+H-1}, u_{T+H-1}, c_{T+H-1}, y_{T+H})$    (3.1)

3: Estimate the state representation function and cost constants by solving

       $\hat{N}, \hat{b}_0 \in \operatorname*{argmin}_{N=N^\top, b_0} \sum_{t=H}^{T+H-1} \left(\|h_t\|_N^2 + b_0 - \overline{c}_t\right)^2,$    (3.2)

   where $h_t = [y_{(t-H+1):t}; u_{(t-H):(t-1)}]$ and $\overline{c}_t := \sum_{\tau=t}^{t+d_x-1}(c_\tau - \|u_\tau\|_{R^*}^2)$
4: Find $\hat{M} \in \operatorname{argmin}_{M \in \mathbb{R}^{d_x \times H(d_y+d_u)}} \|M^\top M - \hat{N}\|_F$
5: Compute $\hat{z}_t = \hat{M}[y_{(t-H+1):t}; u_{(t-H):(t-1)}]$ for all $t \geq H$, so that the data are converted to $\mathcal{D}_{\text{state}}$:

       $(\hat{z}_H, u_H, c_H, \ldots, \hat{z}_{T+H-1}, u_{T+H-1}, c_{T+H-1}, \hat{z}_{T+H})$

6: Run SysId (3.4) or CoSysId (Algorithm 2) to obtain dynamics matrices $(\hat{A}, \hat{B})$
7: Estimate the cost function by solving

       $\widetilde{Q}, \hat{b} \in \operatorname*{argmin}_{Q=Q^\top, b} \sum_{t=H}^{T+H-1} (\|\hat{z}_t\|_Q^2 + \|u_t\|_{R^*}^2 + b - c_t)^2$    (3.3)

8: Truncate negative eigenvalues of $\widetilde{Q}$ to zero to obtain $\hat{Q} \succcurlyeq 0$
9: Find feedback gain $\hat{K}$ from $(\hat{A}, \hat{B}, \hat{Q}, R^*)$ by solving DARE (2.8) and (2.6)
10: **Return:** policy $\hat{\pi} = (\hat{M}, \hat{K})$
```

Algorithm 1 consists of three main steps. Lines 3 to 5 correspond to cost-driven representation function learning. Lines 6 to 8 correspond to latent model learning, where the system dynamics can be identified either explicitly, by ordinary least squares (SysId), or implicitly, by future cost prediction (CoSysId, Algorithm 2). Line 9 corresponds to the policy optimization procedure in the latent model; in LQG this amounts to solving DAREs. Below we elaborate on cost-driven representation learning, SysId, and CoSysId in order.

3.1 Cost-driven representation function learning

The procedure of cost-driven representation function learning is consistent with (Tian et al., 2023). The main idea is to perform quadratic regression (3.2) to the d_x-step cumulative costs; this step corresponds to the value prediction in MuZero. By the μ-observability of $(A^*, (Q^*)^{1/2})$ (Assumption 1.5), the cost observability Gram matrix satisfies

$$\overline{Q}^* := \sum_{t=0}^{d_x-1}((A^*)^t)^\top Q^* (A^*)^t \succcurlyeq \mu^2 I.$$

Under zero control and zero noise, starting from x, the d_x-step cumulative cost is precisely $\|x\|^2_{\overline{Q}^*}$. Hence, with the impact of zero-mean control inputs and zero-mean noises averaged out, $\hat{N}$ estimates $N^* = (M^*)^\top \overline{Q}^* M^*$; up to an orthogonal transformation, $\hat{M}$ recovers $M^{*\prime} := (\overline{Q}^*)^{1/2} M^*$, the representation function under an equivalent parameterization, termed as the *normalized parameterization* in (Tian et al., 2023), where

$$A^{*\prime} = (\overline{Q}^*)^{1/2} A^* (\overline{Q}^*)^{-1/2}, \quad B^{*\prime} = (\overline{Q}^*)^{1/2} B, \quad C^{*\prime} = C^* (\overline{Q}^*)^{-1/2},$$
$$w_t' = (\overline{Q}^*)^{1/2} w_t, \quad Q^{*\prime} = (\overline{Q}^*)^{-1/2} Q^* (\overline{Q}^*)^{-1/2}.$$

Due to the following proposition, the algorithm does not need to know the dimension d_x of the latent model; it can discover d_x from the eigenvalues of $\hat{N}$.

Proposition 2. *Under i.i.d. control inputs $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $t \geq 0$, $\lambda_{\min}(\mathrm{Cov}(z_t^*)) = \Omega(\nu^2)$ for $t \geq d_x$, where ν is defined in Assumption 1.2. As long as $H \geq \frac{\log(a\alpha(\overline{A}^*))}{\log(\rho(\overline{A}^*)^{-1})}$ for some dimension-free constant $a > 0$, M^* has rank d_x and $\sigma_{\min}(M^*) \geq \Omega(\nu H^{-1/2})$.*

Proof. For $t \geq d_x$, unrolling the Kalman filter gives

$$\begin{aligned} z_t^* &= A^* z_{t-1}^* + B^* u_{t-1} + L^* i_t \\ &= A^*(A^* z_{t-2}^* + B^* u_{t-2} + L^* i_{t-1}) + L^* i_t \\ &= [B^*, \ldots, (A^*)^{d_x - 1} B^*][u_{t-1}; \ldots; u_{t-d_x}] + (A^*)^{d_x} z_{t-d_x}^* + [L^*, \ldots, (A^*)^{d_x-1} L^*][i_t; \ldots; i_{t-d_x+1}], \end{aligned}$$

where $(u_\tau)_{\tau=t-d_x}^{t-1}$, $z_{t-d_x}^*$ and $(i_\tau)_{\tau=t-d_x+1}^{t}$ are independent. For $H \geq d_x$, the matrix multiplied by $[u_{t-1}; \ldots; u_{t-d_x}]$ is precisely the controllability matrix $\Phi_c(A^*, B^*)$. Then,

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

By the ν-controllability of (A^*, B^*), $\mathrm{Cov}(z_t^*)$ is full-rank and $\lambda_{\min}(\mathrm{Cov}(z_t^*)) \geq \sigma_u^2 \nu^2$. Since $z_t^* = M^* h_t + \delta_t$ by (2.10), we have

$$\mathrm{Cov}(M^* h_t) = \mathrm{Cov}(z_t^* - \delta_t) = \mathrm{Cov}(z_t^*) + \mathrm{Cov}(\delta_t) - \mathrm{Cov}(z_t^*, \delta_t) - \mathrm{Cov}(\delta_t, z_t^*).$$

Then,

$$\begin{aligned} \|\mathrm{Cov}(z_t^*, \delta_t)\|_2 = \|\mathrm{Cov}(\delta_t, z_t^*)\|_2 &= \|\mathbb{E}[z_t^* \delta_t^\top]\|_2 \\ &\overset{(i)}{\leq} \|\mathbb{E}[z_t^* (z_t^*)^\top]\|_2^{1/2} \cdot \|\mathbb{E}[\delta_t \delta_t^\top]\|_2^{1/2} \\ &= \|\mathrm{Cov}(z_t^*)\|_2^{1/2} \cdot \|\mathrm{Cov}(\delta_t)\|_2^{1/2}, \end{aligned}$$

where (i) is due to Lemma 6. Hence, by Weyl's inequality,

$$\lambda_{\min}(\mathrm{Cov}(M^* h_t)) \geq \lambda_{\min}(\mathrm{Cov}(z_t^*)) - 2\|\mathrm{Cov}(z_t^*)\|_2^{1/2} \cdot \|\mathrm{Cov}(\delta_t)\|_2^{1/2}.$$

Since $\|\mathrm{Cov}(z_t^*)\|_2 = \mathcal{O}(1)$ due to the stability of A^* and $\delta_t = (\overline{A}^*)^H z_{t-H}^*$, there exists some dimension-free constant $a > 0$ such that as long as $H \geq \frac{\log(a\kappa(\overline{A}^*))}{\log(\rho(\overline{A}^*)^{-1})}$,

$$\lambda_{\min}(\mathrm{Cov}(M^* h_t)) \geq \sigma_u^2 \nu^2 / 2.$$

On the other hand,

$$\mathbb{E}[M^* h_t h_t^\top (M^*)^\top] \preccurlyeq \|\mathbb{E}[h_t h_t^\top]\|_2 M^* (M^*)^\top.$$

Since $h_t = [y_{(t-H+1):t}; u_{(t-H):(t-1)}]$ and $(\mathrm{Cov}(y_t))_{t\geq 0}$, $(\mathrm{Cov}(u_t))_{t\geq 0}$ have $\mathcal{O}(1)$ operator norms, by Lemma 7, $\|\mathrm{Cov}(h_t)\|_2 = \|\mathbb{E}[h_t h_t^\top]\|_2 = \mathcal{O}(H)$. Hence,

$$0 < \sigma_u^2 \nu^2 / 2 \leq \lambda_{\min}(\mathrm{Cov}(M^* h_t)) = \mathcal{O}(H) \sigma_{d_x}^2 (M^*).$$

Since $M^* \in \mathbb{R}^{d_x \times d_h}$, this implies that $\mathrm{rank}(M^*) = d_x$ and $\sigma_{\min}(M^*) = \Omega(\nu H^{-1/2})$. □

Proposition 2 is an adaption of Proposition 2 in (Tian et al., 2023) to the infinite-horizon LTI setting. Necessarily, this implies that by our choice of H, $d_h = H(d_y + d_u) \geq d_x$. Moreover, since $\overline{Q}^* \succcurlyeq \mu^2 I$, $N^* = (M^*)^\top \overline{Q}^* M^*$ is a $d_h \times d_h$ matrix with rank d_x, and $\lambda_{\min}^+(N^*) \geq \lambda_{\min}(\overline{Q}^*) \sigma_{\min}^2(M^*) = \Omega(\mu^2 \nu^2 H^{-1})$. Hence, if $\hat{N}$ is sufficiently close to N^*, by setting an appropriate threshold on the eigenvalues of $\hat{N}$, the dimension of the latent model equals the number of eigenvalues above it.

To find an approximate factorization of $\hat{N}$, let $\hat{N} = U \Lambda U^\top$ be its eigenvalue decomposition, where the diagonal elements of Λ are listed in descending order, and U is an orthogonal matrix. Let Λ_{d_x} be the top-left $d_x \times d_x$ block of Λ and U_{d_x} be the left d_x columns of U. By the Eckart-Young-Mirsky theorem, $\hat{M} = \max(\Lambda_{d_x}, 0)^{1/2} U_{d_x}^\top$, where "max" applies elementwise, is the solution to Line 4 of Algorithm 1, that is, the best approximate factorization of $\hat{N}$ among $d_x \times d_h$ matrices in terms of the Frobenius norm approximation error.

In the next two subsections, we move on to discussing the learning of latent dynamics, including the explicit approach SysId and the implicit approach CoSysId.

3.2 Explicit learning of system dynamics

Explicit learning of the system dynamics simply minimizes the *transition prediction error* in the latent space (Subramanian et al., 2020), or more generally, the statistical distances between the predicted and estimated distributions of the next latent state, like the KL divergence (Hafner et al., 2019a). In linear systems, it suffices to use the ordinary least squares as the SysId procedure, that is, to solve

$$(\hat{A}, \hat{B}) \in \operatorname*{argmin}_{A,B} \sum_{t=H}^{T+H-1} \|A\hat{z}_t + Bu_t - \hat{z}_{t+1}\|^2. \tag{3.4}$$

In this linear regression, if $(\hat{z}_t)_{t\geq H}$ are the optimal state estimates $(z_t^*)_{t\geq H}$ (2.9), then Simchowitz et al. (2018) has shown finite-sample guarantees for obtaining $(\hat{A}, \hat{B})$. Here, $\hat{z}_t$ contains errors resulting from the representation function $\hat{M}$ and the residual error δ_t in (2.10), but as long as T and H are large enough, SysId still has a finite-sample guarantee, as will be shown in Lemma 5. We refer to the algorithm that instantiates Algorithm 1 with SysId as CoReL-E (Cost-driven state Representation Learning). As the time-varying counterpart in (Tian et al., 2023), it provably solves LQG control without model-knowledge, as will be shown in Theorem 1.

