A UNIFIED FRAMEWORK FOR LEARNING, STABILIZATION, AND OPTIMIZATION IN ONLINE NONLINEAR CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online stochastic optimal control with unknown nonlinear dynamics. We address three limitations of existing approaches: prespecified parametric dynamics; reliance on a priori bounded trajectories or a known stabilizing controller; and restrictive noise assumptions, such as bounded disturbances or known covariance structures. Specifically, we propose Gibbs-Randomized Model-Based Adaptive Control (`GR-MBAC`), a two-stage algorithm that first learns and certifies a stabilizing feedback gain and then improves control performance through sieve-based dynamics estimation, noise-covariance estimation, and residual-policy updates. Under mild assumptions, we derive finite-sample bounds for the dynamics and covariance estimation errors and establish regret guarantees relative to the optimal finite-horizon controller. We further specialize the theoretical analysis to linear quadratic regulator, nonlinear Fourier dynamics, and nonlinear transient price impact. Under the corresponding conditions, the proposed controller achieves a sublinear regret bound that scales as the square root of the time horizon and is asymptotically optimal.

1 INTRODUCTION

In this paper, we study the online adaptive control of an unknown nonlinear stochastic system:

$$\widetilde{x}_{t+1} = f(\widetilde{x}_t, \widetilde{u}_t) + \widetilde{w}_t, \;\; 0 \le t \le N-1, \tag{1}$$

where N denotes the time horizon, $\widetilde{x}_t \in \mathbb{R}^{d_x}$ denotes the state, $\widetilde{u}_t \in \mathbb{R}^{d_u}$ denotes the feedback control, f denotes the unknown transition, and $\widetilde{w}_t \sim \mathcal{N}(0, \Sigma_w)$ denotes the Gaussian process noise with an unknown covariance matrix Σ_w. Such online control of unknown nonlinear dynamical systems is a foundational challenge in financial economics, cyber-physical systems, and beyond, drawing extensive research from financial mathematics (Chen et al., 2022; 2026; Cao et al., 2026; Coache & Jaimungal, 2024), control theory (Annaswamy et al., 2023; Vinod et al., 2022; Tsiamis & Pappas, 2022; Goel & Hassibi, 2023; Zhang et al., 2026; Bayraktar & Chen, 2023; Cohen et al., 2025; Han et al., 2025; Gao et al., 2026), and statistical learning (Muehlebach et al., 2026; Musavi et al., 2024; Park et al., 2025; Mania et al., 2022; Sattar & Oymak, 2022; Wagenmaker et al., 2023) communities. Despite this extensive body of work, existing approaches suffer from one or more of the following fundamental limitations:

(i) Known parametric structure. Many approaches assume a known structure for the nonlinear function f. For instance, Muehlebach et al. (2026), Musavi et al. (2024), Park et al. (2025), Mania et al. (2022), Sattar & Oymak (2022), Cohen et al. (2025), and Chen et al. (2022) assumed that $f(x_t, u_t) = \theta\phi(x_t, u_t)$, where the feature map ϕ is fully known and only the weight matrix θ is unknown; Annaswamy et al. (2023) assumed that $f(x_t, u_t)$ exhibits control-affine nonlinear structure (i.e., $f(x_t, u_t) = A(x_t) + B(x_t)u_t$ with unknown functions $A(\cdot), B(\cdot)$); Tsiamis & Pappas (2022), Goel & Hassibi (2023), and Han et al. (2025) assumed that $f(x_t, u_t)$ admits linear or nearly-linear structure; Bayraktar & Chen (2023) assumed that $f(x_t, u_t)$ is fully known a priori, with model uncertainty confined entirely to the distribution of real-valued noise term; Coache & Jaimungal (2024) made no structural assumptions on $f(x_t, u_t)$, but sacrificed algorithmic convergence and regret guarantees.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(ii) Trajectory boundedness assumption. Some existing papers rely on a priori restrictions on state growth. For instance, Self et al. (2022) and Brown et al. (2024) assumed that state-action trajectories remain confined to compact sets. Mania et al. (2022) and Wagenmaker et al. (2023) assumed that the feature map $\phi(x, u)$ is uniformly bounded over all states and admissible actions. This is a strong restriction: even linear dynamics $f(x, u) = Ax + Bu$ with $A \neq 0$ violate this boundedness condition of feature map $\phi(x, u) = (x^\top, u^\top)^\top$ over the entire state space $\mathbb{R}^{d_x}$. Alternatively, Musavi et al. (2024) imposed stability conditions ensuring bounded states under bounded inputs and disturbances, while Sattar & Oymak (2022) assumed that a stabilizing controller is already available before learning begins. However, requiring state-action trajectories to remain in a fixed compact set for all time is restrictive in the presence of unbounded disturbances; see Theorem 4 of Lu & Mo (2025).

(iii) Restrictive noise assumptions. Existing literature generally imposes heavily idealized assumptions on the underlying stochastic disturbances w_t. For instance, Vinod et al. (2022), Zhang et al. (2026), Annaswamy et al. (2023), and Gao et al. (2026) assumed that noise $w_t \equiv 0$ for all $t \geq 0$; Musavi et al. (2024), Mania et al. (2022), Sattar & Oymak (2022), and Chen et al. (2022) assumed that noise w_t is uniformly bounded almost surely for all $t \geq 0$; Muehlebach et al. (2026), Park et al. (2025), and Wagenmaker et al. (2023) assumed that w_t follows a Gaussian distribution with a fully known covariance matrix Σ; Bayraktar & Chen (2023) assumed that w_t is one-dimensional. In reality, however, physical, engineered, and financial systems are inescapably subject to high-dimensional, unbounded disturbances with unknown covariance (e.g., Gaussian or sub-Gaussian noise). While such unconstrained noise is ubiquitous in nature, it poses severe theoretical hurdles where prior works fall short.

We develop an online, stability-preserving adaptive control method based on sieve estimation for the unknown nonlinear stochastic system (1). During an initial exploration stage, Algorithm 1 learns and certifies a feedback gain $\widehat{\mathbf{K}}$. Once certified, the gain is held fixed while the algorithm sequentially estimates the unknown dynamics using a possibly growing sieve, estimates the unknown noise covariance, and updates the control policy based on the fitted model. Specifically, our main contributions are summarized as follows.

- We introduce Gibbs-Randomized Model-Based Adaptive Control (`GR-MBAC`; Algorithm 1), which learns and certifies a stabilizing feedback gain $\widehat{\mathbf{K}}$ during initial exploration and then holds it fixed for subsequent control. The technical challenge is to establish both the validity of the data-dependent certificate and finite-time acceptance despite model-estimation uncertainty. Under Assumption 1, Lemma 1 establishes that, with probability at least $1 - \delta_{\mathrm{sta}}$, certification succeeds by $\widetilde{N}$ and the learned gain $\widehat{\mathbf{K}}$ renders the closed-loop system globally exponentially input-to-state stable, as characterized by the quadratic Lyapunov inequalities (37) and (38).
- With the certified gain $\widehat{\mathbf{K}}$ held fixed, `GR-MBAC` estimates the unknown dynamics f over a possibly growing sieve and learns the noise covariance Σ_w from the same controlled trajectory. Theorems 1 and 2 provide explicit bounds on the errors in estimating dynamics f and covariance Σ_w from finitely many observations. The analysis handles adaptive data dependence and sieve misspecification through block small-ball arguments, self-normalized martingale concentration, and a Gibbs loss comparison. For covariance estimation, a held-out residual construction preserves the martingale structure of noise–model-error cross terms, while a weighted error analysis controls the cumulative influence of inaccurate historical model estimates.
- We design a composite control policy that combines stabilization $\widehat{\mathbf{K}}x_t$, performance improvement $u_{\mathrm{opt},t}$, and exploration e_t^{id}. Using a Bellman decomposition, Theorem 3 establishes a finite-horizon regret bound that quantifies the effects of model and covariance estimation errors, numerical policy-optimization error, and exploration. Appendix D specializes these guarantees to LQR, nonlinear Fourier dynamics, and nonlinear transient price impact. Under the respective conditions, the resulting $\widetilde{O}(H^{-1/2})$ average-regret bounds establish asymptotic optimality for the control stage following successful certification.

For further discussion of related work on learning for control, see Appendix A.2.

2 PRELIMINARIES

The notations used in this paper are given in Appendix A.1. We now introduce the problem studied in this paper. For positive integers $d_x, d_u, N \geq 2$, a function $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}^{d_x}$, a symmetric positive definite matrix Σ_w, and a distribution $\widetilde{\nu}_0$, consider an optimal control problem:

$$\begin{aligned} \min_{\pi \in \Pi} \quad & \mathcal{J}^N(\pi) := \mathbb{E}\left[\frac{1}{N}\sum_{t=0}^{N-1} c(\widetilde{x}_t, \widetilde{u}_t) + \widetilde{h}(\widetilde{x}_N)\right] \\ \text{s.t.} \quad & \widetilde{x}_{t+1} = f(\widetilde{x}_t, \widetilde{u}_t) + \widetilde{w}_t, \qquad \widetilde{u}_t \sim \pi_t, \qquad \widetilde{w}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_w), \qquad \widetilde{x}_0 \sim \widetilde{\nu}_0, \\ & \widetilde{I}_t := (\widetilde{x}_0, \widetilde{u}_0, \ldots, \widetilde{x}_{t-1}, \widetilde{u}_{t-1}, \widetilde{x}_t), \qquad t = 0, \ldots, N-1. \end{aligned} \tag{2}$$

Here, $\{\widetilde{x}_t\}_{t=0}^N$ denote the states, $\{\widetilde{w}_t\}_{t=0}^{N-1}$ denote the process noises, and $\{\widetilde{u}_t\}_{t=0}^{N-1}$ denote the controller inputs. In addition, Π denotes the class of admissible randomized policies $\pi = \{\pi_t\}_{t=0}^{N-1}$, where for each t, π_t denotes a conditional probability distribution of $\widetilde{u}_t$ given $\sigma(\widetilde{I}_t)$. Furthermore, $c : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \mapsto \mathbb{R}$ and $\widetilde{h} : \mathbb{R}^{d_x} \mapsto \mathbb{R}$ denote the stage and terminal cost functions, respectively. To simplify the analysis, we denote the distribution of $\widetilde{x}_t$ by $\widetilde{\nu}_t$ for $t = 1, \ldots, N-1$.