3.3 Implicit learning of system dynamics (MuZero-style)

An important ingredient of latent model learning in MuZero (Schrittwieser et al., 2020) is to *implicitly* learn the transition function by minimizing the cost prediction error at *future latent states* generated from the transition function. Let $z_t = Mh_t$ denote the latent state given by representation function M at step t. Let $z_{t,0} = z_t$ and $z_{t,i} = Az_{t,i-1} + Bu_{t+i-1}$ for $i \geq 1$ be the future latent state predicted by dynamics (A, B) from z_t after i steps of transition. For a trajectory of length $T + H$ like (3.1), the loss that considers ℓ steps into the future is given by

$$\textstyle\sum_{t=H}^{T+H-K-1} \sum_{i=0}^{\ell} (\|z_{t,i}\|_Q^2 + \|u_t\|_{R^*}^2 + b - c_t)^2.$$

This loss involves powers of A up to A^ℓ; with the squared norm, the powers double, making the minimization over A hard to solve and analyze for $\ell \geq 2$. In LQG control, our finding is that it suffices to take $\ell = 1$. As mentioned in §1, MuZero also predicts optimal values and optimal actions; in LQG, to handle the case $Q^* \not\succ 0$, like cost-driven representation learning (see §3.1), we adopt the *cumulative costs* and use the normalized parameterization. Thus, the optimization problem we aim to solve is given by

$$\min_{M,A,B,b} \textstyle\sum_{t=H}^{T+H-1} \big((\|Mh_t\|^2 + b - \bar{c}_t)^2 + (\|AMh_t + Bu_t\|^2 + b - \bar{c}_{t+1})^2 \big). \tag{3.5}$$

To convexify the optimization problem (3.5), we define $N := M^\top M$ and $N_1 := [AM, B]^\top [AM, B]$. Then, (3.5) becomes

$$\min_{N,N_1,b} \textstyle\sum_{t=H}^{T+H-1} \big((\|h_t\|_N^2 + b - \bar{c}_t)^2 + (\|[h_t; u_t]\|_{N_1}^2 + b - \bar{c}_{t+1})^2 \big). \tag{3.6}$$

This minimization problem is convex in N, N_1, and b, and has a closed-form solution; essentially, it consists of two linear regression problems coupled by b. As a relaxation, we can decouple the two regression problems by allowing b to take different values in them; this works since the separate solutions for b in the two regression problems are close to each other for large T, and b is a term accounting for the estimation error, not part of the representation function. This decoupling further simplifies the analysis: the first regression problem is exactly cost-driven representation learning (§3.1), and the second is cost-driven system identification (CoSysId, Algorithm 2). The algorithm that instantiates Algorithm 1 with CoSysId is called CoReL-I (Cost-driven state Representation and Dynamic Learning). Like CoReL-E, this MuZero-style latent model learning method provably solves LQG control, as we will show in Theorem 1.

CoSysId has similar steps to cost-driven representation learning (§3.1), except that in Line 5 of Algorithm 2, it requires fitting a matrix $\hat{S}_0$. This is because the cost is *invariant* to the orthogonal transformations of latent states, and the approximate factorization steps recover

Algorithm 2 CoSysId: Cost-driven system identification

```
Input: data $\mathcal{D}_{\text{raw}}$, representation function $\hat{M}$
Estimate the system dynamics by
    $\hat{N}_1, \hat{b}_1 \in \operatorname*{argmin}_{N_1 = N_1^\top, b_1} \sum_{t=H}^{T+H-1} \left( \|[h_t; u_t]\|_{N_1}^2 + b_1 - \bar{c}_{t+1} \right)^2 \qquad (3.7)$
Find $\hat{M}_1 \in \operatorname{argmin}_{M_1 \in \mathbb{R}^{d_x \times (Hd_y + (H+1)d_u)}} \|M_1^\top M_1 - \hat{N}_1\|_F$
Split $\hat{M}_1$ to $[\widetilde{M}, \widetilde{B}]$ at column $H(d_y + d_u)$ and set $\widetilde{A} = \widetilde{M}\hat{M}^\dagger$.
Find alignment matrix $\hat{S}_0$ by
    $\hat{S}_0 \in \operatorname*{argmin}_{S_0 \in \mathbb{R}^{d_x \times d_x}} \sum_{t=H}^{T+H-1} \|S_0 \hat{M}_1 [h_t; u_t] - \hat{M} h_{t+1}\|^2 \qquad (3.8)$
Return: system dynamics estimate $(\hat{A}, \hat{B}) = (\hat{S}_0 \widetilde{A}, \hat{S}_0 \widetilde{B})$
```

M^* and M_1^* up to orthogonal transformations S and S_1, but there is no guarantee for the two transformations to be the same. MuZero bypasses this problem by predicting multiple steps of costs into the future, but analyzing such an optimization function involves the additional complexity of dealing with powers of A. Here, we instead estimate the $S_0 = SS_1^\top$ to align such two transformations. We note that although CoSysId needs the output $\hat{M}$ from cost-driven representation learning, the two quadratic regressions (3.2) and (3.7) are not coupled and can be solved in parallel.

Discussion on CoSysId. In CoSysId (Algorithm 2), the covariates of the quadratic regression in (3.7) are $([h_t; u_t])_{t \geq H}$. One may wonder if we can pursue an alternative approach by fixing M to be $\hat{M}$, and using $([\hat{z}_t; u_t])_{t \geq H}$ as covariates, which have a much lower dimension, though the two quadratic regressions cannot be solved in parallel anymore. Specifically, the new quadratic regression we need to solve is given by

$$\hat{N}_2, \hat{b}_2 \in \operatorname*{argmin}_{N_2 = N_2^\top, b_2} \sum_{t=H}^{T+H-1} \left(\|[\hat{z}_t; u_t]\|_{N_2}^2 + b_2 - \bar{c}_{t+1} \right)^2,$$

where $\hat{z}_t = \hat{M} h_t$ is an approximation of Sz_t^*. The ground truth for $\hat{N}_2$ is $N_2^* = [SA^*S^\top, SB^*]^\top [SA^*S^\top, SB^*]$, so its approximate factorization recovers $[S_2 A^* S^\top, S_2 B^*]$ for some orthogonal matrix S_2. In a similar way to CoSysId, we still need to fit an alignment matrix $S_3 = SS_2^\top$ to align the coordinates. Let $\widetilde{A}, \widetilde{B}$ denote the system parameters recovered from $\hat{N}_2$. The linear regression we now need to solve is from $([\widetilde{A}, \widetilde{B}][\hat{z}_t; u_t])_{t=H}^{T+H-1}$ to $(\hat{z}_{t+1})_{t=H}^{T+H-1}$. However, without further assumptions, $[A^*, B^*]$ does not necessarily have full row rank, and hence, neither does $[\widetilde{A}, \widetilde{B}]$, in which case recovering the entire S_3 is impossible.

On the other hand, for CoSysId (Algorithm 2), the ground truth of $\hat{M}_1$ is $M_1^* = [A^* M^*, B^*]$, which is guaranteed to have full row rank by the same argument as the proof of Proposition 2, since $M_1^*[h_t; u_t]$ estimates z_{t+1}^*, which has full-rank covariance. Hence, recovering $S_0 = SS_1^\top$ is feasible.

4 Theoretical guarantees

The following Theorem 1 shows that both CoReL-E and CoReL-I are guaranteed to solve unknown LQG control with a finite number of samples.

Theorem 1. *Given an unknown LQG control problem satisfying Assumption 1, let $M^{*\prime}$ and $(A^{*\prime}, B^{*\prime}, Q^{*\prime}, R^{*})$ be the optimal state representation function and the true system parameters under the normalized parameterization. For a given $p \in (0,1)$, if we run* CoReL-E *(Algorithm 1 with (3.4)) or* CoReL-I *(Algorithm 1 with Algorithm 2) for $T \geq \mathrm{poly}(d_x, d_y, d_u, \log(T/p))$, $H = \Omega(\log(HT))$, and $\sigma_u = \Theta(1)$, then there exists an orthogonal matrix $S \in \mathbb{R}^{d_x \times d_x}$, such that with probability at least $1-p$, the representation function $\hat{M}$ satisfies*

$$\|\hat{M} - SM^{*\prime}\|_2 = \mathcal{O}(\mathrm{poly}(H, d_x, d_u, d_y, \log(T/p))T^{-1/2}),$$

and the feedback gain $\hat{K}$ satisfies

$$J^{\hat{K}}(SA^{*\prime}S^{\top}, SB^{*\prime}, SQ^{*\prime}S^{\top}, R^{*}) - J^{*}(SA^{*\prime}S^{\top}, SB^{*\prime}, SQ^{*\prime}S^{\top}, R^{*})$$
$$= \mathcal{O}(\mathrm{poly}(H, d_x, d_u, d_y, \log(T/p))T^{-1}).$$

We defer the proof of Theorem 1 to §4.5. Compared with the time-varying setting in (Tian et al., 2023), the bounds here do not have a *separation* between the initial steps and future steps, where for the initial several steps, as the system has not been fully excited, the bounds were much worse. This is due to the fact that in the time-invariant setting, the representation function and the latent model are both stationary. On the other hand, to learn such stationary functions across different time steps, we need to aggregate correlated data along a single trajectory, which poses new significant challenges for the analysis. A major effort to overcome such difficulties involves proving a new result on the persistency of excitation (Lemma 1) using the small-ball method (Mendelson, 2015; Simchowitz et al., 2018), which will be discussed further in §4.2 with more details.

Compared with common system identification methods based on learning Markov parameters (Oymak and Ozay, 2019; Simchowitz et al., 2019), the error bounds of the system parameters produced by CoReL-I (or CoReL-E) have the same dependence on T, but worse dependence on system dimensions. Moreover, to establish persistency of excitation, CoReL-I (or CoReL-E) requires a larger burn-in period. These relative sample inefficiencies are the price we pay for cost-driven state representation learning, which is only supervised by *scalar-valued* costs that are *quadratic* in history, instead of *vector-valued* observations that are *linear* in history. Hence, we have to address the more challenging problem of *quadratic regression*, which lifts the dimension of the optimization problem. On the other hand, cost-driven state representation learning avoids learning the observation-reconstruction function C^{*}, and can learn task-relevant representations in more complex settings, as demonstrated by empirical studies.

4.1 Proposition on multi-step cumulative costs

The following proposition shows the relationship between $c_{t+\ell}$ and h_t, which is important for analyzing CoReL-E and CoReL-I. It is the counterpart of Proposition 3 in (Tian et al., 2023) in the LTI setting.

Proposition 3. *Given an LQG control problem satisfying Assumption 1, let M^* be the state representation function under the normalized parameterization. Let $\alpha = \max(\alpha(A^*), \alpha(\overline{A}^*))$ and $\rho = \max(\rho(A^*), \rho(\overline{A}^*))$. If we apply $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$, then for any $t \geq H$,*

$$\overline{c}_t := \sum\nolimits_{\tau=t}^{t+d_x-1} (c_\tau - \|u_\tau\|_{R^*}^2) = \|M^* h_t\|^2 + \overline{\delta}_t + \overline{b}^* + \overline{e}_t,$$

where $\delta_t = \mathcal{O}(\alpha^2 \rho^H \log(T/p))$ is a small error term, $\overline{b}^* = \mathcal{O}(d_x)$ is a positive constant, and $\overline{e}_t$ is a zero-mean subexponential random variable with $\|\overline{e}_t\|_{\psi_1} = \mathcal{O}(d_x^{3/2})$. Moreover, let $\overline{f}_t = [\mathrm{svec}(h_t h_t^\top); 1]$. As long as $H \geq \max(d_x - 1, \frac{a \log(\alpha T \log(T/p))}{\log(1/\rho)})$ for some problem-dependent constant $a > 0$, $(\overline{e}_t)_{t \geq H}$ satisfy

$$\Big\| \sum\nolimits_{t=H}^{T+H-1} \overline{f}_t \overline{e}_t \Big\| = \mathcal{O}(d_x^{3/2} d_h H^{1/2} T^{1/2} \log^{1/2}(1/p)).$$

Proof. Using $\Phi_{c,\ell}$ as a shorthand for $\Phi_{c,\ell}(A^*, B^*)$ below, by definition, we have

$$\begin{aligned} c_{t+\ell} - \|u_{t+\ell}\|_{R^*}^2 &= \|x_{t+\ell}\|_{Q^*}^2 \\ &= \Big\| (A^*)^\ell x_t + \Phi_{c,\ell}[u_{t+\ell-1}; \ldots; u_t] + \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \\ &\overset{(i)}{=} \|(A^*)^\ell x_t\|_{Q^*}^2 + \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2, \end{aligned}$$

where (i) is due to the independence of the three terms. Substituting $x_t = z_t^* + (x_t - z_t^*)$ and $z_t^* = M^* h_t + \delta_t$ into the above equation, we have

$$\begin{aligned} c_{t+\ell} - \|u_{t+\ell}\|_{R^*}^2 &= \|(A^*)^\ell z_t^*\|_{Q^*}^2 + \|(A^*)^\ell (x_t - z_t^*)\|_{Q^*}^2 + 2\langle (A^*)^\ell z_t^*, (A^*)^\ell (x_t - z_t^*) \rangle_{Q^*} \\ &\quad + \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \\ &= \|(A^*)^\ell (M^* h_t + \delta_t)\|_{Q^*}^2 + \|(A^*)^\ell (x_t - z_t^*)\|_{Q^*}^2 + 2\langle (A^*)^\ell z_t^*, (A^*)^\ell (x_t - z_t^*) \rangle_{Q^*} \\ &\quad + \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \\ &= \|(A^*)^\ell M^* h_t\|_{Q^*}^2 + \delta_{t,\ell} + b_\ell^* + e_{t,\ell}, \end{aligned}$$

where $\delta_{t,\ell} := \|(A^*)^\ell \delta_t\|_{Q^*}^2 + 2\langle (A^*)^\ell M^* h_t, (A^*)^\ell \delta_t \rangle_{Q^*}$ is a small term, and

$$\begin{aligned} b_\ell^* &:= \mathbb{E}\Big[\|(A^*)^\ell (x_t - z_t^*)\|_{Q^*}^2 + \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \Big], \\ e_{t,\ell} &:= \|(A^*)^\ell (x_t - z_t^*)\|_{Q^*}^2 + 2\langle (A^*)^\ell z_t^*, (A^*)^\ell (x_t - z_t^*) \rangle_{Q^*} \\ &\quad + \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum\nolimits_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 - b_\ell^*. \end{aligned}$$

Note that b_ℓ^* is not a function of time step t and $e_{t,\ell}$ is a zero-mean subexponential random variable with $\|e_{t,\ell}\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$. Define filtration

$$\mathcal{F}_t := \sigma(x_0, y_0, u_0, x_1, y_1, \ldots, u_{t-1}, x_t, y_t).$$

Then, $e_{t,\ell} \in \mathcal{F}_{t+\ell}$. Under the normalized parameterization, where $\sum_{\ell=0}^{d_x-1}((A^*)^\ell)^\top Q^*(A^*)^\ell = I$, we have

$$\bar{c}_t = \sum_{\tau=t}^{t+d_x-1}(c_\tau - \|u_\tau\|_{R^*}^2) = \|M^* h_t\|^2 + \bar{\delta}_t + \bar{b}^* + \bar{e}_t,$$

where

$$\bar{\delta}_t := \sum_{\ell=0}^{d_x-1} \delta_{t,\ell} = \|\delta_t\|^2 + 2\big\langle (A^*)^\ell M^* h_t, (A^*)^\ell \delta_t \big\rangle,$$

$$\bar{b}^* := \sum_{\ell=0}^{d_x-1} \bar{b}_\ell = \mathbb{E}\Big[\|x_t - z_t^*\|^2 + \sum_{\ell=0}^{d_x-1} \|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \Big],$$

$$\begin{aligned}\bar{e}_t := \sum_{\ell=0}^{d_x-1} e_{t,\ell} = \|x_t - z_t^*\|^2 + 2\langle z_t^*, x_t - z_t^* \rangle \\ + \sum_{\ell=0}^{d_x-1} \Big(\|\Phi_{c,\ell} u_{(t+\ell-1):t}\|_{Q^*}^2 + \Big\| \sum_{i=1}^{\ell} (A^*)^{i-1} w_{t+\ell-i} \Big\|_{Q^*}^2 \Big) - \bar{b}^*.\end{aligned}$$

Since $\delta_t = (\overline{A}^*)^H z_{t-H}^*$,

$$\|\bar{\delta}_t\| = \mathcal{O}(\alpha^2 \rho^{2H} \log(T/p) + \|A^*\|^{2\ell} \alpha \rho^H \log(T/p)) = \mathcal{O}(\alpha^2 \rho^H \log(T/p)).$$

Moreover, constant $\bar{b}^* = \mathcal{O}(d_x)$, $\bar{e}_t$ is a zero-mean subexponential random variable with $\|\bar{e}_t\|_{\psi_1} = \mathcal{O}(d_x^{3/2})$, and the random process $(\bar{e}_t)_{t \geq H}$ is adapted to filtration $(\mathcal{F}_{t+d_x-1})_{t \geq H}$.