In this paper, f and Σ_w are unknown, and we want to design an online control algorithm for solving problem (2). Our motivation is threefold. **(1)** Much of the existing literature assumes that the unknown dynamics admit a known finite-dimensional representation, such as $f(x, u) = \theta\phi(x, u)$ with a prespecified feature map ϕ (see Muehlebach et al. (2026); Musavi et al. (2024); Park et al. (2025); Mania et al. (2022); Sattar & Oymak (2022); Cohen et al. (2025); Chen et al. (2022)). In contrast, our algorithm learns the dynamics using a sequence of increasingly expressive sieve classes, allowing approximation error at each stage without requiring any prespecified finite-dimensional model to represent the true dynamics exactly. **(2)** We aim to develop a unified online control algorithm applicable to a broad range of systems, including LQ systems, inventory and pricing models, and nonlinear industrial systems such as inverted pendulums. These systems often require substantially different feature representations of their dynamics, and existing control algorithms are typically tailored to each system. **(3)** Many existing works assume that state-action trajectories remain in compact sets (Self et al., 2022; Brown et al., 2024; Mania et al., 2022; Wagenmaker et al., 2023), or require a known stabilizing controller for the unknown system (Musavi et al., 2024; Sattar & Oymak, 2022). However, almost-sure confinement to a fixed compact set cannot generally be guaranteed under additive Gaussian noise (see Theorem 4 of Lu & Mo (2025)). A known stabilizing controller also requires prior information about the system that may be unavailable when its dynamics are unknown. Remarkably, our algorithm requires neither a prespecified compact set containing the trajectories nor a known stabilizing controller, while it learns and certifies a stabilizing feedback gain from online observations during its first stage. A simplified version of our algorithm is given below, and the complete algorithm is provided in Appendix B.1.

`GR-MBAC` at a glance.

Require: sieve specifications $(\phi^{(t)}, \Theta_t)$, update period $2M$, inverse temperatures η_t, residual radius B_{ctl}, horizon N, exploration horizon $\widetilde{N}$, and contraction threshold q_1, exploration amplitudes $\{\sigma_{\text{id},t}\}_{t\geq 0}$, independent exploration vectors $\{\zeta_t^{\text{id}}\}_{t\geq 0} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{d_u})$, an exploration policy $\{\pi_n\}_{n=0}^{N_u}$.

1: `Initialize`: observe $\widetilde{x}_0$ and set $\tau = N$.
2: **for** $n = 0, \ldots, \widetilde{N}$ **do**
3: `LearnCertify`: sample $\widetilde{u}_{n-1}$ from policy π_n and observe $\widetilde{x}_n$; compute a model estimate g_n of f and its confidence radius r_n using $\{(\widetilde{x}_i, \widetilde{u}_i, \widetilde{x}_{i+1})\}_{i=0}^{n-1}$.
4: `CertifyGain`: using (g_n, r_n) and optimization tolerance $\widetilde{\epsilon}_n$, compute a candidate gain K_n, a Lyapunov matrix P_n, and a certified upper bound U_n. If $U_n \leq q_1$, set $\tau = n$ and exit the loop.
5: **end for**
6: `Restart`: set $\widehat{\mathbf{K}} = K_\tau$, $H = N - \tau$, $x_0 = \widetilde{x}_\tau$, $\nu_0 = \widetilde{\nu}_\tau$ and $h = (N/H)\widetilde{h}$. Initialize a model draw θ_{-1}, noise covariance $\widehat{\Sigma}_{w,-1}$, and residual policy $\mu_{\theta_{-1}}$.
7: **for** $t = 0, \ldots, H-1$ **do**
8: get $\widetilde{x}_{\tau+t}$ and set $x_t = \widetilde{x}_{\tau+t}$.

if $t > 0$ and $\mathrm{mod}(t, 2M) = 0$ **then**
  `SieveFit`: compute $\widehat{\theta}_t \in \arg\min_{\theta \in \Theta_t} s_t(\theta)$; then call `GreedyCover` to construct $\mathcal{C}_t$.
  `GibbsSample`: sample $\theta_t \sim p_t(\theta) \propto \exp\{-\eta_t s_t(\theta)\}$ on $\mathcal{C}_t$.
  `CovarianceUpdate`: calculate $r_{t,i} = x_{i+1} - f_{\theta_t}(x_i, u_i)$ and update $\widehat{\Sigma}_{w,t}$.
  `FiniteHorizonSolve`: compute an approximation optimal residual policy $\mu_{\theta_t} = (\mu_{\theta_t}^{(0)}, \ldots, \mu_{\theta_t}^{(H-1)})$ for the fitted finite-horizon model with $(f_{\theta_t}, \widehat{\Sigma}_{w,t})$.
else
  `Hold`: keep θ_t, $\widehat{\Sigma}_{w,t}$, and μ_{θ_t} unchanged.
end if
`Deploy`: clip $\mu_{\theta_t}^{(t)}$: $u_{\mathrm{opt},t} = \mathrm{Proj}_{\mathbb{B}^{d_u}(0, B_{\mathrm{ctl}})}(\mu_{\theta_t}^{(t)}(x_t))$, draw $e_t^{\mathrm{id}} = \sigma_{\mathrm{id},t}\zeta_t^{\mathrm{id}}$, and apply $u_t = \widehat{\mathbf{K}}x_t + u_{\mathrm{opt},t} + e_t^{\mathrm{id}}$.
end for

Here, τ denotes the number of steps used by `GR-MBAC` to learn the stabilizing gain $\widehat{\mathbf{K}}$. For constants $\mathscr{K}, \mathscr{L} > 0$, $0 < \underline{v} \le \overline{v}$, and $0 < q_2 < q_1 < 1$, we define

$$\mathfrak{S} := \{(K, P) : K \in \mathbb{R}^{d_u \times d_x}, P \in \mathbb{R}^{d_x \times d_x}, \|K\|_{\mathrm{op}} \le \mathscr{K},\ \underline{v} I_{d_u} \preceq P \preceq \overline{v} I_{d_u}\}. \tag{3}$$

The following assumption gives conditions under which Algorithm 1 can learn a stabilizing controller within the given exploration horizon $\widetilde{N}$.

Assumption 1. It holds that

(a) $f(0,0) = 0$ and $\|f(x,u) - f(x,u')\|_2 \le \mathscr{L}\|u - u'\|_2$ for any $x \in \mathbb{R}^{d_x}, u \in \mathbb{R}^{d_u}$.

(b) There exists an unknown $(K_\star, P_\star) \in \mathfrak{S}$ that satisfies $\|f(x, K_\star x)\|_{P_\star} \le (q_1 - q_2)\|x\|_{P_\star}$ for any $x \in \mathbb{R}^{d_x}$.

Remark 1. Notably, $f(0,0) = 0$ in Assumption 1(a) can be imposed without loss of generality by translating the state and input to move any equilibrium to the origin. In addition, Assumption 1(b) requires an unknown linear feedback $K_\star$ under which the nominal closed-loop system admits a globally contracting quadratic Lyapunov function. Indeed, stabilizable linear systems satisfy it for suitably chosen q_1, q_2 and $\mathfrak{S}$. Beyond linear systems, Assumption 1(b) also holds under established matrix-inequality criteria for globally Lipschitz nonlinear systems (Nugroho et al., 2021) and discrete-time Luré systems with sector-bounded nonlinearities (Kim & Braatz, 2014). These classes include discretized flexible-link robot models with sinusoidal nonlinearities and dynamic neural-network models, respectively.

Lemma 1 (see Appendix B) guarantees that, under Assumption 1, Algorithm 1 finds and certifies a stabilizing gain $\widehat{\mathbf{K}}$ within $\widetilde{N}$ steps with probability at least $1 - \delta_{\mathrm{conf}}$. Hence, we now study the performance of Algorithm 1 after finding a stabilizing controller $\widehat{\mathbf{K}}$. Let $\mathscr{G}_\tau = \sigma(\widetilde{x}_0, \widetilde{u}_0, \ldots, \widetilde{x}_\tau)$ and set

$$H := N - \tau, \qquad x_t := \widetilde{x}_{\tau+t}, \qquad u_t := \widetilde{u}_{\tau+t}, \qquad w_t := \widetilde{w}_{\tau+t}, \qquad 0 \le t < H,$$
$$\nu_0 := \widetilde{\nu}_\tau, \qquad h := \frac{N}{H}\widetilde{h}, \qquad \mathbb{P}_\tau := \mathbb{P}(\,\cdot \mid \mathscr{G}_\tau), \qquad \mathbb{E}_\tau = \mathbb{E}[\,\cdot \mid \mathscr{G}_\tau].$$

The residual cost $\mathbb{E}_\tau\left[\frac{1}{N}\sum_{n=\tau}^{N-1} c(\widetilde{x}_n, \widetilde{u}_n) + \widetilde{h}(\widetilde{x}_N)\right]$ can be equivalently rewritten as

$$\frac{H}{N}\mathbb{E}_\tau\left[\frac{1}{H}\sum_{t=0}^{H-1} c(x_t, u_t) + h(x_H)\right]. \tag{4}$$

Let $\mathcal{T}_H := \{2jM : j = 1, \ldots, \lfloor (H-1)/(2M) \rfloor\}$ and define

$$\mathscr{F}_{-1} := \mathscr{G}_\tau, \qquad \mathscr{F}_t := \sigma\left(\mathscr{G}_\tau \cup \sigma(x_0, u_0, \ldots, x_t, u_t)\right), \qquad t = 0, \ldots, H-1,$$
$$\mathscr{D}_s := \sigma\left(\mathscr{G}_\tau \cup \sigma\big((x_i, u_i, x_{i+1})_{i=0}^{n_s-1}\big)\right), \qquad n_s := s - M - 1, \quad s \in \mathcal{T}_H.$$

For each update $s \in \mathcal{T}_H$, let Ξ_s denote all auxiliary random variables used to compute the controller in `FiniteHorizonSolve`, and set $\mathscr{R}_s^{\mathrm{opt}} := \sigma(\Xi_s)$. Define

$$\mathscr{H}_{-1} := \mathscr{F}_{-1},\ \mathscr{H}_t := \sigma\left(\mathscr{F}_t \cup \sigma(\zeta_0^{\mathrm{id}}, \ldots, \zeta_t^{\mathrm{id}}) \cup \bigcup_{\substack{s \in \mathcal{T}_H \\ s \le t}} \left(\sigma(\theta_s) \cup \mathscr{R}_s^{\mathrm{opt}}\right)\right), \qquad t = 0, \ldots, H-1.$$

3 THEORETICAL RESULTS

In this section, we establish explicit bounds on the estimation errors for the dynamics f and the noise covariance Σ_w, and then derive a regret bound for Algorithm 1 with respect to the objective in (4). We begin by specifying the sieve models used to estimate the unknown dynamics. Let f^{pub} denote the known component of f. For each $t \in \mathcal{T}_H$, let $\phi^{(t)} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_t}$ be the prespecified feature map and define $f_\theta(x,u) := f^{\mathrm{pub}}(x,u) + \theta\phi^{(t)}(x,u)$ for $\theta \in \Theta_t$. For a target accuracy $\Delta_t > 0$ and failure probability $\alpha_{\mathrm{app},t} \in (0,1)$, we define a sieve oracle $\theta_t^* \in \Theta_t$ satisfying

$$\mathbb{P}_\tau\left(\frac{1}{n_t}\sum_{i=0}^{n_t-1}\left\|f(x_i,u_i) - f_{\theta_t^*}(x_i,u_i)\right\|_2^2 > \Delta_t^2\right) \le \alpha_{\mathrm{app},t}. \tag{5}$$

Indeed, for a β-Hölder residual $f - f^{\mathrm{pub}}$ on a box $\mathcal{K}_t$, tensor-product Chebyshev sieves of coordinatewise degree m_t (with $d_t = (m_t+1)^{d_x+d_u}$ and sufficiently large Θ_t), achieve uniform approximation error $\widetilde{O}\left(d_t^{-\beta/(d_x+d_u)}\right)$ (Chen, 2007). If $\mathcal{K}_t$ contains the observed trajectory with $\mathbb{P}_\tau$-probability at least $1-\alpha_{\mathrm{app},t}$, taking $\Delta_t = C_t d_t^{-\beta/(d_x+d_u)}$ ensures (5). Thus, roughly speaking, (5) relaxes global uniform approximation to a high-probability bound on trajectory-averaged errors. Moreover, Lemma 3 in Appendix C.1 provides an explicit choice of $\mathcal{K}_t$ containing the observed state–action trajectory with $\mathbb{P}_\tau$-probability at least $1-\alpha_{\mathrm{app},t}$.

Next, we introduce the following assumption, which specifies persistence of excitation and is used to establish an estimation error bound for θ_t and $\widehat{\Sigma}_{w,t}$.

Assumption 2. Fix $t \in \mathcal{T}_H$ with $t < H$, and denote $D_t = d_x d_t$. There are $k_t \in [n_t]$, matrices $\Gamma_{\mathrm{PE},t} \succ 0$, $\Gamma_{\mathrm{up},t} \succ 0$, and $\kappa_{\mathrm{PE},t} \ge 1$ such that, for every $v \in \mathbb{S}^{d_t-1}$ and $0 \le j \le n_t - k_t$, with $m_{q,j,t}(v) := \frac{1}{k_t}\sum_{r=1}^{k_t}\mathbb{E}\left[|v^\top\phi_{j+r-1}^{(t)}|^q \mid \mathscr{H}_{j-1}\right]$, one has almost surely $m_{2,j,t}(v) \ge v^\top\Gamma_{\mathrm{PE},t}v$, and $m_{4,j,t}(v) \le \kappa_{\mathrm{PE},t}m_{2,j,t}(v)^2$. Moreover, it almost surely holds that $\mathbb{E}\left[\sum_{i=0}^{n_t-1}\phi_i^{(t)}(\phi_i^{(t)})^\top \middle| \mathscr{H}_{-1}\right] \preceq n_t\Gamma_{\mathrm{up},t}$.

Remark 2. The second-moment lower bound and fourth-moment upper bound in Assumption 2 have been widely used in the analysis of sample covariance matrices (Koltchinskii & Mendelson, 2015) and least-squares estimation (Mendelson, 2015; Ziemann & Tu, 2022). Meanwhile, such moment conditions are met for bilinear systems with Gaussian inputs and noise, and for analytic nonlinear systems under suitable excitation and stability conditions (see Sattar et al. (2022); Musavi et al. (2024)). In fact, the two moment conditions may be replaced by

$$\frac{1}{k_t}\sum_{r=1}^{k_t}\mathbb{P}\left(|v^\top\phi_{j+r-1}^{(t)}| \ge \sqrt{v^\top\Gamma_{\mathrm{PE},t}v/2} \,\middle|\, \mathscr{H}_{j-1}\right) \ge \frac{1}{4\kappa_{\mathrm{PE},t}}, \tag{6}$$

which is commonly used in finite-sample system identification (see Matni & Tu (2019); Li et al. (2024)). It has been verified for linear, bilinear, and analytic nonlinear systems under their respective assumptions; see Simchowitz et al. (2018); Sattar et al. (2022); Musavi et al. (2024). The upper bound on second moments in Assumption 2 is commonly used to control the empirical Gram matrix; see Matni & Tu (2019). This bound can be verified for linear systems and continuous features in the nonlinear setting considered by Musavi et al. (2024); Simchowitz et al. (2018). We also verify Assumption 2 for the LQR, Fourier, and nonlinear price impact examples under the conditions stated for each system in Appendix D.

Theorem 1. *Under Assumptions 1 and 2, for any $t \in \mathcal{T}_H$, choose $\alpha_{\mathrm{up},t}, \delta_{\mathrm{sb},t}, \delta_{\mathrm{mg},t} \in (0,1)$, $\lambda_t > 0$ and $\epsilon_t > 0$. Let $\mathcal{M}_t := \mathcal{M}(\Theta_t, \|\cdot\|_{\mathrm{F}}, \varepsilon_{\mathrm{cov},t})$ be the packing number of Θ_t, and assume*

$$n_t \ge \frac{10k_t}{(p_t^{\mathrm{sb}})^2}\left\{\log\frac{1}{\delta_{\mathrm{sb},t}} + 2d_t\log\frac{10}{p_t^{\mathrm{sb}}} + \log\det(2\overline{\Gamma}_t\Gamma_{\mathrm{PE},t}^{-1})\right\}. \tag{7}$$

Then θ_t in Algorithm 1 satisfies

$$\mathbb{P}_\tau(\|\theta_t - \theta_t^*\|_{\mathrm{F}} \ge \epsilon_t) \le q_t := \alpha_{\mathrm{up},t} + \delta_{\mathrm{sb},t} + \alpha_{\mathrm{app},t} + \delta_{\mathrm{mg},t} + \mathcal{M}_t\exp\left(-\eta_t\mathfrak{G}_t\right). \tag{8}$$

Here, all quantities appearing in the bound are defined explicitly in the proof; see Appendix C.1.

Proposition 1. *Under Assumptions 1, 2, and 4 with the parameters chosen in* (48)*, there exist constants $C > 0$ and $t_0 \geq 2(M+1)$ independent of H, such that, for every $t \in \mathcal{T}_H$ with $t \geq t_0$, it holds that*

$$\mathbb{P}_\tau\left(\sup_{(x,u)\in\mathcal{K}_t} \|f(x,u) - f_{\theta_t}(x,u)\|_2 \leq C\left[\frac{[\log t]^{(a+s)/2}}{d_t^{s/(d_x+d_u)-(b+1)/2}} + d_t^{1+b/2}\sqrt{\frac{\log t}{t}} \right] \right) \geq 1 - (t+1)^{-2}. \tag{9}$$

Here, all quantities appearing in the bound are defined explicitly in the proof; see Appendix C.2.

Theorem 2. *Under Assumptions 1 and 2, for every $t \in \mathcal{T}_H$, $\alpha_t \in (0,1)$, and $\epsilon_t^{\mathrm{cov}}, \zeta_t > 0$, it holds that*

$$\begin{aligned} &\mathbb{P}_\tau\Big(\|\widehat{\Sigma}_{w,t} - \Sigma_w\|_{\mathrm{op}} > \beta_t(\alpha_t) + \zeta_t + 2\epsilon_t^{\mathrm{cov}} \Big) \\ &\leq \alpha_t + \min\{1, \Lambda_t(\alpha_t,\zeta_t)\} + 2d_x^2 \exp\left(-\frac{t^2(\epsilon_t^{\mathrm{cov}})^2}{32 d_x^2 V_{\mathrm{cr},t}(\alpha_t,\zeta_t)} \right) \\ &\quad + 2\,9^{d_x} \exp\left[-c_{\mathrm{HW}} t \min\left\{ \frac{(\epsilon_t^{\mathrm{cov}})^2}{8\|\Sigma_w\|_{\mathrm{op}}^2}, \frac{\epsilon_t^{\mathrm{cov}}}{4\|\Sigma_w\|_{\mathrm{op}}} \right\} \right]. \end{aligned} \tag{10}$$

Here, all quantities appearing in the bound are defined explicitly in the proof; see Appendix C.3.

Remark 3. Under suitable choices of the sieve dimensions and tuning parameters, the right-hand sides of (8) and (10) converge to zero for linear systems and a broad class of nonlinear systems. Appendix D verifies this convergence and derives explicit rates for LQR, nonlinear Fourier dynamics, and nonlinear transient price impact under the respective conditions. Moreover, existing identification studies often develop estimation and exploration procedures tailored to specified model structures. For instance, Simchowitz et al. (2018) analyzed ordinary least squares for linear systems; Mania et al. (2022) combined trajectory planning, tracking, and least-squares estimation for nonlinear systems with known features; and Sattar & Oymak (2022) analyzed gradient descent for known parametric nonlinear models. In pricing and finance, Chen et al. (2022) combined parametric demand regression with empirical noise-distribution estimation for joint pricing and inventory control; Cao et al. (2026) used regularized maximum likelihood to estimate price sensitivity in the Avellaneda–Stoikov market-making model; and Neuman & Zhang (2023) estimated price-impact kernels through regularized least squares for a convolution inverse problem. Nevertheless, our framework uses the same Gibbs-based dynamics estimator f_{θ_t} and residual-based covariance estimator $\widehat{\Sigma}_{w,t}$ across linear and nonlinear systems, with suitable sieve representations and parameters. Remarkably, our unified framework also applies to models studied in the aforementioned literature. Again, Appendix D provides detailed verifications for LQR, nonlinear Fourier dynamics, and nonlinear transient price-impact models; other models can be treated likewise.