However, the concentration of $\sum_{t=H}^{T+H-1} \overline{f}_t \bar{e}_t$ or even $\sum_{t=H}^{T+H-1} \bar{e}_t$ is highly nontrivial, in that $(\bar{e}_t)_{t \geq H}$ is not a martingale difference sequence. Below we develop the idea that random variables that are widely separated in a mixing stochastic process are nearly independent to show the concentration of $\sum_{t=H}^{T+H-1} \overline{f}_t \bar{e}_t$. Specifically, we partition the time steps into $\overline{H} = 2(H + d_x - 1) = \mathcal{O}(H)$ blocks. For partition $H \leq i < H + \overline{H}$, the indices are given by $(i + j\overline{H})_{j \geq 0}$. To obtain independent random variables $(g_{i+j\overline{H}})_{j \geq 0}$, we apply the Gram-Schmidt process to $(\overline{f}_{i+j\overline{H}} \bar{e}_{i+j\overline{H}})_{j \geq 0}$, which is adapted to $(\mathcal{F}_{i+j\overline{H}+d_x-1})_{j \geq 0}$. That is,

$$g_{i+(j+1)\overline{H}} = \overline{f}_{i+(j+1)\overline{H}} \bar{e}_{i+(j+1)\overline{H}} - \mathbb{E}[\overline{f}_{i+(j+1)\overline{H}} \bar{e}_{i+(j+1)\overline{H}} \mid \mathcal{F}_{i+j\overline{H}+d_x-1}].$$

Then,

$$\sum_{t=H}^{T+H-1} \overline{f}_t \bar{e}_t = \sum_{t=H}^{T+H-1} g_t + \sum_{t=H}^{T+H-1} \mathbb{E}[\overline{f}_{i+(j+1)\overline{H}} \bar{e}_{i+(j+1)\overline{H}} \mid \mathcal{F}_{i+j\overline{H}+d_x-1}]. \tag{4.1}$$

Since each dimension of $\overline{f}_t$ is subexponential with mean and the subexponential norm both bounded by $\mathcal{O}(1)$, each dimension of $g_{i+j\overline{H}}$ is $\frac{1}{2}$-sub-Weibull, with the sub-Weibull norm being $\mathcal{O}(d_x^{3/2})$. By applying sub-Weibull concentration (Hao et al., 2019, Theorem 3.1) to each of the $d_h(d_h+1)/2$ dimensions of $(g_{i+j\overline{H}})_{j \geq 0}$, we have

$$\Big\| \sum_{j \geq 0} g_{i+j\overline{H}} \Big\| = \mathcal{O}(d_x^{3/2} d_h (T/\overline{H})^{1/2} \log^{1/2}(1/p))$$

Repeating the above argument for each $H \leq i < H + \overline{H}$, we have

$$\Big\| \sum_{t=H}^{T+H-1} g_t \Big\| = \mathcal{O}(d_x^{3/2} d_h \overline{H} (T/\overline{H})^{1/2} \log^{1/2}(1/p)) = \mathcal{O}(d_x^{3/2} d_h H^{1/2} T^{1/2} \log^{1/2}(1/p)). \tag{4.2}$$

It remains to bound the residuals of the Gram-Schmidt process. To this end, we first express $\overline{f}_{t+\overline{H}}$ and $\overline{e}_{t+\overline{H}}$ into two parts, one adapted to and the other independent of $\mathcal{F}_{t+d_x-1}$. By definition,

$$\begin{aligned} y_{t+k} &= C^*((A^*)^k x_t + \Phi_{c,k} u_{(t+k-1):t} + \textstyle\sum_{i=1}^k (A^*)^{i-1} w_{t+k-i}) + v_{t+k} \\ &= C^*(A^*)^k x_t + \xi^y_{t,k}, \end{aligned}$$

where $\xi^y_{t,k}$ is independent of $\mathcal{F}_t$, and a zero-mean Gaussian random vector with the operator norm of the covariance matrix being $\mathcal{O}(1)$ due to the stability of A^*. Recall that $f_t = \operatorname{svec}(h_t h_t^\top)$ and $h_t = [u_{(t-H):(t-1)}; y_{(t-H+1):t}]$. Let $h_{t+\overline{H}} = s_{t+\overline{H}} + \xi^h_{t+\overline{H}}$, where

$$\begin{aligned} s_{t+\overline{H}} &= [0; C^*(A^*)^{\overline{H}-H-d_x+2} x_{t+d_x-1}; \dots; C^*(A^*)^{\overline{H}-d_x+1} x_{t+d_x-1}], \\ \xi^h_{t+\overline{H}} &= [u_{(t+\overline{H}-H):(t+\overline{H}-1)}; \xi^y_{t+d_x-1,\overline{H}-H-d_x+2}; \dots; \xi^y_{t+d_x-1,\overline{H}-d_x+1}], \end{aligned}$$

and $\xi^h_{t+\overline{H}}$ is independent of $\mathcal{F}_{t+d_x-1}$, and a zero-mean Gaussian random vector with the variance of each dimension bounded by $\mathcal{O}(1)$. Then,

$$\begin{aligned} \overline{f}_{t+H} &= \operatorname{svec}(h_{t+H} h_{t+H}^\top) \\ &= \operatorname{svec}(s_{t+H} s_{t+H} + s_{t+H}(\xi^h_{t+H})^\top + \xi^h_{t+H} s_{t+H}^\top + \xi^h_{t+H}(\xi^h_{t+H})^\top). \end{aligned} \tag{4.3}$$

We now turn to $\overline{e}_t$. Since

$$\begin{aligned} x_{t+1} - z^*_{t+1} &= A^* x_t + B^* u_t + w_t - (A^* z^*_t + B^* u_t + L^*(y_{t+1} - C^*(A^* z^*_t + B^* u_t))) \\ &= A^*(x_t - z^*_t) + w_t - L^*(C^*(A^* x_t + B^* u_t + w_t) + v_{t+1} - C^*(A^* z^*_t + B^* u_t)) \\ &= \overline{A}^*(x_t - z^*_t) + (I - L^* C^*) w_t + v_{t+1}, \end{aligned}$$

we have

$$x_{t+\overline{H}} - z^*_{t+\overline{H}} = (\overline{A}^*)^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \delta^x_{t+\overline{H}},$$

where $\delta^x_{t+\overline{H}}$ is independent of $\mathcal{F}_{t+d_x-1}$, and a zero-mean Gaussian random vector with the operator norm of the covariance matrix bounded by $\mathcal{O}(1)$ due to the stability of $\overline{A}^*$. Since

$$\begin{aligned} z^*_{t+1} &= \overline{A}^* z^*_t + \overline{B}^* u_t + L^* y_{t+1} \\ &= \overline{A}^* z^*_t + \overline{B}^* u_t + L^*(C^*(A^* x_t + B^* u_t + w_t) + v_{t+1}) \\ &= A^* z^*_t + L^* C^* A^*(x_t - z^*_t) + B^* u_t + L^* C^* w_t + L^* v_{t+1}, \end{aligned}$$

we have

$$z^*_{t+\overline{H}} = (A^*)^{\overline{H}-d_x+1} z^*_{t+d_x-1} + \Phi_A(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^z_{t+\overline{H}},$$

where $\Phi_A := \sum_{i=1}^{\overline{H}-d_x+1} (A^*)^{\overline{H}-d_x+1-i} L^* C^* A^* (\overline{A}^*)^{i-1}$ and $\xi^z_{t+\overline{H}}$ is independent of $\mathcal{F}_{t+d_x-1}$, and a zero-mean Gaussian random vector with the operator norm of the covariance matrix bounded by $\mathcal{O}(1)$ due to the stability of A^* and $\overline{A}^*$. Hence, $\overline{e}_{t+\overline{H}}$ can be expressed as

$$\begin{aligned}\overline{e}_{t+\overline{H}} = {} & \|(\overline{A})^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\|^2 \\ & + 2\big\langle (A^*)^{\overline{H}-d_x+1} z^*_{t+d_x-1} + \Phi_A(x_{t+d_x-1} - z^*_{t+d_x-1}) \\ & + \xi^z_{t+\overline{H}}, (\overline{A})^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\big\rangle \\ & + \xi^e_{t+\overline{H}} - \mathbb{E}\Big[\|(\overline{A})^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\|^2\Big],\end{aligned} \tag{4.4}$$

where

$$\begin{aligned}\xi^e_{t+\overline{H}} := {} & \textstyle\sum_{\ell=0}^{d_x-1} \Big(\|\Phi_{c,\ell} u_{(t+\overline{H}+\ell-1):(t+\overline{H})}\|^2_{Q^*} + \Big\| \sum_{i=1}^{\ell} (A^*)^{i-1} w_{t+\overline{H}+\ell-i}\Big\|^2_{Q^*}\Big) \\ & - \mathbb{E}\Big[\textstyle\sum_{\ell=0}^{d_x-1} \Big(\|\Phi_{c,\ell} u_{(t+\overline{H}+\ell-1):(t+\overline{H})}\|^2_{Q^*} + \Big\| \sum_{i=1}^{\ell} (A^*)^{i-1} w_{t+\overline{H}+\ell-i}\Big\|^2_{Q^*}\Big)\Big]\end{aligned}$$

is independent of $\mathcal{F}_{t+d_x-1}$, and a zero-mean subexponential random variable with $\mathcal{O}(d_x^{3/2})$ subexponential norm due to the stability of A^*.

Notice that

$$\mathbb{E}[\overline{f}_t \overline{e}_t] = \mathbb{E}[\overline{f}_t \mathbb{E}[\overline{e}_t \mid \overline{f}_t]] = 0.$$

Then, by substituting (4.3) and (4.4), we have

$$\begin{aligned} & \mathbb{E}[\overline{f}_{t+\overline{H}} \overline{e}_{t+\overline{H}} \mid \mathcal{F}_{t+d_x-1}] \\ & = \mathbb{E}[\overline{f}_{t+\overline{H}} \overline{e}_{t+\overline{H}} \mid \mathcal{F}_{t+d_x-1}] - \mathbb{E}[\overline{f}_{t+\overline{H}} \overline{e}_{t+\overline{H}}] \\ & = \mathbb{E}\Big[\Big(\mathrm{svec}(s_{t+\overline{H}} s_{t+\overline{H}} + s_{t+\overline{H}}(\xi^h_{t+\overline{H}})^\top + \xi^h_{t+\overline{H}} s^\top_{t+\overline{H}} + \xi^h_{t+\overline{H}}(\xi^h_{t+\overline{H}})^\top)\Big) \\ & \quad \cdot \Big(\|(\overline{A})^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\|^2 + 2\big\langle (A^*)^{\overline{H}-d_x+1} z^*_{t+d_x-1} \\ & \quad + \Phi_A(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^z_{t+\overline{H}}, (\overline{A})^{\overline{H}-d_x+1} \cdot (x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\big\rangle \\ & \quad + \xi^e_{t+\overline{H}}\Big) \mid \mathcal{F}_{t+d_x-1}\Big] - \mathbb{E}\Big[\Big(\mathrm{svec}(s_{t+\overline{H}} s_{t+\overline{H}} + s_{t+\overline{H}}(\xi^h_{t+\overline{H}})^\top + \xi^h_{t+\overline{H}} s^\top_{t+\overline{H}} + \xi^h_{t+\overline{H}}(\xi^h_{t+\overline{H}})^\top)\Big) \\ & \quad \cdot \Big(\|(\overline{A})^{\overline{H}-d_x+1}(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\|^2 + 2\big\langle (A^*)^{\overline{H}-d_x+1} z^*_{t+d_x-1} \\ & \quad + \Phi_A(x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^z_{t+\overline{H}}, (\overline{A})^{\overline{H}-d_x+1} \cdot (x_{t+d_x-1} - z^*_{t+d_x-1}) + \xi^x_{t+\overline{H}}\big\rangle + \xi^e_{t+\overline{H}}\Big)\Big],\end{aligned}$$

where the terms completely independent of $\mathcal{F}_{t+d_x-1}$ cancel each other, and all other terms contain at least one of $(A^*)^{\overline{H}-H-d_x+2}$, $(\overline{A}^*)^{\overline{H}-d_x+1}$ and Φ_A, with each of the $d_h(d_h+1)/2$ dimensions being the product of two subexponential random variables and two Gaussian random variables. Hence, with probability at least $1-p$,

$$\begin{aligned}\|\mathbb{E}[\overline{f}_{t+\overline{H}} \overline{e}_{t+\overline{H}} \mid \mathcal{F}_{t+d_x-1}]\| & = \mathcal{O}(\alpha d_x^{3/2} d_h \rho^{\overline{H}-H-d_x+2} \log^3(T/p)) \\ & = \mathcal{O}(\alpha d_x^{3/2} d_h \rho^{H} \log^3(T/p)).\end{aligned} \tag{4.5}$$

Finally, combining (4.1) with the bounds in (4.2) and (4.5), we have

$$\|\textstyle\sum_{t=H}^{T+H-1} \overline{f}_t \overline{e}_t\| = \mathcal{O}(d_x^{3/2} d_h (H^{1/2} T^{1/2} \log^{1/2}(1/p) + \alpha \rho^H T \log^3(T/p))).$$

As long as $H \geq \frac{a \log(\alpha T \log(T/p))}{\log(1/\rho)}$ for some problem-dependent constant $a > 0$,

$$\| \textstyle\sum_{t=H}^{T+H-1} \overline{f}_t \overline{e}_t \| = \mathcal{O}(d_x^{3/2} d_h H^{1/2} T^{1/2} \log^{1/2}(1/p)).$$

□

4.2 Persistency of excitation

Central to the analysis of CoReL-E and CoReL-I is the finite-sample characterization of the *quadratic regression* problem. To solve (3.2), notice that

$$\|h_t\|_N^2 = \langle N, h_t h_t^\top \rangle_F = \langle \mathrm{svec}(N), \mathrm{svec}(h_t h_t^\top) \rangle,$$

so this quadratic regression is essentially a linear regression problem in terms of $[\mathrm{svec}(N); b_0]$. A major difficulty in the analysis is to establish persistency of excitation for $([\mathrm{svec}(h_t h_t^\top); 1])_{t \geq H}$, meaning that the minimum eigenvalue of the Gram matrix $\sum_{t=H}^{T+H-1} [\mathrm{svec}(h_t h_t^\top); 1][\mathrm{svec}(h_t h_t^\top)^\top, 1]$ grows linearly in the size T of the data. This is needed to ensure the uniqueness and convergence of the parameter estimation.

A linear lower bound on $\lambda_{\min}(\sum_{t=H}^{T+H-1} h_t h_t^\top)$ is a known result for the identification of partially observable linear dynamical systems, see the recent overview in (Tsiamis et al., 2022). In our case, however, elements of $\mathrm{svec}(h_t h_t^\top)$ are *products* of Gaussians, making the analysis difficult. If $(h_t)_{t \geq H}$ are independent, which is the case if they are from multiple independent trajectories, the result has been established in (Jadbabaie et al., 2021) and (Tian et al., 2023). It can also be proved with the matrix Azuma inequality (Tropp, 2012). Here, by contrast, we need to aggregate correlated data to estimate a set of *stationary* parameters. In sum, the difficulty we face results from both products of Gaussians and the data dependence.

In principle, given enough burn-in time, state x_t, and hence observation y_t and truncated history h_t, converge to the steady-state distributions, and samples with an interval of the order of mixing time are approximately independent (Levin and Peres, 2017). Hence, intuitively, a linear lower bound is viable. However, the bound yielded by such an analysis deteriorates as the system is less stable and the mixing time increases, which is qualitatively incorrect for linear systems. To eschew such dependence, (Simchowitz et al., 2018) introduces the so-called *small-ball* method. We take the same route, while establishing different arguments to handle the products of Gaussians.

Let us first recall the block martingale small-ball condition (Simchowitz et al., 2018, Definition 2.1).

Definition 1 (Block martingale small-ball (BMSB) condition)**.** *Let $(f_t)_{t \geq 1}$ be a stochastic process in $\mathbb{R}^d$ adapted to filtration $(\mathcal{F}_t)_{t \geq 1}$. We say $(f_t)_{t \geq 1}$ satisfies the (k, Γ, q)-BMSB condition for $k \in \mathbb{N}^+$, $\Gamma \succ 0$ and $q > 0$, if for any $t \geq 1$, for any fixed unit vector $v \in \mathbb{R}^d$, $\frac{1}{k} \sum_{i=1}^{k} \mathbb{P}(| \langle f_{t+i}, v \rangle | \geq \|v\|_\Gamma \mid \mathcal{F}_t) \geq q$ almost surely.*

The key Lemma 1 below shows that $([\mathrm{svec}(h_t h_t^\top); 1])_{t \geq H}$ satisfies the BMSB condition.

Lemma 1. *Let $h_t = [y_{(t-H+1):t}; u_{(t-H):(t-1)}]$ be the H-step history at time step $t \geq H$ in system (2.1) with $u_t \sim \mathcal{N}(0, \sigma_u^2 I)$ for $t \geq 0$. Define filtration $\mathcal{F}_t := \sigma(x_0, y_0, u_0, x_1, y_1, \ldots, u_{t-1}, x_t, y_t)$. Define $f_t := \operatorname{svec}(h_t h_t^\top)$ and $\overline{f}_t := [f_t; 1]$, adapted to $(\mathcal{F}_t)_{t \geq H}$. Recall that for square matrix A, $\alpha(A) := \sup_{k \geq 0} \|(A)^k\|_2 \rho(A)^{-k}$. As long as $H \geq \frac{a_1 \log(d_h \alpha(A^*) \log(T/p))}{\log(\rho(A^*)^{-1})}$ for some dimension-free constant a_1, $(\overline{f}_t)_{t \geq H}$ is $(k, \gamma^2 I, q)$-BMSB for $k = 4H$, $\gamma = \Theta(d_h^{-3/2})$, and $q = \Theta(d_h^{-3})$, where $\Theta(\cdot)$ hides the dependence on dimension-free constants.*

Proof. Since svec is a bijection, every vector $w \in \mathbb{R}^{d_h(d_h+1)/2}$ corresponds to a symmetric matrix $D \in \mathbb{R}^{d_h \times d_h}$ with Frobenius norm $\|w\|$. Then, for any unit vector $v = [w; s]$ with $w \in \mathbb{R}^{d_h(d_h+1)/2}$ and $s \in \mathbb{R}$,

$$\left\langle \overline{f}_{t+i}, v \right\rangle = \langle f_{t+i}, w \rangle + s = \langle \operatorname{svec}(h_{t+i} h_{t+i}^\top), \operatorname{svec}(D) \rangle + s = h_{t+i}^\top D h_{t+i} + s.$$

Take $\Gamma = \gamma^2 I$ for some $\gamma > 0$ to be specified later. Then, $\|v\|_\Gamma = \gamma$. It suffices to show that for $i > \overline{H}$ for some $\overline{H} > 0$,

$$\mathbb{P}(|h_{t+i}^\top D h_{t+i} + s| \geq \gamma \mid \mathcal{F}_t) \geq q,$$

since if so, we have

$$\frac{1}{2\overline{H}} \sum_{i=1}^{2\overline{H}} \mathbb{P}(|h_{t+i}^\top D h_{t+i} + s| \geq \gamma \mid \mathcal{F}_t) \geq \frac{1}{2\overline{H}} \sum_{i=\overline{H}+1}^{2\overline{H}} \mathbb{P}(|h_{t+i}^\top D h_{t+i} + s| \geq \gamma \mid \mathcal{F}_t) \geq q/2,$$

which means $(\overline{f}_t)_{t \geq H}$ is $(2\overline{H}, \gamma^2 I, q/2)$-BMSB.

Now let us take a close look at

$$h_{t+i} = [y_{(t+i-H+1):(t+i)}; u_{(t+i-H):(t+i-1)}].$$

Since

$$y_{t+i} = C^*(A^*)^i x_t + \sum_{j=1}^i C^*(A^*)^j (B^* u_{t+i-j} + w_{t+i-j}) + v_{t+i},$$

$y_{t+i} \mid \mathcal{F}_t$ is Gaussian with mean $C^*(A^*)^i x_t$ and covariance determined by $\sum_{j=1}^i C^*(A^*)^j (B^* u_{t+i-j} + w_{t+i-j}) + v_{t+i}$, where we note that v_{t+i} is independent of all other random variables and has full-rank covariance. Hence, for $i \geq H$, $h_{t+i} \mid \mathcal{F}_t$ is Gaussian and has full-rank covariance. Then intuitively, since $\|D\|_F = 1$, $|h_{t+i}^\top D h_{t+i}| \mid \mathcal{F}_t$ is a well-behaved random variable that can exceed some $\gamma > 0$ with a positive probability q. Formally, let $\mu_{t,i} := \mathbb{E}[h_{t+i} \mid \mathcal{F}_t]$. By Lemma 3, for $i \geq H$, there exists some absolute constant $a > 0$, such that

$$\mathbb{E}[|(h_{t+i} - \mu_{t,i})^\top D (h_{t+i} - \mu_{t,i}) + s| \mid \mathcal{F}_t] \geq a \min\{\sigma_u, \sigma_v\} d_h^{-3/2}.$$

By triangle inequality, we have

$$\begin{aligned} |(h_{t+i} - \mu_{t,i})^\top D (h_{t+i} - \mu_{t,i}) + s| &= |h_{t+i}^\top D h_{t+i} + \mu_{t,i}^\top D \mu_{t,i} - 2 h_{t+i}^\top D \mu_{t,i} + s| \\ &\leq |h_{t+i}^\top D h_{t+i} + s| + |\mu_{t,i}^\top D \mu_{t,i}| + 2 |h_{t+i}^\top D \mu_{t,i}|. \end{aligned}$$

Hence,

$$\mathbb{E}[|h_{t+i}^\top D h_{t+i} + s| \mid \mathcal{F}_t] \ge a \min\{\sigma_u, \sigma_v\} d_h^{-3/2} - \mathbb{E}[|\mu_{t,i}^\top D \mu_{t,i}| + 2|h_{t+i}^\top D \mu_{t,i}| \mid \mathcal{F}_t].$$

Now we argue that for large enough i, $\mathbb{E}[|\mu_{t,i}^\top D \mu_{t,i}| + 2|h_{t+i}^\top D \mu_{t,i}|]$ is negligible. Since matrix A^* is stable, with probability at least $1-p$, $\|x_t\| = \mathcal{O}(d_x^{1/2} \log(T/p))$ for all $t \ge 0$. Hence,

$$\|C^*(A^*)^i x_t\| = \mathcal{O}(\alpha(A^*)\rho(A^*)^i d_x^{1/2} \log(T/p)),$$

where we recall that $\alpha(A^*) := \sup_{k\ge 0} \|(A^*)^k\|_2 \rho(A^*)^{-k}$ and $\|C^*\|_2$, $\|A^*\|_2$ are hidden in $\mathcal{O}(\cdot)$. Then, for $i \ge H$,

$$\begin{aligned}\mathbb{E}[|\mu_{t,i}^\top D \mu_{t,i}| + 2|h_{t+i}^\top D \mu_{t,i}| \mid \mathcal{F}_t] &= |\langle \mu_{t,i}\mu_{t,i}^\top, D\rangle_F| + 2\mathbb{E}[|\langle \mu_{t,i} h_{t+i}^\top, D\rangle_F| \mid \mathcal{F}_t] \\ &\le \|\mu_{t,i}\mu_{t,i}^\top\|_F \cdot \|D\|_F + 2\mathbb{E}[\|\mu_{t,i} h_{t+i}^\top\|_F \cdot \|D\|_F \mid \mathcal{F}_t] \\ &= \|\mu_{t,i}\|^2 + 2\|\mu_{t,i}\| \cdot \mathbb{E}[\|h_{t+i}\| \mid \mathcal{F}_t].\end{aligned}$$

By definition, $\mu_{t,i}$ is the concatenation of $(C^*(A^*)^j x_t)_{i-H+1 \le j \le i}$ and zero vectors. Hence,

$$\|\mu_{t,i}\| = \mathcal{O}(d_h^{1/2} \alpha(A^*) \rho(A^*)^i \log(T/p)).$$

Choose $H \ge \frac{a_1 \log(d_h \alpha(A^*) \log(T/p))}{\log(\rho(A^*)^{-1})}$ for some dimension-free constant $a_1 > 0$, such that for $i > 2H$, we have

$$\|\mu_{t,i}\|^2 + 2\|\mu_{t,i}\| \cdot \mathbb{E}[\|h_{t+i}\| \mid \mathcal{F}_t] \le a \min\{\sigma_u, \sigma_v\} d_h^{-3/2}/2.$$

Then, we have the desired lower bound that

$$\mathbb{E}[|h_{t+i}^\top D h_{t+i} + s| \mid \mathcal{F}_t] \ge a \min\{\sigma_u, \sigma_v\} d_h^{-3/2}/2.$$

On the other hand, since

$$|h_{t+i}^\top D h_{t+i} + s| = |\langle D, h_{t+i} h_{t+i}^\top\rangle_F + s| \le \|D\|_F \|h_{t+i} h_{t+i}^\top\|_F + |s| \le h_{t+i}^\top h_{t+i} + |s|,$$

we have $\mathbb{E}[|h_{t+i}^\top D h_{t+i} + s|^2 \mid \mathcal{F}_t] \le 2\mathbb{E}[\|h_{t+i}\|^4 \mid \mathcal{F}_t] + 2s^2$. Since $\|h_{t+i}\| \mid \mathcal{F}_t$ is sub-Gaussian with

$$\| \|h_{t+i}\| \mid \mathcal{F}_t\|_{\psi_2} = \mathcal{O}(\|\mathbb{E}[h_{t+i} h_{t+i}^\top \mid \mathcal{F}_t]\|_2^{1/2}) = \mathcal{O}(1),$$

$\mathbb{E}[|h_{t+i}^\top D h_{t+i} + s|^2 \mid \mathcal{F}_t] = \mathcal{O}(1)$. By the Paley-Zygmund inequality, for $\beta \in [0,1]$ we have

$$\mathbb{P}(|h_{t+i}^\top D h_{t+i} + s| \ge \beta a \min\{\sigma_u, \sigma_v\} d_h^{-3/2}/2 \mid \mathcal{F}_t) = \Omega((1-\beta)^2 a^2 d_h^{-3}),$$

where the dependence on σ_u, σ_v is hidden in $\Omega(\cdot)$. By taking $\beta = 1/2$, we can see that $(f_t)_{t \ge H}$ satisfies the $(k, \gamma^2 I, q)$-BMSB condition for $k = 4H$, $\gamma = \Theta(d_h^{-3/2})$ and $q = \Theta(d_h^{-3})$. □

Crucial for its proof is Lemma 3, a lower bound on the expectation of Gaussian quadratic forms, which might be of independent interest. Then with Lemma 1, following the analysis in (Simchowitz et al., 2018, Appendix D), we can show that if additionally, T is large enough, then with high probability,

$$\lambda_{\min}\Big(\sum_{t=H}^{T+H-1} f_t f_t^\top\Big) = \Omega(\gamma^2 q^2 T) = \Omega(d_h^{-9} T),$$

which establishes the persistency of excitation.