We now derive a regret bound for Algorithm 1 with respect to the objective in (4); to this end, we first introduce some notation. For any policy $\pi \in \Pi$, define

$$\mathcal{J}^H(\pi) := \mathbb{E}_\tau\left[\frac{1}{H}\sum_{t=0}^{H-1} c(x_t,u_t) + h(x_H) \right],\ u_t \sim \pi_{t+\tau}.$$

Define $\mathcal{J}^{H,*} := \inf_{\pi\in\Pi} \mathcal{J}^H(\pi)$ and the finite-horizon optimal value functions

$$V_H^*(x) := h(x), V_t^*(x) := \inf_{u\in\mathbb{R}^{d_u}} \left\{ \frac{1}{H}c(x,u) + \mathbb{E}_\tau[V_{t+1}^*(f(x,u)+w_t)] \right\},\quad 0 \leq t < H.$$

Let $\pi^{\mathrm{A}} := (\pi_t^{\mathrm{A}})_{t=0}^{N-1}$ denote the adapted policy generated by Algorithm 1. The finite-horizon Bellman residual is defined as

$$\mathcal{D}_t^H(x,u) := \frac{1}{H}c(x,u) + \mathbb{E}_\tau[V_{t+1}^*(f(x,u)+w_t) - V_t^*(x)]. \tag{11}$$

For any $\alpha \in (0,1)$ and $s \in \mathcal{T}_H$, there exists a compact set $\mathcal{K}_s(\alpha) \subseteq \mathbb{R}^{d_x} \times \mathbb{R}^{d_u}$ such that

$$\mathbb{P}\left(\bigcap_{s\leq t<(s+2M)\wedge H} \{(x_t,u_t) \in \mathcal{K}_s(\alpha)\} \right) \geq 1-\alpha. \tag{12}$$

Indeed, such a compact set exists because

$$\lim_{r\to\infty} \mathbb{P}\left(\max_{s\le t<(s+2M)\wedge H} \max\{\|x_t\|_\infty, \|u_t\|_\infty\} \le r \right) = 1.$$

For $s \in \mathcal{T}_H$, define

$$\chi_s(\alpha) := \sup_{(x,u)\in\mathcal{K}_s(\alpha)} \|f(x,u) - f_{\theta_s^*}(x,u)\|_2,$$

$$\mathcal{G}_s(a_s, c_s, \alpha) := \{\|\theta_s - \theta_s^*\|_{\mathrm{F}} \le a_s,\ \|\widehat{\Sigma}_{w,s} - \Sigma_w\|_{\mathrm{op}} \le c_s\} \cap \{(x_t, u_t) \in \mathcal{K}_s(\alpha),\ s \le t < (s+2M)\wedge H\}$$

with $a_s, c_s \ge 0$.

Assumption 3. At the fixed confidence level α, there is a common $p > 1$ such that, for every update $s \in \mathcal{T}_H$, finite deterministic numbers $\beta_s^{\mathrm{Bell}}(\alpha), C_s^{\mathrm{id}} \ge 0$ can be chosen so that, on $\mathcal{G}_s(a_s, c_s, \alpha)$,

$$\mathcal{D}_t^H(x_t, u_t) \le \beta_s^{\mathrm{Bell}}(\alpha) + C_s^{\mathrm{id}} \|e_t^{\mathrm{id}}\|_2^2 \tag{13}$$

holds for every $s \le t < (s+2M)\wedge H$. In addition, we have $C_s^{\mathrm{bad}} := \max_{s\le t<(s+2M)\wedge H} \|\mathcal{D}_t^H(x_t,u_t)\|_{L^p} < \infty$, and $B_0^{\mathrm{init}} := \sum_{t=0}^{(2M)\wedge H-1} \mathbb{E}_\tau \mathcal{D}_t^H(x_t,u_t) < \infty$.

Remark 4. Assumption 3 is widely used in online control and reinforcement learning; see Muehlebach et al. (2026); Kakade et al. (2020); Wang et al. (2025) and the reference therein. Furthermore, Assumption 3 has been verified in many studies. Equation (13) has been established for LQR under sufficiently accurate model estimates (Mania et al., 2019; Simchowitz & Foster, 2020) and for smooth nonlinear parametric systems (Wang et al., 2025). The conditions on C_s^{bad} and B_0^{init} hold for stable linear feedback systems with quadratic costs, Gaussian noise, and suitable initial-state moments, or for nonlinear systems with bounded stage and terminal costs Kakade et al. (2020); Wang et al. (2025). In Appendix D, we also verify Assumption 3 for the LQR, Fourier, and nonlinear price impact examples under the conditions stated for each system.

For later use, denote

$$\delta_s := \mathbb{P}(\mathcal{G}_s(a_s,c_s,\alpha)^c), \qquad \ell_{s,H} := 2M \wedge (H-s),$$

$$\mathrm{Err}_{\mathrm{exp}}(H) := B_0^{\mathrm{init}} + \sum_{s\in\mathcal{T}_H} \left[\ell_{s,H}\Big(C_s^{\mathrm{bad}} \delta_s^{1-1/p} + \beta_s^{\mathrm{Bell}}(\alpha) \Big) + d_u C_s^{\mathrm{id}} \sum_{t=s}^{s+\ell_{s,H}-1} \sigma_{\mathrm{id},t}^2 \right]. \tag{14}$$

Theorem 3. *Under Assumptions 1—3, it holds that*

$$\mathcal{J}^H(\pi^{\mathrm{A}}) - \mathcal{J}^{H,*} \le \mathrm{Err}_{\mathrm{exp}}(H). \tag{15}$$

The proof of Theorem 3 is given in Appendix C.4. This theorem translates the statistical guarantees of Theorems 1 and 2 into control-performance bounds. In Appendix D, we derive explicit regret rates for three systems: LQR, the Fourier model, and nonlinear transient price impact. Under the respective assumptions, all three specializations achieve, after successful stabilization,

$$\mathcal{J}^H(\pi^{\mathrm{A}}) - \mathcal{J}^{H,*} \le \widetilde{O}(H^{-1/2}).$$

For LQR (Appendix D.1), this recovers the rate of Cohen et al. (2019); Simchowitz & Foster (2020), which is minimax optimal in its horizon dependence up to logarithmic factors. For nonlinear transient price impact (Appendix D.3), our bound matches the horizon dependence of the parametric guarantee in Muehlebach et al. (2026), while also accounting for noise-covariance estimation. The Fourier example (Appendix D.2) further demonstrates how our framework handles model approximation:

$$f(x,u) = ax + b_\star u + \sum_{\ell\ge1} \alpha_{\ell,\star} \sin(\ell x), \qquad |\alpha_{\ell,\star}| \le A_{\mathrm{F}} e^{-\nu_{\mathrm{F}}\ell},$$

where a is known, while $b_\star$ and the Fourier coefficients are unknown. This model permits infinitely many nonzero Fourier coefficients and therefore need not admit an exact finite Fourier representation. It thus extends beyond the fixed-dimensional parametric guarantees of Boffi et al. (2021);

Wang et al. (2025); Muehlebach et al. (2026). Since our framework allows nonzero sieve approximation error, we use a growing sieve with $d_H = \left\lceil \frac{\log H}{2\nu_{\mathrm{F}}} \right\rceil + 1$, whose truncation error satisfies $\Delta_H = \widetilde{O}(H^{-1/2})$. Controlling this error together with the growing model complexity yields the stated $\widetilde{O}(H^{-1/2})$ average-regret bound. Moreover, Table 1 in Appendix B.2 provides a detailed comparison of our regret bounds and assumptions with those of existing methods (e.g., Cohen et al. (2019); Hazan et al. (2020); Simchowitz & Foster (2020); Boffi et al. (2021); Kakade et al. (2020); Wang et al. (2025); Muehlebach et al. (2026); Park & Van Roy (2015)), highlighting the broad applicability and flexibility of our online control framework.

4 Numerical Experiments

We conduct three experiments to evaluate the performance of Algorithm 1. For each experiment, we compare the performance of three algorithms: {A-KNOWN, A-SIEVE, SPECIALIST}. Here, A-KNOWN is a realization of Algorithm 1 knowing that f can be expressed by some special basis functions and setting such basis functions as $\phi^{(t)}$. A-SIEVE is a realization of Algorithm 1 without any prior knowledge of f and setting $\phi^{(t)}$ as a general sieve class with growing d_t. SPECIALIST is an existing adaptive control algorithm. We use this algorithm as a baseline method to compare the performance with those of proposed A-KNOWN and A-SIEVE. Furthermore, we use the true dynamics f and the true noise covariance matrix Σ_w to compute an optimal or near-optimal controller for the corresponding control problem. We refer to this controller as the oracle controller and evaluate the cumulative cost of each algorithm relative to it. In each experiment, we first randomly sample tasks from a distribution and run the three algorithms and the oracle controller on each task to generate multiple state-action trajectories. For a fair comparison, in each task, we first let A-KNOWN and A-SIEVE learn the stabilizing gain $\widehat{\mathbf{K}}$. Then, we restart the two algorithms from the same initial state as the ones of SPECIALIST and oracle controller to test the performance of A-KNOWN and A-SIEVE. For each algorithm, we report the averaged cumulative regret $H(\mathcal{J}^H(\pi^A) - \mathcal{J}^H(\pi^{\mathrm{oracle}}))$ and the averaged stage cost $c(x_t, u_t)$ at each time step t over the test trajectories. Here, π^A denotes the policy generated by Algorithm $A \in \{\text{A-KNOWN, A-SIEVE, Specialist}\}$ and π^{oracle} denotes the policy of the oracle controller. We also report 95% Student-t confidence bands for the regret bound and the stage cost. In addition, we present state-action trajectories of the three algorithms and the oracle controller in a testing task. For A-KNOWN and A-Sieve, we also present the number of the steps that each algorithm find the controller stabilizing the true model and refer to the number as stabilization time.

We first show the experimental results for the LQR problem. Due to the page limit, the experimental reports on the pendulum system and the pricing inventory problem are presented in Appendices F.2 and F.3. Consider a family of discrete-time, three-dimensional linear systems

$$x_{t+1} = \begin{bmatrix} 1 & 0.1 & 0 \\ 0 & 1 & 0.1 \\ 0.1a & 0 & 1 - 0.1\gamma \end{bmatrix} x_t + \begin{bmatrix} 0 \\ 0 \\ 0.1b \end{bmatrix} u_t + w_t, \qquad x_t \in \mathbb{R}^3, \quad u_t \in \mathbb{R} \tag{16}$$

with

$$a, b \sim \mathrm{Unif}[0.8, 1.2], \qquad \gamma \sim \mathrm{Unif}[0.10, 0.30],$$
$$w_t \sim \mathcal{N}(0, \sigma^2 I_3), \qquad \sigma \sim \mathrm{Unif}[0.02, 0.04], \quad x_0 \sim \mathrm{Unif}([-1, 1]^3).$$

The stage cost and terminal cost functions are defined by

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t, \qquad h(x_H) \equiv 0, \qquad t = 0, \ldots, H-1$$

with $H = 100$, $Q = \mathrm{diag}(1, 0.1, 0.1)$, and $R = 0.1$. In each task, we sample a model and a common initial state from the aforementioned distributions for the three algorithms and the oracle controller. A-KNOWN implements Algorithm 1 with $\phi^{(t)}(x, u) = [x^\top, u^\top]^\top$ for any $t = 0, \ldots, H-1$ and uses the Riccati recursion method (see Bertsekas et al. (1995)) to realize `FiniteHorizonSolve`. A-SIEVE sets $\phi^{(t)}$ as the following form:

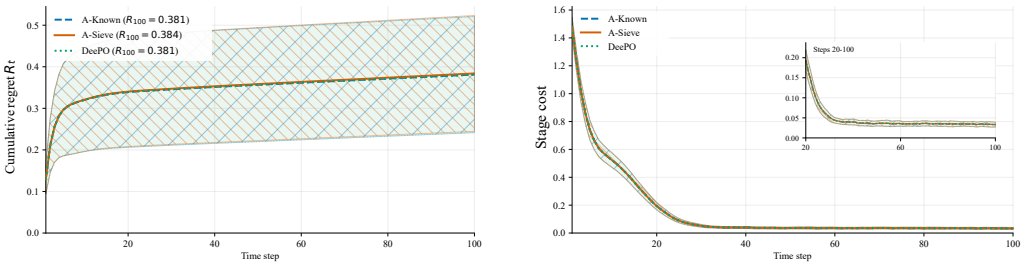
$$z = (x^\top, u)^\top = (z_1, \ldots, z_4)^\top \in \mathbb{R}^4, \quad m_t = \left\lfloor \log_2 \left(1 + \frac{1}{32} \left\lfloor \frac{t}{2M} \right\rfloor \right) \right\rfloor,$$

$$\phi^{(t)}(z) = \mathrm{col}\left(z,\ \{\sin z_i\}_{i=1}^{4},\ \chi(z_1), \chi(z_2),\ \left\{\frac{2^{-j}}{(j+2)^2}\sin((j+2)z_1)\right\}_{j=1}^{m_t}\right)$$

with $T_2(s) = 2s^2 - 1$, $\chi(s) := T_2(\sin s) - T_2(0) = 2\sin^2 s, s \in \mathbb{R}$. Meanwhile, A-SIEVE uses the iterative linear quadratic regulator (iLQR) method (see Li & Todorov (2004)) to realize `FiniteHorizonSolve`. SPECIALIST is the DeePO method, which is widely used in LQ problems with unknown models; see (Zhao et al., 2025). Furthermore, we set the oracle controller as the LQR controller designed by the true model parameters (Bertsekas et al. (1995)). We test the three algorithms and the oracle controller via 20 tasks with 50 trajectories collected per task. In each task, we first let A-KNOWN and A-SIEVE learn the stabilizing gain $\widehat{\mathbf{K}}$ by using the method shown in Appendix B.1.1 and B.1.2; see Appendix F.1 for the formal realization. Then, we restart A-KNOWN and A-SIEVE to the initial state and test their performances with the same parameters as follows:

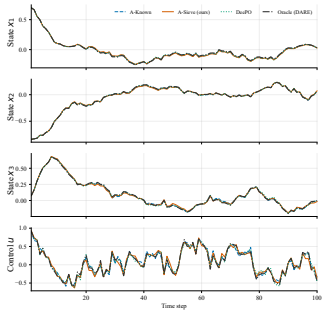
$$M = 3, \quad e_t^{\mathrm{id}} \sim \mathcal{N}(0, 0.05^2), \quad \eta_t = 1, \quad B_{\mathrm{ctl}} = \log H, \quad \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \quad f^{\mathrm{pub}} \equiv \mathbf{0}_3$$

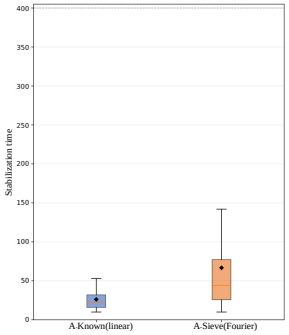
with $t = 0, \ldots, H-1$. The experimental results are shown below.



(a) Cumulative regret.

(b) Stage cost.





(c) Representative state and control input trajectories.

(d) Box plot of the times that A-KNOWN and A-SIEVE use for find a stabilizing gain $\widehat{\mathbf{K}}$.

Figure 1: Experimental performance of A-KNOWN, A-SIEVE, and SPECIALIST on 20 tasks with 50 trajectories per task.

Figures 1a–1c compare the performances of the three algorithms in the cumulative regret, stage cost, and the state and control input trajectories. The final cumulative regrets of A-KNOWN, A-SIEVE, and DeePO are approximately 0.381, 0.384, and 0.381, respectively. Meanwhile, the three algorithms also exhibit similar stage costs and trajectories of states and control inputs. Furthermore, these results show that, on the tested LQR systems, A-SIEVE can achieve performance nearly comparable to both A-KNOWN and the DeePO algorithm without prior information of the task models. Figure 1d shows that for each task, both A-SIEVE and A-KNOWN can learn the stabilizing gain $\widehat{\mathbf{K}}$ in 150 steps. Although A-SIEVE has no prior knowledge of the task models and uses much more steps (to learn $\widehat{\mathbf{K}}$) than the one used by A-KNOWN, it still learns the stabilizing gains successfully for every unknown model.

AI USE STATEMENT

In this work, we used generative AI tools for checking and correcting mathematical proofs, as well as for editing and checking experiment-related materials. We have not used generative AI tools for proposing the algorithm, designing the experiments, generating synthetic data, implementing the core method, or determining the reported results. All AI-assisted work was manually reviewed by the authors. All mathematical proofs, experimental results, claims, and other reported content were checked and finalized by the authors. We take responsibility for the final content of this work, including all text, claims, proofs, data, code, and other artifacts.

REFERENCES

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

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Aurélien Alfonsi, Antje Fruth, and Alexander Schied. Optimal execution strategies in limit order books with general shape functions. *Quantitative finance*, 10(2):143–157, 2010.

Anuradha M Annaswamy, Anubhav Guha, Yingnan Cui, Sunbochen Tang, Peter A Fisher, and Joseph E Gaudio. Integration of adaptive control and reinforcement learning for real-time control and learning. *IEEE Transactions on Automatic Control*, 68(12):7740–7755, 2023.

Venkataramanan Balakrishnan, Stephen Boyd, and Silvano Balemi. Branch and bound algorithm for computing the minimum stability degree of parameter-dependent linear systems. *International Journal of Robust and Nonlinear Control*, 1(4):295–317, 1991.

Erhan Bayraktar and Tao Chen. Nonparametric adaptive robust control under model uncertainty. *SIAM Journal on Control and Optimization*, 61(5):2737–2760, 2023.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186 (2):345–366, 2015.

Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Tao Bian and Zhong-Ping Jiang. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 71:348–360, 2016.

Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pp. 471–483. PMLR, 2021.

Ilija Bogunovic and Andreas Krause. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34:3004–3015, 2021.

Stephen Boyd, Laurent El Ghaoui, Eric Feron, and Venkataramanan Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

William Brown, Christos Papadimitriou, and Tim Roughgarden. Online stackelberg optimization via nonlinear control. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 697–749. PMLR, 2024.

Joan Bruna, Benjamin Peherstorfer, and Eric Vanden-Eijnden. Neural galerkin schemes with active learning for high-dimensional evolution equations. *Journal of Computational Physics*, 496: 112588, 2024.

Jialun Cao, David Šiška, Łukasz Szpruch, and Tanut Treetanthiploet. Logarithmic regret in the ergodic avellaneda–stoikov market making model. *SIAM Journal on Financial Mathematics*, 17 (3):877–925, 2026.

Boxiao Chen, David Simchi-Levi, Yining Wang, and Yuan Zhou. Dynamic pricing and inventory control with fixed ordering cost and incomplete demand information. *Management Science*, 68 (8):5684–5703, 2022.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Ying Chen, Ulrich Horst, and Hoang Hai Tran. Optimal trade execution under endogenous order flow. *Operations Research*, 74(1):72–92, 2026.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in Neural Information Processing Systems*, 31, 2018.

Anthony Coache and Sebastian Jaimungal. Reinforcement learning with dynamic convex risk measures. *Mathematical Finance*, 34(2):557–587, 2024.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pp. 1300–1309. PMLR, 2019.

Samuel N Cohen, Christoph Knochenhauer, and Alexander Merkel. Optimal adaptive control with separable drift uncertainty. *SIAM Journal on Control and Optimization*, 63(2):1348–1373, 2025.

Gianbiagio Curato, Jim Gatheral, and Fabrizio Lillo. Optimal execution with non-linear transient market impact. *Quantitative Finance*, 17(1):41–54, 2017.

Claudio De Persis and Pietro Tesi. Formulas for data-driven control: Stabilization, optimality, and robustness. *IEEE Transactions on Automatic Control*, 65(3):909–924, 2019.

Claudio De Persis and Pietro Tesi. Low-complexity learning of linear quadratic regulators from noisy data. *Automatica*, 128:109548, 2021.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Florian Dörfler, Pietro Tesi, and Claudio De Persis. On the certainty-equivalence approach to direct data-driven lqr design. *IEEE Transactions on Automatic Control*, 68(12):7989–7996, 2023.

Guillermo Gallego and Garrett Van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science*, 40(8):999–1020, 1994.

Weinan Gao, Yebin Wang, and Kyriakos G Vamvoudakis. Output-feedback learning-based adaptive optimal control of nonlinear systems. *Automatica*, 187:112875, 2026.

Gautam Goel and Babak Hassibi. Regret-optimal estimation and control. *IEEE Transactions on Automatic Control*, 68(5):3041–3053, 2023.

Jiequn Han, Arnulf Jentzen, et al. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in mathematics and statistics*, 5(4):349–380, 2017.

Yinbin Han, Meisam Razaviyayn, and Renyuan Xu. Policy gradient converges to the globally optimal policy for nearly linear-quadratic regulators. *SIAM Journal on Control and Optimization*, 63 (4):2936–2963, 2025.

Elad Hazan, Sham Kakade, and Karan Singh. The nonstochastic control problem. In *Algorithmic Learning Theory*, pp. 408–421. PMLR, 2020.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pp. 12790–12822. PMLR, 2023.