Lower bound for Gaussian quadratic forms.

Lemma 2. *Let $z_1, z_2, \ldots, z_d$ be independent standard Gaussian random variables. Let $v = [v_1, v_2, \ldots, v_{d+1}]^\top \in \mathbb{S}^d$ be a $(d+1)$-dimensional unit vector. Then,*

$$\inf_{v \in \mathbb{S}^d} \mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] \geq 0.8 d^{-3/2}.$$

Proof. Let us consider the value of v_{d+1}. Since $\mathbb{E}[z_i^2] = 1$ for all $1 \leq i \leq d$, we have

$$\mathbb{E}\Big[\Big|\sum_{i=1}^d v_i z_i^2\Big|\Big] \leq \sum_{i=1}^d |v_i| \leq \sqrt{d \sum_{i=1}^d v_i^2} \leq \sqrt{d(1 - v_{d+1}^2)}.$$

Then,

$$\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] \geq |v_{d+1}| - \mathbb{E}\Big[\Big|\sum_{i=1}^d v_i z_i^2\Big|\Big] \geq |v_{d+1}| - \sqrt{d(1 - v_{d+1}^2)}.$$

Hence, if $|v_{d+1}| \geq 2\sqrt{d/(4d+1)}$, we have $\sqrt{d(1 - v_{d+1}^2)} \leq |v_{d+1}|/2$. It follows that

$$\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] \geq \frac{|v_{d+1}|}{2} \geq \sqrt{\frac{d}{4d+1}} \geq \frac{1}{\sqrt{5}}.$$

Below we consider the case where $|v_{d+1}| < 2\sqrt{d/(4d+1)}$. Let $\mathrm{sign}(\cdot)$ denote the sign function. Let $\mathcal{I}^+ := \{i : \mathrm{sign}(v_i) = 1,\ 1 \leq i \leq d\}$ and $\mathcal{I}^- := \{i : \mathrm{sign}(v_i) = -1,\ 1 \leq i \leq d\}$ be the index sets of positive and negative values among $(v_i)_{i=1}^d$. Then,

$$\begin{aligned}\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] &= \mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d |v_i|\,\mathrm{sign}(v_i) z_i^2\Big|\Big] \\ &= \mathbb{E}\Big[\Big|v_{d+1} + \sum_{i\in\mathcal{I}^+} |v_i| z_i^2 - \sum_{j\in\mathcal{I}^-} |v_j| z_j^2\Big|\Big].\end{aligned}$$

For a given v, since $(z_i^2)_{i=1}^d$ have identical distributions, $\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i\in\mathcal{I}^+} |v_i| z_i^2 - \sum_{j\in\mathcal{I}^-} |v_j| z_j^2\Big|\Big]$ has the same value under permutations of $(v_i)_{i\in\mathcal{I}^+}$ and $(v_j)_{j\in\mathcal{I}^-}$. Summing over all the permutations of $(v_i)_{i\in\mathcal{I}^+}$ and $(v_j)_{j\in\mathcal{I}^-}$ gives

$$\begin{aligned}&d\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] \\ &\geq \mathbb{E}\Big[\Big|d \cdot v_{d+1} + \big(\sum_{i\in\mathcal{I}^+} |v_i|\big) \sum_{i\in\mathcal{I}^+} z_i^2 - \big(\sum_{j\in\mathcal{I}^-} |v_j|\big) \sum_{j\in\mathcal{I}^-} z_j^2\Big|\Big].\end{aligned}$$

Hence,

$$\mathbb{E}\Big[\Big|v_{d+1} + \sum_{i=1}^d v_i z_i^2\Big|\Big] \geq \frac{1}{d}\Big(\sum_{i=1}^d |v_i|\Big)\mathbb{E}\Big[\Big|d \cdot v_{d+1} + \sum_{i=1}^d \mathrm{sign}(v_i) z_i^2\Big|\Big]$$

Since $\sum_{i=1}^d |v_i| \geq (\sum_{i=1}^d v_i^2)^{1/2} = (1 - v_{d+1}^2)^{1/2}$, we have

$$\mathbb{E}\big[\big|v_{d+1} + \textstyle\sum_{i=1}^d v_i z_i^2\big|\big] \geq \frac{(1 - v_{d+1}^2)^{1/2}}{d} \inf_{w \in \{\pm 1\}^d} \mathbb{E}\big[\big|d \cdot v_{d+1} + \textstyle\sum_{i=1}^d w_i z_i^2\big|\big].$$

It remains to lower bound $\inf_{w \in \{\pm 1\}^d} \mathbb{E}\big[\big|d \cdot v_{d+1} + \sum_{i=1}^d w_i z_i^2\big|\big]$. By symmetry, for any pair $w_i \neq w_j$, the expectation remains the same if we interchange z_i and z_j. Hence, for any random variable x,

$$\mathbb{E}[|x + z_i - z_j|] = \frac{1}{2}(\mathbb{E}[|x + z_i - z_j|] + \mathbb{E}[|x + z_i - z_j|]) \geq \mathbb{E}[|x|].$$

We shall apply this symmetry trick in the following to cancel terms with opposite signs. Let p denote the number of $+1$'s and q denote the number of -1's in w, such that $p + q = n$. If $p \neq q$, by the symmetry trick,

$$\mathbb{E}\big[\big|d \cdot v_{d+1} + \textstyle\sum_{i=1}^d w_i z_i^2\big|\big] \geq \mathbb{E}\big[\big|d \cdot v_{d+1} + \textstyle\sum_{i=1}^{|p-q|} z_i^2\big|\big] \geq \mathrm{Var}\big(\textstyle\sum_{i=1}^{|p-q|} z_i^2\big) = 2|p - q| \geq 2.$$

If $p = q$, again, the symmetry trick yields

$$\mathbb{E}[|d \cdot v_{d+1} + \textstyle\sum_{i=1}^d w_i z_i^2|] \geq \mathbb{E}[|d \cdot v_{d+1} + z_1^2 - z_2^2|] \geq \mathrm{Var}\big(z_1^2 - z_2^2\big) = 4.$$

Hence, regardless of p and q, we have $\inf_{w \in \{\pm 1\}^d} \mathbb{E}[|d \cdot v_{d+1} + \sum_{i=1}^d w_i z_i^2|] \geq 2$. Then,

$$\mathbb{E}\Big[\Big|v_{d+1} + \textstyle\sum_{i=1}^d v_i z_i^2\Big|\Big] \geq 2 \cdot \frac{(1 - v_{d+1}^2)^{1/2}}{d}.$$

Since $|v_{d+1}| < 2\sqrt{d/(4d+1)}$,

$$\mathbb{E}\Big[\Big|v_{d+1} + \textstyle\sum_{i=1}^d v_i z_i^2\Big|\Big] = 2 \cdot \frac{1}{\sqrt{4d+1} \cdot d} = 0.8 d^{-3/2}.$$

Hence, overall we have

$$\inf_{v \in \mathbb{S}^d} \mathbb{E}\Big[\Big|v_{d+1} + \textstyle\sum_{i=1}^d v_i z_i^2\Big|\Big] \geq 0.8 d^{-3/2}.$$

□

From the proof, we can see that without v_{d+1}, we have the improved bound $\inf_{v \in \mathbb{S}^{d-1}} \mathbb{E}[|\sum_{i=1}^d v_i z_i^2|] \geq 2d^{-1}$.

Based on Lemma 2, we can prove the more general Lemma 3 below.

Lemma 3. *Let x be a d-dimensional zero-mean Gaussian random vector with covariance Σ. For any $d \times d$ symmetric matrix A and constant $b \in \mathbb{R}$ that satisfy $\|A\|_F^2 + b^2 = 1$, there exists an absolute constant $a > 0$, such that $\mathbb{E}[|x^\top A x + b|] \geq a \lambda_{\min}(\Sigma) d^{-3/2}$.*

Proof. Let $y := \Sigma^{-1/2}x$. Then y is a standard Gaussian random vector, and $x^\top Ax = y^\top \Sigma^{1/2} A \Sigma^{1/2} y$. Let $U^\top \Lambda U$ be the eigenvalue decomposition of $\Sigma^{1/2}A\Sigma^{1/2}$. Then,

$$x^\top Ax = y^\top U^\top \Lambda U y = z^\top \Lambda z,$$

where $z := Uy$ is still a standard Gaussian random vector.

By the unitary invariance of the Frobenius norm,

$$\|\Lambda\|_F = \|U^\top \Lambda U\|_F = \|\Sigma^{1/2}A\Sigma^{1/2}\|_F \geq \lambda_{\min}(\Sigma)\|A\|_F.$$

Hence,

$$\|\Lambda\|_F^2 + b^2 \geq \lambda_{\min}^2(\Sigma)\|A\|_F^2 + b^2 \geq \lambda_{\min}^2(\Sigma) \wedge 1.$$

Therefore, by Lemma 2, there exists an absolute constant $a > 0$, such that

$$\inf_{\|A\|_F^2+b^2=1} \mathbb{E}[|x^\top Ax + b|] \geq \inf_{\|\Lambda\|_F^2+b^2 \geq \lambda_{\min}^2(\Sigma)\wedge 1} \mathbb{E}[|z^\top \Lambda z + b|] \geq a(\lambda_{\min}(\Sigma) \wedge 1)d^{-3/2}.$$

□

4.3 Quadratic regression bound

The following quadratic regression bound is at the core of proving Theorem 1. Its proof builds on a new persistency of excitation result (Lemma 1). We retain $(e_t)_{t\geq 1}$ in the bound, as in our problem $(e_t)_{t\geq 1}$ may not correspond to a martingale, and may contain an additional small error term resulting from using $M^* h_t$ to approximate z_t^*. For notational convenience, we note that the $h_t, c_t, \mathcal{F}_t$ in Lemma 4 and its proof slightly abuse the notation, which use different variables from the rest of the paper. Hence, the indices start with $t = 1$, rather than $t = H$ in the CoReL-E and CoReL-I algorithms.