Petros A Ioannou and Jing Sun. *Robust adaptive control*. Courier Corporation, 2012.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

Kwang-Ki K Kim and Richard D Braatz. Observer-based output feedback control of discrete-time luré systems with sector-bounded slope-restricted nonlinearities. *International Journal of Robust and Nonlinear Control*, 24(16):2458–2472, 2014.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

He Kong. Noise covariance estimation via autocovariance least-squares with deadbeat filters. *Automatica*, 153:111049, 2023.

Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on Neural Networks*, 9(5):987–1000, 1998.

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

Weiwei Li and Emanuel Todorov. Iterative linear quadratic regulator design for nonlinear biological movement systems. In *First International Conference on Informatics in Control, Automation and Robotics*, volume 2, pp. 222–229. SciTePress, 2004.

Yingying Li, Jing Yu, Lauren Conger, Taylan Kargin, and Adam Wierman. Learning the uncertainty sets of linear control systems via set membership: A non-asymptotic analysis. In *Forty-first International Conference on Machine Learning*, 2024.

Yiwen Lu and Yilin Mo. Almost surely $\sqrt{T}$ regret for adaptive lqr. *IEEE Transactions on Automatic Control*, 70(8):5145–5159, 2025.

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

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *Journal of Machine Learning Research*, 23(32):1–30, 2022.

Nikolai Matni and Stephen Tu. A tutorial on concentration bounds for system identification. In *2019 IEEE 58th conference on decision and control (CDC)*, pp. 3741–3749. IEEE, 2019.

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

Michael Muehlebach, Zhiyu He, and Michael Jordan. The sample complexity of online reinforcement learning: A multi-model perspective. In *International Conference on Learning Representations*, volume 2026, pp. 119924–119960, 2026.

Negin Musavi, Ziyao Guo, Geir Dullerud, and Yingying Li. Identification of analytic nonlinear dynamical systems with non-asymptotic guarantees. *Advances in Neural Information Processing Systems*, 37:85500–85522, 2024.

Arnold Neumaier. Complete search in continuous global optimization and constraint satisfaction. *Acta numerica*, 13:271–369, 2004.

Eyal Neuman and Yufei Zhang. Statistical learning with sublinear regret of propagator models. *arXiv preprint arXiv:2301.05157*, 2023.

Sebastian A Nugroho, Suyash C Vishnoi, Ahmad F Taha, and Christian G Claudel. Feedback stabilization and output tracking for discrete-time lipschitz nonlinear systems via iterative convex approximations. In *2021 American Control Conference (ACC)*, pp. 4824–4829. IEEE, 2021.

Anna A Obizhaeva and Jiang Wang. Optimal trading strategy and supply/demand dynamics. *Journal of Financial markets*, 16(1):1–32, 2013.

Brian J Odelson, Murali R Rajamani, and James B Rawlings. A new autocovariance least-squares method for estimating noise covariances. *Automatica*, 42(2):303–308, 2006.

Beomsoo Park and Benjamin Van Roy. Adaptive execution: Exploration and learning of price impact. *Operations Research*, 63(5):1058–1076, 2015.

Hyuk Park, Grani A Hanasusanto, and Yingying Li. Robust system identification: Non-asymptotic guarantees and connection to regularization. In *13th International Conference on Learning Representations, ICLR 2025*, pp. 102144–102166. International Conference on Learning Representations, ICLR, 2025.

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

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23(140):1–49, 2022.

Yahya Sattar, Samet Oymak, and Necmiye Ozay. Finite sample identification of bilinear dynamical systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 6705–6711. IEEE, 2022.

Ryan Self, Moad Abudia, SM Nahid Mahmud, and Rushikesh Kamalapurkar. Model-based inverse reinforcement learning for deterministic systems. *Automatica*, 140:110242, 2022.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pp. 8937–8948. PMLR, 2020.

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*, pp. 439–473. PMLR, 2018.

Yuda Song and Wen Sun. Pc-mlp: Model-based reinforcement learning with policy cover guided exploration. In *International Conference on Machine Learning*, pp. 9801–9811. PMLR, 2021.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Sho Takemori and Masahiro Sato. Approximation theory based methods for rkhs bandits. In *International Conference on Machine Learning*, pp. 10076–10085. PMLR, 2021.

Yuval Tassa, Tom Erez, and Emanuel Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4906–4913. IEEE, 2012.

Anastasios Tsiamis and George J Pappas. Online learning of the kalman filter with logarithmic regret. *IEEE Transactions on Automatic Control*, 68(5):2774–2789, 2022.

Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. In *Conference on Learning Theory*, pp. 3820–3857. PMLR, 2022.

Abraham P Vinod, Arie Israel, and Ufuk Topcu. On-the-fly control of unknown nonlinear systems with sublinear regret. *IEEE Transactions on Automatic Control*, 68(5):2729–2742, 2022.

Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 36:15406–15455, 2023.

Feicheng Wang and Lucas Janson. Exact asymptotics for linear quadratic adaptive control. *Journal of Machine Learning Research*, 22(265):1–112, 2021.

James Wang, Bruce D Lee, Ingvar Ziemann, and Nikolai Matni. Logarithmic regret for nonlinear control. *arXiv preprint arXiv:2501.10261*, 2025.

Hengde Zhang, Yunxiao Ren, Zhisheng Duan, Zhiyong Sun, and Guanrong Chen. Deep koopman iterative learning and stability-guaranteed control for unknown nonlinear time-varying systems. *Automatica*, 190:113054, 2026.

Lingyi Zhang, David Sidoti, Adam Bienkowski, Krishna R Pattipati, Yaakov Bar-Shalom, and David L Kleinman. On the identification of noise covariances and adaptive kalman filtering: A new look at a 50 year-old problem. *IEEE Access*, 8:59362–59388, 2020.

Yanxia Zhang and Lei Guo. A limit to the capability of feedback. *IEEE Transactions on Automatic Control*, 47(4):687–692, 2002.

Feiran Zhao, Florian Dörfler, Alessandro Chiuso, and Keyou You. Data-enabled policy optimization for direct adaptive learning of the lqr. *IEEE Transactions on Automatic Control*, 70(11):7217–7232, 2025.

Mo Zhou and Jianfeng Lu. A policy gradient framework for stochastic optimal control problems with global convergence guarantee. *SIAM Journal on Control and Optimization*, 63(4):2605–2631, 2025.

Ingvar Ziemann and Stephen Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

A Literature and Model-Specific Comparisons

A.1 Notation

For a set A, we denotes the number of elements in A by $|A|$. Let $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ denote the set of natural numbers, $\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$ denote the set of integers, and $\mathbb{R}$, $\mathbb{C}$ denote the sets of real and complex numbers, respectively. For $n \in \mathbb{N}_+$, let $[n]$ be $\{1, \ldots, n\}$. For any $x, y \in \mathbb{R}$, let $x \wedge y$ be $\min\{x, y\}$ and $x \vee y$ be $\max\{x, y\}$. For any positive integers m and n, we denote by $\mathbb{R}^m$ the set of real m-dimensional column vectors and by $\mathbb{R}^{m\times n}$ the set of real $m \times n$ matrices. For a square matrix $A \in \mathbb{R}^{n\times n}$, let $\det(A)$ denote its determinant and $\rho(A)$ denote its spectral radius. Consider a sample space Ω, For any event $A \subseteq \Omega$, let $\mathbf{1}\{A\} : \Omega \mapsto \{0, 1\}$ denote the indicator function:

$$\mathbf{1}\{A\}(\omega) := \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A. \end{cases}$$

For $\mu \in \mathbb{R}^m$ and a symmetric positive definite matrix $\Sigma \in \mathbb{R}^{m\times m}$, let $\mathcal{N}(\mu, \Sigma)$ denote the Gaussian distribution with mean μ and covariance matrix Σ. For a probability distribution ν, we use $X \sim \nu$ to denote that the random variable X has distribution ν. For an index set $\mathcal{I}$, we use $X_i \overset{\text{iid}}{\sim} \nu$, $i \in \mathcal{I}$, to denote that the random variables $(X_i)_{i\in\mathcal{I}}$ are independent and identically distributed with common distribution ν. For a vector $x = (x_1, \ldots, x_m)^\top \in \mathbb{R}^m$, denote its Euclidean norm and maximum norm by $\|x\| = \sqrt{\sum_{i=1}^m x_i^2}$, $\|x\|_\infty = \max_{1\le i\le m} |x_i|$, respectively. For a matrix $A \in \mathbb{R}^{m\times n}$, denote its Frobenius norm by $\|A\|_{\mathrm{F}}$, spectral norm by $\|A\|_{\mathrm{op}}$, and infinity norm by $\|A\|_\infty$, respectively. For a symmetric positive definite matrix $P \in \mathbb{R}^{m\times m}$ and $x \in \mathbb{R}^m$, define the weighted Euclidean norm by $\|x\|_P := \sqrt{x^\top P x}$. For $d \in \mathbb{N}_+$ and $r \ge 0$, define the closed Euclidean ball by

$$\mathbb{B}^d(0, r) := \{x \in \mathbb{R}^d : \|x\|_2 \le r\}, \qquad \mathbb{B}^d := \mathbb{B}^d(0, 1).$$

For a set $\Theta \subseteq \mathbb{R}^{d_x \times d}$, define its Frobenius diameter by

$$\operatorname{diam}_{\mathrm{F}}(\Theta) := \sup_{\theta,\theta' \in \Theta} \|\theta - \theta'\|_{\mathrm{F}}.$$

For $1 \le p < \infty$ and a real-valued random variable X with $\mathbb{E}[|X|^p] < \infty$, define its L^p norm by $\|X\|_{L^p} := (\mathbb{E}[|X|^p])^{1/p}$. For a random vector $X \in \mathbb{R}^m$, define its L^p norm by $\|X\|_{L^p} := (\mathbb{E}[\|X\|_2^p])^{1/p}$. For $a \in \mathbb{R}$, let $\lfloor a \rfloor$ denote $\max\{k \in \mathbb{Z} : k \le a\}$. On a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, for a random vector $X : \Omega \mapsto \mathbb{R}^m$, let $\sigma(X)$ denote $\sigma(X) := \big\{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathcal{B}(\mathbb{R}^m)\big\}$, where $\mathcal{B}(\mathbb{R}^m)$ is the Borel σ-algebra of $\mathbb{R}^m$. More generally, for random variables $\{X_i\}_{i \in \mathcal{I}}$ on the same probability space, where $\mathcal{I}$ is an index set, $\sigma(X_i : i \in \mathcal{I})$ denotes the smallest σ-algebra making every X_i measurable. Similarly, For any collection $\mathcal{C}$ of subsets of the sample space Ω, define $\sigma(\mathcal{C}) := \bigcap \{\mathscr{G} : \mathscr{G} \text{ is a } \sigma\text{-algebra on } \Omega,\ \mathcal{C} \subseteq \mathscr{G}\}$. For a positive integer n, let $\mathbb{S}^{n-1}$ denote the Euclidean unit sphere in $\mathbb{R}^n$: $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. For an open set $\mathcal{U} \subseteq \mathbb{R}^{d_x + d_u}$ and a function $v : \mathcal{U} \mapsto \mathbb{R}^{d_x}$, we say that v is differentiable at $z \in \mathcal{U}$ if there exists a linear map $Dv(z) : \mathbb{R}^{d_x + d_u} \mapsto \mathbb{R}^{d_x}$ such that

$$\lim_{\substack{h \to 0 \\ h \neq 0}} \frac{\|v(z+h) - v(z) - Dv(z)(h)\|_2}{\|h\|_2} = 0.$$

This map is unique and is called the total derivative. Its matrix representation is

$$Dv(z) = \left[\frac{\partial v_a(z)}{\partial z_b}\right]_{1 \le a \le d_x,\ 1 \le b \le d_x + d_u}.$$

For an open set $\mathcal{U} \subseteq \mathbb{R}^{d_x + d_u}$ and a map $f : \mathcal{U} \to \mathbb{R}^{d_x}$, let $\mathbf{D}^0 f := f$. For $i \ge 1$, the ith total derivative of f at $z \in \mathcal{U}$ is the i-linear map

$$\mathbf{D}^i f(z) : (\mathbb{R}^{d_x + d_u})^i \to \mathbb{R}^{d_x},$$

defined by

$$\mathbf{D}^i f(z)[h_1, \ldots, h_i] := \frac{\partial^i}{\partial a_1 \cdots \partial a_i} f\left(z + \sum_{j=1}^{i} a_j h_j\right)\Bigg|_{a_1 = \cdots = a_i = 0}.$$

In particular, $\mathbf{D} f(z)$ is the Jacobian matrix of f at z. For $r \in \mathbb{N}$, $C^r(\mathcal{U}; \mathbb{R}^{d_x})$ denotes the class of functions whose partial derivatives of all orders up to r exist and are continuous on $\mathcal{U}$; C^0 denotes continuous functions. For $v \in C^r(\mathcal{U}; \mathbb{R}^{d_x})$ and $1 \le j \le r$, the jth total derivative is defined by

$$D^j v(z)[h_1, \ldots, h_j] := \frac{\partial^j}{\partial a_1 \cdots \partial a_j} v\left(z + \sum_{\ell=1}^{j} a_\ell h_\ell\right)\Bigg|_{a_1 = \cdots = a_j = 0}, \quad h_\ell \in \mathbb{R}^{d_x + d_u}, \quad a_\ell \in \mathbb{R}.$$

Set $D^0 v := v$. For a j-linear map $A : (\mathbb{R}^{d_x + d_u})^j \to \mathbb{R}^{d_x}$, define

$$\|A\|_{\mathrm{op}} := \sup_{\substack{h_1, \ldots, h_j \in \mathbb{R}^{d_x + d_u} \\ \|h_1\|_2, \ldots, \|h_j\|_2 \le 1}} \|A[h_1, \ldots, h_j]\|_2, \qquad j \ge 1,$$

and use the convention $\|D^0 v(z)\|_{\mathrm{op}} := \|v(z)\|_2$. Let $\mathcal{K} \subseteq \mathbb{R}^{d_x + d_u}$ be nonempty and compact, and let $L \ge 0$. For $r \in \mathbb{N}$, define

$$C^r(\mathcal{K}, L) := \bigcup_{\substack{\mathcal{U} \supseteq \mathcal{K} \\ \mathcal{U} \text{ open in } \mathbb{R}^{d_x + d_u}}} \Big\{ v|_{\mathcal{K}} :\ v \in C^r(\mathcal{U}; \mathbb{R}^{d_x}),\ \max_{0 \le j \le r} \sup_{z \in \mathcal{K}} \|D^j v(z)\|_{\mathrm{op}} \le L \Big\}.$$

Here $v|_{\mathcal{K}}$ denotes the restriction of v to $\mathcal{K}$. For smoothness order $s > 0$, define the Hölder class $\mathcal{H}^s(\mathcal{K}, L)$ as

$$\bigcup_{\substack{\mathcal{U} \supseteq \mathcal{K} \\ \mathcal{U} \text{ open in } \mathbb{R}^{d_x + d_u}}} \Bigg\{ v|_{\mathcal{K}} : v \in C^{\lceil s \rceil - 1}(\mathcal{U}; \mathbb{R}^{d_x}),$$

$$\max_{\substack{j \in \mathbb{N} \\ 0 \le j < s}} \sup_{z \in \mathcal{K}} \|D^j v(z)\|_{\mathrm{op}} + \sup_{\substack{z, z' \in \mathcal{K} \\ z \neq z'}} \frac{\|D^{\lceil s \rceil - 1} v(z) - D^{\lceil s \rceil - 1} v(z')\|_{\mathrm{op}}}{\|z - z'\|_2^{s - \lceil s \rceil + 1}} \le L \Bigg\}. \tag{17}$$

The last supremum is defined to be zero when $\mathcal{K}$ contains only one point. For a finite ordered collection of vectors $v_1, \ldots, v_m$, define

$$\mathrm{col}(v_1, \ldots, v_m) := \begin{bmatrix} v_1^\top & \cdots & v_m^\top \end{bmatrix}^\top.$$

In particular, if each v_j is a scalar, then

$$\mathrm{col}(v_1, \ldots, v_m) = (v_1, \ldots, v_m)^\top.$$

Definition 1 (Exponential incremental input-to-state stability)**.** Let $F : \mathbb{R}^{d_x} \times \mathbb{N} \to \mathbb{R}^{d_x}$. The dynamics $x_{t+1} = F(x_t, t)$ are exponentially incrementally input-to-state stable if there exist $\beta \geq 1$, $\gamma > 0$, and $\rho \in (0, 1)$ such that, for every $x_0, y_0 \in \mathbb{R}^{d_x}$ and every input sequence $\{v_t\}_{t \geq 0} \subseteq \mathbb{R}^{d_x}$, the trajectories

$$x_{t+1} = F(x_t, t) + v_t, \qquad y_{t+1} = F(y_t, t)$$

satisfy

$$\|x_t - y_t\|_2 \leq \beta \rho^t \|x_0 - y_0\|_2 + \gamma \sum_{j=0}^{t-1} \rho^{t-1-j} \|v_j\|_2, \qquad t \geq 1.$$

Definition 2 (Łojasiewicz gradient condition)**.** Let $\mathcal{L}_{\pi_{\mathrm{id}}}(\theta)$ denote the population identification loss induced by the initial policy π_{id}. Suppose that $\mathcal{L}_{\pi_{\mathrm{id}}}$ is continuously differentiable on an open neighborhood U of the true parameter $\theta^\star$, and that $\theta^\star$ minimizes $\mathcal{L}_{\pi_{\mathrm{id}}}$ on U. We say that $\mathcal{L}_{\pi_{\mathrm{id}}}$ satisfies the local Łojasiewicz gradient condition at $\theta^\star$ if there exist constants $c > 0$, $\alpha \in (0, 1)$, and an open neighborhood $V \subseteq U$ of $\theta^\star$ such that

$$\left|\mathcal{L}_{\pi_{\mathrm{id}}}(\theta) - \mathcal{L}_{\pi_{\mathrm{id}}}(\theta^\star)\right|^\alpha \leq c \left\|\nabla_\theta \mathcal{L}_{\pi_{\mathrm{id}}}(\theta)\right\|_2, \qquad \forall\, \theta \in V. \tag{18}$$

Definition 3 ((κ, γ)-strong stability)**.** Consider the discrete-time linear system $x_{t+1} = Ax_t + Bu_t + w_t$ under the feedback law $u_t = Kx_t$. Let $\kappa \geq 1$ and $\gamma \in (0, 1]$. A feedback gain $K \in \mathbb{R}^{d_u \times d_x}$ is said to be (κ, γ)-strongly stable for (A, B) if $\|K\|_2 \leq \kappa$ and there exist matrices $H, L \in \mathbb{R}^{d_x \times d_x}$, with H invertible, such that

$$A + BK = HLH^{-1}, \qquad \|L\|_2 \leq 1 - \gamma, \qquad \|H\|_2 \, \|H^{-1}\|_2 \leq \kappa, \tag{19}$$

where $\|\cdot\|_2$ denotes the spectral norm.

A.2 Related work

Learning for control. Recently, there has been growing interest in understanding control from a learning-theoretic perspective. The manifold approaches to learning for control can be broadly categorized as indirect (i.e., based on system identification followed by model-based control design) versus direct (which bypass the system identification step entirely). In fact, the indirect controller design for Linear-Quadratic Regulator (LQR) has a rich history with well-understood tools for identification and control, while representative approaches including optimism-in-face-of-uncertainty (Dean et al., 2020), certainty-equivalence control (Simchowitz et al., 2018; Mania et al., 2019; Wang & Janson, 2021), and adaptive dynamic programming (Bian & Jiang, 2016). In contrast, indirect controller design for nonlinear systems is a relatively emerging research direction with two approaches. One approach is to completely decouple controller design into two independent steps: nonlinear system identification (see Musavi et al. (2024); Park et al. (2025); Mania et al. (2022); Sattar & Oymak (2022)) followed by model-based optimal control (see He et al. (2023); Agarwal et al. (2020); Sun et al. (2019); Song & Sun (2021) for model-based RL). The other approach alternates between system identification and model-based optimal control, continuously enhancing estimation accuracy and control performance; see Muehlebach et al. (2026); Wagenmaker et al. (2023); Annaswamy et al. (2023); Vinod et al. (2022); Zhang et al. (2026); Bayraktar & Chen (2023); Cohen et al. (2025); Chen et al. (2022). Different from the indirect design, direct methods entirely bypass (known as an "end-to-end" method) intermediate modeling errors and directly optimizes control performance using sample trajectories. An emerging line of direct methods obtains the LQR directly from a single episode of persistently exciting data; see De Persis & Tesi (2019); Zhao et al. (2025); Dörfler et al. (2023); De Persis & Tesi (2021). However, direct methods for nonlinear systems remain a nascent field, with existing results being relatively restrictive and largely limited to extensions of linear control paradigms; see Han et al. (2025); Gao et al. (2026).