Lemma 4. *Let $(h_t^*)_{t\geq 1}$ be a sequence of d-dimensional Gaussian random vectors adapting to filtration $(\mathcal{F}_t)_{t\geq 1}$ with $\|\mathbb{E}[h_t^*(h_t^*)^\top]\|_2^{1/2} \leq \sigma$. Define random variable $c_t = (h_t^*)^\top N^* h_t^* + b^* + e_t$, where $N^* \in \mathbb{R}^{d\times d}$ is a positive semidefinite matrix and $b^* \in \mathbb{R}$ is a constant. Assume σ and $\|N^*\|_2$ are $\mathcal{O}(1)$. Define $h_t = h_t^* + \delta_t$, where the perturbation vector δ_t can be correlated with h_t^* and its ℓ_2 norm is sub-Gaussian with $\mathbb{E}[\|\delta_t\|] \leq \epsilon$, $\|\|\delta_t\|\|_{\psi_2} \leq \epsilon$. Define $f_t^* := \operatorname{svec}(h_t^*(h_t^*)^\top)$ and $\overline{f}_t^* := [f_t^*; 1]$. Assume that $(\overline{f}_t^*)_{t\geq 1}$ satisfies $(k, \gamma^2 I, q)$-BMSB condition and $\epsilon \leq \min(\sigma d^{1/2}, a_0\gamma\sigma^{-1}d^{-1}\log^{-2}(T/p))$ for some absolute constant $a_0 > 0$. Consider*

$$(\hat{N}, \hat{b}) \in \operatorname*{argmin}_{N=N^\top, b} \textstyle\sum_{t=1}^T (c_t - \|h_t\|_N^2 - b)^2. \tag{4.6}$$

Then, as long as $T \geq a_1 k d^2 q^{-2} \log(d/(\gamma q p))$ for some dimension-free constant $a_1 > 0$, we have that with probability at least $1-p$,

$$\begin{aligned}\|\hat{N} - N^*\|_F = \mathcal{O}\Big(&\epsilon(\gamma q)^{-1} d^{1/2} \log(T/p) \\ &+ \epsilon(\gamma q)^{-2} d^{1/2} T^{-1} \log(T/p) \textstyle\sum_{t=1}^T \|e_t\| + (\gamma q)^{-2} T^{-1} \Big\| \textstyle\sum_{t=1}^T \overline{f}_t^* e_t \Big\|\Big),\end{aligned}$$

where σ and $\|N^*\|_2$ are problem-dependent constants hidden in $\mathcal{O}(\cdot)$.

Proof. Regression (4.6) can be written as

$$\operatorname*{argmin}_{\mathrm{svec}(N),b} \sum\nolimits_{t=1}^{T} \left(c_t - \mathrm{svec}(h_t h_t^\top)^\top \mathrm{svec}(N) - b\right)^2.$$

Define $f_t := \mathrm{svec}(h_t h_t^\top)$ and $\overline{f}_t := [f_t; 1]$. It is a linear regression problem with extended covariates $\overline{f}_t$, which can be further rewritten as

$$\operatorname*{argmin}_{\mathrm{svec}(N),b} \sum\nolimits_{t=1}^{T} \left(c_t - \overline{f}_t^\top [\mathrm{svec}(N); b]\right)^2. \tag{4.7}$$

Let $\overline{F} := [\overline{f}_1, \overline{f}_2, \ldots, \overline{f}_T]^\top$ be the $T \times \frac{d(d+1)}{2}$ matrix whose tth row is $f_t^\top$. Define $\overline{F}^*$ similarly by replacing $\overline{f}_t$ by $\overline{f}_t^*$. Solving linear regression (4.7) gives

$$\overline{F}^\top \overline{F}[\mathrm{svec}(\hat{N}); \hat{b}] = \sum\nolimits_{t=1}^{T} \overline{f}_t c_t.$$

Substituting $c_t = (\overline{f}_t^*)^\top [\mathrm{svec} N^*; b^*] + e_t$ into the above equation yields

$$\overline{F}^\top \overline{F}[\mathrm{svec}(\hat{N}); \hat{b}] = \overline{F}^\top \overline{F}^* [\mathrm{svec}(N^*); b^*] + \overline{F}^\top \xi,$$

where ξ denotes the vector whose tth element is e_t. Rearranging the terms, we have

$$\overline{F}^\top \overline{F}[\mathrm{svec}(\hat{N} - N^*); \hat{b} - b^*] = \overline{F}^\top (\overline{F}^* - \overline{F})[\mathrm{svec}(N^*); b^*] + \overline{F}^\top \xi. \tag{4.8}$$

Next, we show that $\lambda_{\min}(\overline{F}^\top \overline{F}) = \Omega(\gamma^2 q^2 T)$, which we achieve by showing $(\overline{f}_t)_{t\geq 1}$ satisfies the BMSB condition. By our assumption, $(\overline{f}_t^*)_{t\geq 1}$ satisfies $(k, \gamma^2 I, q)$-BMSB condition, meaning that for any fixed unit vector $v \in \mathbb{R}^{\frac{d(d+1)}{2}+1}$, it holds almost surely that

$$\frac{1}{k} \sum\nolimits_{i=1}^{k} \mathbb{P}\left(|\langle \overline{f}_{t+i}^*, v\rangle| \geq \gamma \mid \mathcal{F}_t\right) \geq q.$$

For any fixed unit vector $v \in \mathbb{R}^{\frac{d(d+1)}{2}+1}$, we have

$$|\langle \overline{f}_t, v\rangle| = |\langle \overline{f}_t^*, v\rangle + \left\langle \overline{f}_t - \overline{f}_t^*, v\right\rangle| \geq |\langle \overline{f}_t^*, v\rangle| - |\langle \overline{f}_t - \overline{f}_t^*, v\rangle| \geq |\langle \overline{f}_t^*, v\rangle| - \|\overline{f}_t - \overline{f}_t^*\|.$$

To bound $\|\overline{f}_t - \overline{f}_t^*\| = \|f_t - f_t^*\| = \|h_t h_t^\top - h_t^*(h_t^*)^\top\|_F$, we have

$$\|h_t^*(h_t^*)^\top\|_F \overset{(i)}{\leq} 2\|h_t^*(h_t^*)^\top - h_t h_t^\top\|_2 = 2\|h_t^*(h_t^* - h_t)^\top + (h_t^* - h_t)h_t^\top\|_2 \leq 2(\|h_t^*\| + \|h_t\|)\|\delta\|,$$

where (i) follows from the fact that matrix $h_t^*(h_t^*)^\top - h_t h_t^\top$ has at most rank two. Since h_t^* is Gaussian with $\|\mathbb{E}[h_t^*(h_t^*)^\top]\|_2^{1/2} \leq \sigma$, $\|h^*\|$ is sub-Gaussian with its mean and sub-Gaussian norm bounded by $\mathcal{O}(\sigma d^{1/2})$. Since $\|\delta\|$ is sub-Gaussian with its mean and sub-Gaussian norm bounded by $\epsilon \leq \sigma d^{1/2}$, we conclude that $\|h^*(h^*)^\top - hh^\top\|_2$ is subexponential with its mean and subexponential norm bounded by $\mathcal{O}(\epsilon\sigma d^{1/2})$. Hence, with probability at least $1 - p$,

$$\|h^*(h^*)^\top - hh^\top\|_2 = \mathcal{O}(\epsilon\sigma d^{1/2} \log(T/p)).$$

Then, for all $1 \le t \le T$, since $\|\overline{f}_t - \overline{f}_t^*\| = \mathcal{O}(\epsilon\sigma d^{1/2}\log(T/p))$, there exists an absolute constant $a_0 > 0$, such that as long as $\epsilon \le \frac{a_0\gamma}{\sigma d^{1/2}\log(T/p)}$, $\|\overline{f}_t - \overline{f}_t^*\| \le \gamma/2$. It follows that

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{P}(|\left\langle \overline{f}_{t+i}, v\right\rangle| \ge \gamma/2 \mid \mathcal{F}_t) \ge \frac{1}{k}\sum_{i=1}^{k} \mathbb{P}(|\left\langle \overline{f}_{t+i}^*, v\right\rangle| \ge \gamma \mid \mathcal{F}_t) \ge q,$$

which means that $(\overline{f}_t)_{1\le t\le T}$ is $(k, \gamma^2 I/4, q)$-BMSB. Following the analysis in (Simchowitz et al., 2018, Appendix D), by lower bounding $\inf_{v:\|v\|=1}\sum_{t=1}^{T}\langle v, f_t\rangle^2$ using a covering argument (Simchowitz et al., 2018, Lemma 4.1), we can show that for a given $p \in (0,1)$, as long as $T \ge a_1 k d^2 q^{-2}\log(d/(\gamma q p))$ for some absolute constant $a_1 > 0$, then with probability at least $1-p$, we have

$$\lambda_{\min}\left(\sum_{t=1}^{T}\overline{f}_t\overline{f}_t^\top\right) = \Omega(\gamma^2 q^2 T).$$

Hence, we have $\lambda_{\min}(\overline{F}^\top\overline{F}) = \Omega(\gamma^2 q^2 T)$.

Now we return to (4.8). By inverting $\overline{F}^\top\overline{F}$, we obtain

$$\begin{aligned}\|[\mathrm{svec}(\hat{N} - N^*); \hat{b} - b^*]\| &= \|\overline{F}^\dagger(\overline{F}^* - \overline{F})[\mathrm{svec}(N^*); b^*] + \overline{F}^\dagger\xi\| \\ &\le \underbrace{\|\overline{F}^\dagger(\overline{F}^* - \overline{F})[\mathrm{svec}(N^*); b^*]\|}_{(a)} + \underbrace{\|\overline{F}^\dagger\xi\|}_{(b)}.\end{aligned} \tag{4.9}$$

Term (a) is upper bounded by

$$\begin{aligned}\sigma_{\min}(\overline{F})^{-1}\|(\overline{F}^* - \overline{F})[\mathrm{svec}(N^*); b^*]\| &= \mathcal{O}(\sigma_{\min}(\overline{F})^{-1})\|(F^* - F)\mathrm{svec}(N^*)\| \\ &= \mathcal{O}((\gamma q)^{-1}T^{-1/2})\|(F^* - F)\mathrm{svec}(N^*)\|.\end{aligned}$$

Using arguments similar to those in (Mhammedi et al., 2020, Section B.2.13), we have

$$\begin{aligned}\|(F^* - F)\mathrm{svec}(N^*)\|^2 &= \sum_{t=1}^{T}\left\langle \mathrm{svec}(h_t^*(h_t^*)^\top) - \mathrm{svec}(h_t h_t^\top), \mathrm{svec}(N^*)\right\rangle^2 \\ &= \sum_{t=1}^{T}\left\langle h_t^*(h_t^*)^\top - h_t h_t^\top, N^*\right\rangle_F^2 \\ &\le \|N^*\|_2^2\sum_{t=1}^{T}\|h_t^*(h_t^*)^\top - h_t h_t^\top\|_*^2 \\ &\overset{(i)}{\le} 4\|N^*\|_2^2\sum_{t=1}^{T}\|h_t^*(h_t^*)^\top - h_t h_t^\top\|_2^2,\end{aligned}$$

where (i) follows from the fact that the matrix $h_t^*(h_t^*)^\top - h_t h_t^\top$ has at most rank two. Since

$$\|h^*(h^*)^\top - hh^\top\|_2^2 = \mathcal{O}(\epsilon^2\sigma^2 d\|\log^2(T/p)),$$

term (a) in (4.9) is bounded by

$$\mathcal{O}\big((\gamma q)^{-1}T^{-1/2}\epsilon\sigma\|N^*\|_2 d^{1/2}T^{1/2}\log(T/p)\big) = \mathcal{O}((\gamma q)^{-1}d^{1/2}\epsilon\log(T/p)).$$

Now we consider term (b) in (4.9):

$$(b) = \|\overline{F}^\dagger\xi\| \le \lambda_{\min}(\overline{F}^\top\overline{F})^{-1}\|\overline{F}^\top\xi\| = \mathcal{O}((\gamma q)^{-2}T^{-1})\Big\|\sum_{t=1}^{T}\overline{f}_t e_t\Big\|.$$

Since

$$\Big\| \sum_{t=1}^T \overline{f}_t e_t \Big\| \le \Big\| \sum_{t=1}^T \overline{f}_t^* e_t \Big\| + \sum_{t=1}^T \|\overline{f}_t - \overline{f}_t^*\| \|e_t\|.$$

we have

$$(b) = \mathcal{O}\Big((\gamma q)^{-2} T^{-1} \Big\| \sum_{t=1}^T \overline{f}_t^* e_t \Big\| + \epsilon (\gamma q)^{-2} d^{1/2} T^{-1} \log(T/p) \sum_{t=1}^T \|e_t\| \Big).$$

Combining the bounds on (a) and (b), we show that as long as $T \ge a_1 k d^2 q^{-2} \log(d/(\gamma q p))$, with probability at least $1 - p$,

$$\begin{aligned}&\|[\mathrm{svec}(\hat{N} - N^*); \hat{b} - b^*]\| \\ =\ &\mathcal{O}(\epsilon (\gamma q)^{-1} d^{1/2} \log(T/p) + \epsilon (\gamma q)^{-2} d^{1/2} T^{-1} \log(T/p) \sum_{t=1}^T \|e_t\| + (\gamma q)^{-2} T^{-1} \Big\| \sum_{t=1}^T \overline{f}_t e_t \Big\|).\end{aligned}$$

□

4.4 Perturbed linear regression bound

Identifying the time-invariant latent dynamics involves linear regression with *correlated data* and *perturbed measurements*. The following Lemma 5 extends the previous linear system identification result in (Simchowitz et al., 2018) to the case with noises in both input and output variables. In Lemma 5, γ and q are treated as dimension-free constants (in contrast to Lemma 4), which is indeed the case in our application of Lemma 5 to $(z_t^*)_{t \ge H}$ in analyzing SysId (3.4) for CoReL-E in §4.5.1, and in analyzing alignment matrix estimation (3.8) in Algorithm 2 for CoReL-I in §4.5.2. Note that the bound in Lemma 5 is worse than that in the time-varying setting in (Tian et al., 2023), due to the treatment of correlated data.

Lemma 5. *Let $(x_t^*)_{t \ge 1}$ be a sequence of d_1-dimensional Gaussian random vectors adapted to a filtration $(\mathcal{F}_t)_{t \ge 1}$ with $\|\mathbb{E}[x_t^* (x_t^*)^\top]\|_2^{1/2} \le \sigma$. Define $y_t^* = A^* x_t^* + e_t$, where $A^* \in \mathbb{R}^{d_2 \times d_1}$ and $e_t \mid \mathcal{F}_t$ is Gaussian with zero mean and $\|\mathbb{E}[e_t e_t^\top]\|_2^{1/2} \le \epsilon$. Define $y_t = y_t^* + \delta_t^y$ and $x_t = x_t^* + \delta_t^x$, where the perturbation vectors δ_t^x and δ_t^y can be correlated with x_t^* and y_t^*, and their ℓ_2 norms are sub-Gaussian with $\mathbb{E}[\|\delta_t^x\|] \le \epsilon_x$, $\|\|\delta_t^x\|\|_{\psi_2} \le \epsilon_x$ and $\mathbb{E}[\|\delta_t^y\|] \le \epsilon_y$, $\|\|\delta_t^y\|\|_{\psi_2} \le \epsilon_y$. Assume that $(x_t^*)_{t \ge 1}$ satisfies the $(k, \gamma^2 I, q)$-BMSB condition, $\epsilon_x \le a_0 \gamma^2 q^2 / \sigma$ for some absolute constant $a_0 > 0$, $\sigma, \epsilon, \epsilon_x, \epsilon_y$ are $\mathcal{O}(1)$, and k, γ, q are $\Theta(1)$. Consider*

$$\hat{A} \in \operatorname*{argmin}_{A \in \mathbb{R}^{d_2 \times d_1}} \sum_{t=1}^T \|y_t - A x_t\|^2. \tag{4.10}$$

Then, as long as $T \ge a_1 k q^{-2} (\log(1/p) + d_1 \log(10/q) + d_1 \log(\sigma \gamma^{-1} d_1 \log(T/p)))$ for some absolute constant $a_1 > 0$, we have that with probability at least $1 - p$,

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

Proof. Let $X \in \mathbb{R}^{T \times d_1}$ denote the matrix whose tth row is $x_t^\top$. Define $X^*, Y, E, \Delta_x, \Delta_y$ similarly. To solve the regression problem, we set its gradient to be zero and substitute in $Y = X^* (A^*)^\top + E + \Delta_y$ to obtain

$$\hat{A}(X^\top X) = A^* (X^*)^\top X + E^\top X + \Delta_y^\top X. \tag{4.11}$$

Substituting in $X = X^* + \Delta_x$ gives

$$\begin{aligned}&(\hat{A} - A^*)((X^*)^\top X^*)\\ &= A^*(X^*)^\top \Delta_x - \hat{A}(\Delta_x^\top \Delta_x + \Delta_x^\top X^* + (X^*)^\top \Delta_x) + E^\top X^* + E^\top \Delta_x + \Delta_y^\top X^* + \Delta_y^\top \Delta_x.\end{aligned} \tag{4.12}$$

Now we deal with each term on the right-hand side. Since $(X^*)^\top \Delta_x = \sum_{t=1}^T x_t^* (\delta_t^x)^\top$, by the triangle inequality,

$$\|(X^*)^\top \Delta_x\|_2 \leq \sum_{t=1}^T \|x_t^*(\delta_t^x)^\top\|_2 \leq \sum_{t=1}^T \|x_t^*\| \cdot \|\delta_t^x\|.$$

Since $(x_t^*)_{t\geq 1}$ are Gaussian, with probability at least $1-p$, $\|x_t^*\| = \mathcal{O}(\sigma d_1^{1/2} \log^{1/2}(T/p))$. Since $(\|\delta_t^x\|)_{t\geq 1}$ are sub-Gaussian with $\mathbb{E}[\|\delta_t^x\|] \leq \epsilon_x$ and $\|\|\delta_t^x\|\|_{\psi_2} \leq \epsilon_x$, with probability at least $1-p$, $\|\delta_t^x\| = \mathcal{O}(\epsilon_x \log^{1/2}(T/p))$. Hence,

$$\|(X^*)^\top \Delta_x\|_2 = \mathcal{O}(\epsilon_x \sigma d_1^{1/2} T \log(T/p)).$$

Similarly, with probability at least $1-p$,

$$\begin{aligned}\|\Delta_x^\top \Delta_x\|_2 = \mathcal{O}(\epsilon_x^2 T \log(T/p)), \quad \|E^\top \Delta_x\|_2 = \mathcal{O}(\epsilon \epsilon_x d_2^{1/2} T \log(T/p)),\\ \|\Delta_y^\top X^*\|_2 = \mathcal{O}(\epsilon_y \sigma d_1^{1/2} T \log(T/p)), \quad \|\Delta_y^\top \Delta_x\|_2 = \mathcal{O}(\epsilon_x \epsilon_y T \log(T/p)).\end{aligned}$$

It remains to bound $\|\hat{A}\|_2$. Notice that with probability at least $1-p$,

$$\|(X^*)^\top X^*\|_2 \leq \sum_{t=1}^T \|x_t^*\|^2 = \mathcal{O}(\sigma^2 d_1 \log(T/p) T).$$

Let $T_0 := a_1 k q^{-2}(\log(1/p) + d_1 \log(10/q) + d_1 \log(\sigma\gamma^{-1} d_1 \log(T/p)))$ for some absolute constant $a_1 > 0$. Then, by (Simchowitz et al., 2018, Appendix D), as long as $T \geq T_0$, with probability at least $1-p$, $\lambda_{\min}((X^*)^\top X^*) = \gamma^2 q^2 T/32$. Since $X^\top X = (X^*)^\top X^* + \Delta_x^\top X^* + (X^*)^\top \Delta_x + \Delta_x^\top \Delta_x$,

$$\lambda_{\min}(X^\top X) \geq \lambda_{\min}((X^*)^\top X^*) - \|\Delta_x^\top X^* + (X^*)^\top \Delta_x + \Delta_x^\top \Delta_x\|_2.$$

Hence, there exists an absolute constant $a_0 > 0$, such that as long as $\epsilon_x \leq a_0 \gamma^2 q^2/\sigma$, $\lambda_{\min}(X^\top X) = \Omega(\gamma^2 q^2 T)$, which implies $\|X^\dagger\|_2 = \mathcal{O}(\gamma^{-1} q^{-1} T^{-1/2})$. From (4.11), we have

$$\|\hat{A}\|_2 = (\|A^*\|_2 \|X^*\|_2 + \|E\|_2 + \|\Delta_y\|_2)\|X^\dagger\|_2 = \mathcal{O}(\gamma^{-1} q^{-1}(\sigma \|A^*\|_2 + \epsilon + \epsilon_y)) = \mathcal{O}(1),$$

where in the last equality we treat $\gamma, q, \sigma, \epsilon, \epsilon_y$ as problem-dependent constants.

Finally, by (Simchowitz et al., 2018, Theorem 2.4), as long as $T \geq T_0$,

$$\|E^\top (X^*)^\dagger\|_2 = \mathcal{O}((d_2 + d_1 \log(d_1 \log(T/p)) + \log(1/p))^{1/2} T^{-1/2}).$$

Combining all the above individual bounds with the terms on the right-hand side of (4.12), we have

$$\|\hat{A} - A^*\|_2 = \mathcal{O}((\epsilon_x + \epsilon_y) d_1^{1/2} \log(T/p) + (d_2 + d_1 \log(d_1 \log(T/p)) + \log(1/p))^{1/2} T^{-1/2}),$$

which completes the proof. □

4.5 Proof of the main results

In this section, we prove the sample complexity bounds for CoReL-E and CoReL-I in Theorem 1. As we shall see, the proofs for the two algorithms share similar ideas and tools. We start with the same analysis for both algorithms, and split into separate paragraphs as the analysis diverges.

By Proposition 3,

$$\bar{c}_t := \sum_{\tau=t}^{t+d_x-1}(c_\tau - \|u_\tau\|_{R^*}^2) = \|M^* h_t\|^2 + \bar{\delta}_t + \bar{b}^* + \bar{e}_t,$$

where $\delta_t = \mathcal{O}(\alpha^2\rho^H\log(T/p))$ is a small error term, $\bar{b}^* = \mathcal{O}(d_x)$ is a positive constant, and $\bar{e}_t$ is a zero-mean subexponential random variable with $\|\bar{e}_t\|_{\psi_1} = \mathcal{O}(d_x^{3/2})$. Recall that we define $N^* := (M^*)^\top M^*$, $f_t := \mathrm{svec}(h_t h_t^\top)$, $\bar{f}_t := [\mathrm{svec}(h_t h_t^\top); 1]$, and $d_h := H(d_y + d_u)$. Rewriting the above equation, we have

$$\bar{c}_t = \bar{f}_t^\top[\mathrm{svec}(N^*); \bar{b}^*] + \bar{\delta}_t + \bar{e}_t, \tag{4.13}$$

By Lemma 1, $\bar{f}_t$ is $(k, \gamma^2 I, q)$-BMSB for $k = 4H$, $\gamma = \Theta(d_h^{-3/2})$ and $q = \Theta(d_h^{-3})$. Then, by Lemma 4, $\hat{N}$ obtained by solving regression (3.2) has the guarantee that there exists some absolute constant $a_0 > 0$, such that as long as $T \geq a_0 H^9 (d_y + d_u)^8 \log(H(d_y + d_u)/p)$, with probability at least $1 - p$,

$$\|\hat{N} - N^*\|_F = \mathcal{O}\Big((\gamma q)^{-2} T^{-1} \Big\| \sum_{t=1}^{T} \bar{f}_t(\bar{\delta}_t + \bar{e}_t)\Big\|\Big).$$

By Proposition 3, as long as $H \geq \frac{a_1 \log(\alpha T \log(T/p))}{\log(1/\rho)}$ for some problem-dependent constant $a_1 > 0$,

$$\|\sum_{t=H}^{T+H-1} \bar{f}_t \bar{e}_t\| = \mathcal{O}(d_x^{3/2} d_h H^{1/2} T^{1/2} \log^{1/2}(1/p)).$$

Since $\|\sum_{t=H}^{T+H-1} \bar{f}_t \bar{\delta}_t\| = \mathcal{O}(\alpha^2 d_h \rho^H \log(T/p))$, we have

$$\|\hat{N} - N^*\|_F = \mathcal{O}\Big((\gamma q)^{-2} T^{-1} \big(d_x^{3/2} d_h H^{1/2} T^{1/2} \log^{1/2}(1/p) + \alpha^2 d_h \rho^H \log(T/p)\big)\Big).$$

As long as $H \geq \frac{a_2 \log(\alpha T \log(T/p))}{\log(1/\rho)}$ for some problem-dependent constant $a_2 > 0$, we have

$$\|\hat{N} - N^*\|_F = \mathcal{O}(H^{21/2} d_x^{3/2} (d_y + d_u)^{10} T^{-1/2} \log^{1/2}(1/p)).$$

By (Tu et al., 2016, Lemma 5.4), there exists an orthogonal matrix S, such that $\|\hat{M} - SM^*\|_F$ is of the same order of $\|\hat{N} - N^*\|_F$. To understand the approximation error $\hat{z}_t - Sz_t^*$, recall that $z_t^* = M^* h_t + \delta_t$, where $\delta_t = (\bar{A}^*)^H z_{t-H}^*$. Then,

$$\|\hat{z}_t - Sz_t^*\| = \|(\hat{M} - SM^*)h_t - S\delta_t\| \leq \|\hat{M} - SM^*\|_2 \|h_t\| + \|\delta_t\|.$$

Since $\|h_t\|$ is sub-Gaussian with $\mathbb{E}[\|h_t\|] = \mathcal{O}(d_h^{1/2})$, $\|\|h_t\|\|_{\psi_2} = \mathcal{O}(d_h^{1/2})$, we have $\|\hat{M} - SM^*\|_2 \|h_t\|$ is sub-Gaussian with its mean and sub-Gaussian norm bounded by

$$\mathcal{O}(H^{11} d_x^{3/2} (d_y + d_u)^{21/2} T^{-1/2} \log^{1/2}(1/p)). \tag{4.14}$$

Notice that $\|\delta_t\|$ is sub-Gaussian with mean and sub-Gaussian norm bounded by $\mathcal{O}(\alpha(\overline{A}^*)\rho(\overline{A}^*)^H d_x^{1/2})$, which, by our choice of H, is dominated by (4.14). Hence, for all $t \geq H$, $\|\hat{z}_t - Sz_t^*\|$ is sub-Gaussian with its mean and sub-Gaussian norm bounded by (4.14).

The latent cost is also described in Proposition 1, given by

$$c_t = \|z_t^*\|_{Q^*}^2 + \|u_t\|_{R^*}^2 + b^* + e_t,$$

where $b^* := \mathbb{E}[\|x_t - z_t^*\|_{Q^*}^2]$, $e_t := \|x_t - z_t\|_{Q^*}^2 + 2\langle z_t^*, x_t - z_t^* \rangle_{Q^*}$ is a zero-mean subexponential random variable with $\|e_t\|_{\psi_1} = \mathcal{O}(d_x^{1/2})$, and the random process $(e_t)_{t \geq H}$ is adapted to filtration $(\mathcal{F}_t)_{t \geq H}$. In a similar way to the analysis for $([\mathrm{svec}(h_t h_t^\top); 1])_{t \geq H}$, $([\mathrm{svec}(z_t^*(z_t^*)^\top); 1])_{t \geq H}$ satisfies the $(\Theta(1), \Theta(d_x^{-3/2}), \Theta(d_x^{-3}))$-BMSB condition. By the perturbed quadratic regression bound (Lemma 4), $\widetilde{Q}$ from regression (3.3) has the guarantee that

$$\begin{aligned} \|\widetilde{Q} - SQ^*S^\top\|_F = \mathcal{O}\Big(&H^{11} d_x^{3/2} (d_y + d_u)^{21/2} T^{-1/2} \log^{1/2}(1/p) \\ &\cdot (d_x^{9/2} d_x^{1/2} \log(T/p) + d_x^9 d_x^{1/2} \log(T/p) T^{-1} \textstyle\sum_{t=1}^{T} \|e_t\|) \\ &+ d_x^9 T^{-1} \Big\| \textstyle\sum_{t=H}^{T+H-1} [\mathrm{svec}(\hat{z}_t \hat{z}_t^\top); 1] e_t \Big\|\Big). \end{aligned}$$

With probability at least $1 - p$, $\|e_t\| = \mathcal{O}(d_x^{1/2} \log(T/p))$. By similar analysis to that for $\sum_{t=H}^{T+H-1} \overline{f}_t \overline{e}_t$ in the proof of Proposition 3, we have

$$\textstyle\sum_{t=H}^{T+H-1} [\mathrm{svec}(z_t^*(z_t^*)^\top); 1] e_t = \mathcal{O}(d_x^{1/2} d_x H^{1/2} T^{1/2} \log^{1/2}(1/p)).$$

Since $Q^* \succcurlyeq 0$ and $\hat{Q}$ is the projection of $\widetilde{Q}$ onto positive semidefinite matrices, we have

$$\|\hat{Q} - SQ^*S^\top\|_F \leq \|\widetilde{Q} - SQ^*S^\top\|_F = \mathcal{O}(H^{11} d_x^{23/2} (d_y + d_u)^{21/2} T^{-1/2} \log^{5/2}(T/p)).$$

4.5.1 Remaining proof of Theorem 1 for CoReL-E

We proceed to analyze system identification. The latent dynamics is described in Proposition 1, given by

$$z_{t+1}^* = A^* z_t^* + B^* u_t + L^* i_{t+1},$$

To apply the perturbed linear regression bound (Lemma 5), the noise term $L^* i_{t+1} \mid \mathcal{F}_t$ needs to be a zero-mean Gaussian, which does not hold, since

$$\begin{aligned} i_{t+1} &= y_{t+1} - C^*(A^* z_t^* + B^* u_t) \\ &= C^*((A^* x_t + B^* u_t + w_t) + v_{t+1}) - C^*(A^* z_t^* + B^* u_t) \\ &= C^* A^* (x_t - z_t^*) + C^* w_t + v_{t+1}, \end{aligned}$$

where $x_t - z_t^* \in \mathcal{F}_t$. To solve this problem, we consider a different filtration $(\mathcal{G}_t := \sigma(y_0, u_0, y_1, \ldots, u_{t-1}, y_t))_{t \geq 0}$ that involves only observations and actions. Then, $z_t^* \in \mathcal{G}_t$ and $L^* i_{t+1} \mid \mathcal{F}_t$ is a zero-mean Gaussian with the operator norm of the covariance matrix bounded by $\mathcal{O}(1)$. With this filtration, by the perturbed linear regression bound (Lemma 5), for T greater than a constant polynomial in the problem parameters, we have

$$\begin{aligned}\|[\hat{A},\hat{B}]-S[A^*S^\top,B^*]\|_2 &= \mathcal{O}(H^{11}d_x^{3/2}(d_y+d_u)^{21/2}T^{-1/2}\log^{1/2}(1/p)\cdot d_x^{1/2}\log(T/p))\\ &= \mathcal{O}(H^{11}d_x^{2}(d_y+d_u)^{21/2}T^{-1/2}\log^{3/2}(T/p)).\end{aligned}$$

Hence, $\|\hat{A}-SA^*S^\top\|_2$, $\|\hat{B}-SB^*\|_2$ and $\|\hat{Q}-SQ^*S^\top\|_2$ are all bounded by

$$\mathcal{O}(H^{11}d_x^{23/2}(d_y+d_u)^{21/2}T^{-1/2}\log^{3/2}(T/p)).$$

4.5.2 Remaining proof of Theorem 1 for CoReL-I

We proceed to analyze cost-driven system identification (Algorithm 2). Define $M_1^* := [A^*M^*, B^*]$ as the composition of one-step transition and representation functions and $N_1^* := (M_1^*)^\top M_1^*$, which is estimated by $\hat{N}_1$ in (3.7).

By the same analysis as that of $\hat{N}$, we have

$$\|\hat{N}_1-N_1^*\|_F=\mathcal{O}(H^{1/2}(H(d_y+d_u)+d_x)^{10}d_x^{3/2}T^{-1/2}\log^{1/2}(1/p)).$$

By (Tu et al., 2016, Lemma 5.4), there exists orthogonal matrices S_1, such that $\|\hat{M}_1-S_1M_1^*\|_F$ is on the same order of $\|\hat{N}_1-N_1^*\|_F$. The bound on $\|\hat{M}_1-S_1M_1^*\|_F$ applies to $\|\widetilde{B}-S_1B^*\|_2$ as well. Since $[S_1A^*S^\top SM^*, S_1B^*]=S_1M_1^*$, by the perturbation bounds of the Moore-Penrose inverse (Wedin, 1973), $\|\widetilde{A}-S_1A^*S^\top\|_2$ is also on the same order of $\|\hat{N}_1-N_1^*\|_F$.

To align $\widetilde{A}$ with $SA^*S^\top$, we compute another matrix $\hat{S}_0$ by solving the regression (3.8) from $\hat{M}_1[h_t;u_t]$ to $\hat{M}h_{t+1}$. Since $\hat{M}_1[h_t;u_t]$ and $\hat{M}h_{t+1}$ approximate $S_1z_{t+1}^*$ and Sz_{t+1}^*, respectively, (3.8) is essentially a linear regression that estimates the alignment matrix $SS_1^\top$ with perturbed variables $\hat{M}_1[h_t;u_t]$ and $\hat{M}h_{t+1}$. The ℓ_2 norm of the perturbation on Sz_t^* is given by (4.14). Similarly, the ℓ_2 norm of the other perturbation $\|\hat{M}_1[h_t;u_t]-S_1z_{t+1}^*\|$ is sub-Gaussian with its mean and sub-Gaussian norm bounded by

$$\mathcal{O}(H^{1/2}(H(d_y+d_u)+d_x)^{21/2}d_x^{3/2}T^{-1/2}\log^{1/2}(1/p)).$$

Hence, by the perturbed linear regression bound (Lemma 5), for T greater than a constant polynomial in the problem parameters, we have

$$\begin{aligned}&\|\hat{S}_0-SS_1^\top\|_2\\ &=\mathcal{O}(H^{1/2}(H(d_y+d_u)+d_x)^{21/2}d_x^{3/2}T^{-1/2}\log^{1/2}(1/p)\cdot d_x^{1/2}\log(T/p))\\ &=\mathcal{O}(H^{11}d_x^{2}(d_y+d_u)^{21/2}T^{-1/2}\log^{3/2}(T/p)).\end{aligned}$$

As a result,

$$\begin{aligned}\|\hat{A}-SA^*S^\top\|_2&=\|\hat{S}_0\widetilde{A}-SS_1^\top S_1A^*S^\top\|_2\\ &=\|(\hat{S}_0-SS_1^\top)\widetilde{A}\|_2+\|SS_1^\top(\widetilde{A}-S_1A^*S^\top)\|_2\\ &=\mathcal{O}(H^{11}d_x^{2}(d_y+d_u)^{21/2}T^{-1/2}\log^{3/2}(T/p)),\end{aligned}$$

and $\|\hat{B} - SB^*\|_2$ has the same order. Hence, $\|\hat{A} - SA^*S^\top\|_2$, $\|\hat{B} - SB^*\|_2$ and $\|\hat{Q} - SQ^*S^\top\|_2$ are all bounded by

$$\mathcal{O}(H^{11} d_x^{23/2} (d_y + d_u)^{21/2} T^{-1/2} \log^{5/2}(T/p)).$$

5 Additional related work

(Oymak and Ozay, 2019) studies the identification of partially observable linear dynamical systems from a single trajectory, which presents a finite-sample analysis of identifying the Markov parameter and a perturbation analysis of the Ho-Kalman algorithm (Ho and Kalman, 1966). (Simchowitz et al., 2019) relaxes the stability requirement to marginal stability by using prefiltered least squares to identify the Markov parameter. The method in (Zheng and Li, 2020) applies to unstable systems but requires multiple trajectories. Since the Markov parameter maps control input histories to observations, these methods do not work with costs and use the Markov parameter as an intermediate step to identify the system. By contrast, our methods, entirely driven by the costs and closely connected with empirical methods, directly learn the representation function and the latent model. Directly learning the latent model connects our work to the identification of fully observable linear dynamical systems. (Simchowitz et al., 2018) introduces small-ball conditions to handle correlated data and characterizes the statistical rates for stable and unstable systems, both proving to be useful for our analysis.

Online control of partially observable linear dynamical systems is considered in (Lale et al., 2020, 2021) for stochastic noises and in (Simchowitz et al., 2020) for nonstochastic noises. Reference (Zheng et al., 2021) considers end-to-end sample complexity and is closest to our setup. All these methods rely on the estimation of Markov parameters. For a discussion of the literature in more detail and breadth, we refer the reader to the recent survey (Tsiamis et al., 2022).

6 Conclusion and future work

We studied cost-driven state representation learning for solving unknown infinite-horizon time-invariant LQG control. We established finite-sample guarantees for two methods, which differ in whether the latent state dynamics is learned explicitly by minimizing the transition prediction errors, or implicitly by using the transition for future state and cost predictions, with the latter being motivated by that used in MuZero (Schrittwieser et al., 2020). For MuZero-style latent model learning, our analysis identifies a coordinate misalignment problem in the latent state space, suggesting the value of *multi-step* future state and cost prediction. A limitation of this work is that we only consider state representation based on truncated histories, i.e., frame stacking, as used in MuZero; the *recursive form* of the representation function, as in the Kalman filter, is also used empirically (Ha and Schmidhuber, 2018; Hafner et al., 2019a), and might be worth further investigation.

Many questions remain to be answered in state representation learning for control. Provable generalization of cost-driven state representation learning to nonlinear observation channels or dynamics is a natural consideration. Moreover, with the ubiquity of visual perception in real-world control systems, what if we have a time-varying observation function or multiple observation functions, modeling images taken from different angles? In reality, most of the time we do not have a well-defined cost function; learning task-relevant state representations from demonstrations is another intriguing direction.

Acknowledgement

This work was supported in part by the NSF TRIPODS program (award number DMS-2022448). KZ's work was mainly done while at MIT, and acknowledges partial support from Simons-Berkeley Research Fellowship. KZ also acknowledges support from U.S. Army Research Laboratory (ARL) Grant W911NF-24-1-0085.

References

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

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.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

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

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

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

N Komaroff. Iterative matrix bounds and computational solutions to the discrete algebraic Riccati equation. *IEEE Transactions on Automatic Control*, 39(8):1676–1678, 1994.

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.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

Shahar Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):1–25, 2015.

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.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

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

Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. *Advances in neural information processing systems*, 30, 2017.

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.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

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

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.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

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.

Yi Tian, Kaiqing Zhang, Russ Tedrake, and Suvrit Sra. Can direct latent model learning solve linear quadratic gaussian control? In *Learning for Dynamics and Control Conference*, pages 51–63. PMLR, 2023.

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

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite sample perspective. *arXiv preprint arXiv:2209.05423*, 2022.

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.

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

Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. Mastering Atari games with limited data. *Advances in Neural Information Processing Systems*, 34:25476–25488, 2021.

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.

Yang Zheng and Na Li. Non-asymptotic identification of linear dynamical systems using multiple trajectories. *IEEE Control Systems Letters*, 5(5):1693–1698, 2020.

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.

A Auxiliary results

Lemma 6. *Let x and y be random vectors defined on the same probability space. Then, $\|\mathbb{E}[xy^\top]\|_2^2 \leq \|\mathbb{E}[xx^\top]\|_2 \cdot \|\mathbb{E}[yy^\top]\|_2$.*

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

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

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

$$\|\mathbb{E}[xy^\top]\|_2^2 \leq \|\mathbb{E}[xx^\top]\|_2 \cdot \|\mathbb{E}[yy^\top]\|_2,$$

which completes the proof. □

Lemma 7. *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,

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

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

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

□