B AUXILIARY RESULTS

In this appendix, we first present the formal version of Algorithm GR−MBAC with the necessary definitions and Table 1 for literature comparison. Then, we establish a high-probability guarantee for finding a stabilizing gain and derive the corresponding Lyapunov inequality. Finally, we introduce the definitions and an auxiliary lemma of the block martingale small-ball (BMSB) condition that are needed for the subsequent analysis.

B.1 FORMAL VERSION OF ALGORITHM GR−MBAC

In this subsection, we present the formal version of GR−MBAC. The algorithm includes three computationally challenging steps: LearnCertify, CertifyGain, and FiniteHorizonSolve. We discuss their implementations in Sections B.1.1, B.1.2, and B.1.3, respectively. Before presenting the formal algorithm, we introduce the following definition for simplifying statement.

Definition 4. For a nonempty compact set $T \subseteq \mathbb{R}^D$ with Euclidean metric d, let

$$\mathcal{M}(T, d, \varepsilon) := \max\big\{|A| : A \subseteq T,\ d(a, a') > \varepsilon \text{ for distinct } a, a' \in A\big\}$$

be the ε-packing number of T.

We now introduce the specific realization of $\mathrm{GreedyCover}(T, a_0, \varepsilon)$. Given $a_0 \in T$, set $A_0 = \{a_0\}$. For $r = 0, 1, \ldots$, choose

$$a_{r+1} \in \operatorname*{argmax}_{a \in T} d(a, A_r) \text{ with } d(a, A_r) := \min_{b \in A_r} d(a, b).$$

If $d(a_{r+1}, A_r) \le \varepsilon$, stop; otherwise set $A_{r+1} := A_r \cup \{a_{r+1}\}$. The output $\mathrm{GreedyCover}(T, a_0, \varepsilon)$ is the terminal set A_r. Every added point of A_r has distance greater than ε from all previously selected points. Hence the output A_r of $\mathrm{GreedyCover}(T, a_0, \varepsilon)$ satisfies

$$a_0 \in A_r, \qquad \sup_{a \in T} d(a, A_r) \le \varepsilon, \qquad |A_r| \le \mathcal{M}(T, d, \varepsilon).$$

We now introduce the formal version of Algorithm GR−MBAC.

Algorithm 1 Gibbs-Randomized Model-Based Adaptive Control (GR−MBAC)

Require: sieve specifications $(\phi^{(t)}, \Theta_t)$, update period $2M$, inverse temperatures η_t, residual radius B_{ctl}, horizon N, exploration horizon $\widetilde{N}$, and contraction threshold q_1, exploration amplitudes $\{\sigma_{\mathrm{id},t}\}_{t\ge 0}$, independent exploration vectors $\{\zeta_t^{\mathrm{id}}\}_{t\ge 0} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, I_{d_u})$, an exploration policy $\{\pi_n\}_{n=0}^{N_u}$, cover radii $\{\varepsilon_{\mathrm{cov},t}\}_{t\ge 0}$.

Initialize: observe $\widetilde{x}_0$ and set $\tau = N$.
for $n = 1, \ldots, \widetilde{N}$ **do**
  Explore: sample $\widetilde{u}_{n-1}$ from policy π_n as the current action and observe new state $\widetilde{x}_n$.
  LearnCertify: using $\{(\widetilde{x}_i, \widetilde{u}_i, \widetilde{x}_{i+1})\}_{i=0}^{n-1}$, compute a model estimate g_n of f and confidence radius r_n satisfying (21).
  CertifyGain: using $g_n, r_n, \mathfrak{S}, \widetilde{\epsilon}_n$, compute (K_n, P_n, U_n) satisfying (28).
  if $U_n \le q_1$ **then**
    Set $\tau = n$, $\widehat{\mathbf{K}} = K_n$ and $\widehat{P} = P_n$, and **break**.
  end if
end for
for epoch $t = 0, 1, 2, \ldots, H-1$ **do**
  if $\mathrm{mod}(t, 2M) = 0$ and $t \ne 0$ **then**
    SieveFit: set $f_\theta(x, u) = f^{\mathrm{pub}}(x, u) + \theta\phi^{(t)}(x, u)$ and $s_t(\theta) = \sum_{i=0}^{t-M-2} \|x_{i+1} - f_\theta(x_i, u_i)\|_2^2$. Compute $\widehat{\theta}_t \in \arg\min_{\theta \in \Theta_t} s_t(\theta)$.
    GreedyCover: set $\mathcal{C}_t = \mathrm{GreedyCover}(\Theta_t, \widehat{\theta}_t, \varepsilon_{\mathrm{cov},t})$ as defined in Definition 4.
    GibbsSample: define the discrete Gibbs law $p_t(\theta) = \exp\{-\eta_t s_t(\theta)\} / \sum_{\vartheta \in \mathcal{C}_t} \exp\{-\eta_t s_t(\vartheta)\}$, $\theta \in \mathcal{C}_t$. Sample $\theta_t \sim p_t$.

15: `CovarianceUpdate`: compute, with $k' = t/(2M)$,

$$\widehat{\Sigma}_{w,t} = \frac{k'-1}{k'}\widehat{\Sigma}_{w,t-1} + \frac{1}{k'M}\sum_{i=t-M-1}^{t-2} r_{t,i} r_{t,i}^\top,$$
$$r_{t,i} := x_{i+1} - f_{\theta_t}(x_i, u_i).$$

16: `FiniteHorizonSolve`: solve the following finite-horizon stochastic control problem:

$$\begin{aligned} \min_{\mu=\{\mu^{(i)}\}_{i=0}^{H-1}} \quad & \mathcal{J}_t^{\mathrm{mod}}(\mu) := \mathbb{E}_\tau\left[\frac{1}{H}\sum_{i=0}^{H-1} c\big(x_i, \widehat{\mathbf{K}}x_i + v_i\big) + h(x_H)\right] \\ \text{s.t.} \quad & x_0 \sim \nu_0, \\ & \widehat{w}_i \sim \mathcal{N}(0, \widehat{\Sigma}_{w,t}), && i = 0, \dots, H-1, \\ & v_i = \mathrm{Proj}_{\mathbb{B}^{d_u}(0, B_{\mathrm{ctl}})}\big(\mu^{(i)}(x_0, u_0, \dots, x_i)\big), && i = 0, \dots, H-1, \\ & x_{i+1} = f_{\theta_t}\big(x_i, \widehat{\mathbf{K}}x_i + v_i\big) + \widehat{w}_i, && i = 0, \dots, H-1. \end{aligned} \tag{20}$$

Get an optimal policy μ_{θ_t} of problem (20) and compute the clipped action $u_{\mathrm{opt},t} = \mathrm{Proj}_{\mathbb{B}^{d_u}(0,B_{\mathrm{ctl}})}\big(\mu_{\theta_t}^{(t)}(x_0, u_0, \dots, x_t)\big)$.

17: **else**

18: `Hold`: keep

$$\theta_t = \theta_{t-1},\ \widehat{\Sigma}_{w,t} = \widehat{\Sigma}_{w,t-1},\ \mu_{\theta_t} = \mu_{\theta_{t-1}},$$
$$u_{\mathrm{opt},t} = \mathrm{Proj}_{\mathbb{B}^{d_u}(0,B_{\mathrm{ctl}})}\big(\mu_{\theta_t}^{(t)}(x_0, u_0, \dots, x_t)\big).$$

19: **end if**

20: `Deploy`: sample $\zeta_t^{\mathrm{id}} \sim \mathcal{N}(0, I_{d_u})$ independently, set $e_t^{\mathrm{id}} = \sigma_{\mathrm{id},t}\zeta_t^{\mathrm{id}}$, and apply $u_t = \widehat{\mathbf{K}}x_t + u_{\mathrm{opt},t} + e_t^{\mathrm{id}}$.

21: **end for**

B.1.1 LEARNCERTIFY STEP

At exploration time n, `LearnCertify` uses $\{(\widetilde{x}_i, \widetilde{u}_i, \widetilde{x}_{i+1})\}_{i=0}^{n-1}$ to compute a model g_n and a confidence radius r_n. Write $d = d_x + d_u$, $z = (x^\top, u^\top)^\top$, and $\widetilde{z}_i = (\widetilde{x}_i^\top, \widetilde{u}_i^\top)^\top$. Fix a regularization parameter $\lambda > 0$, a confidence level $\delta_{\mathrm{conf}} \in (0,1)$, and a known noise bound $\Sigma_w \preceq \bar{\sigma}^2 I_{d_x}$. Using the constants in (3), define $\widetilde{C} := \sqrt{\overline{v}(1+\mathscr{K}^2)/\underline{v}}$ and

$$\mathcal{E}_{\mathrm{conf}} := \Big\{\|g_n(z) - f(z)\|_2 \le r_n\|z\|_2, \quad \forall z \in \mathbb{R}^d,\ 0 \le n \le \widetilde{N}\Big\}.$$

We require $\{(g_n, r_n, \widetilde{\epsilon}_n)\}_{n=0}^{\widetilde{N}}$ to satisfy

$$\mathbb{P}(\mathcal{E}_{\mathrm{conf}}) \ge 1 - \delta_{\mathrm{conf}}, \qquad 2\widetilde{C}r_{\widetilde{N}} + \widetilde{\epsilon}_{\widetilde{N}} \le q_2, \tag{21}$$

where $\{\widetilde{\epsilon}_n\}_{n=0}^{\widetilde{N}}$ is the optimization tolerance used by `CertifyGain`. Only the terminal tolerance $\widetilde{\epsilon}_{\widetilde{N}}$ is required to satisfy the condition in (21) for the stability analysis in Lemma 1.

We now introduce some realization methods of `LearnCeertify` for different systems. For linear system, `LearnCertify` can be implemented using regularized least squares and a confidence bound valid simultaneously over time (Abbasi-Yadkori et al., 2011; Lale et al., 2022). Let $f(x,u) = Ax + Bu = \Theta_\star z$ with $\Theta_\star = [A\ B]$ and a known bound $\|\Theta_\star\|_{\mathrm{F}} \le S$, compute the regularized least-squares estimate and its confidence radius as

$$\begin{aligned} & V_n := \lambda I_d + \sum_{i=0}^{n-1} \widetilde{z}_i \widetilde{z}_i^\top, \qquad \widehat{\Theta}_n := \left(\sum_{i=0}^{n-1} \widetilde{x}_{i+1}\widetilde{z}_i^\top\right) V_n^{-1}, \\ & \beta_n := \bar{\sigma}\sqrt{2d_x \log\left(\frac{d_x \det(V_n)^{1/2}}{\delta_{\mathrm{conf}}\lambda^{d/2}}\right)} + \sqrt{\lambda}S, \\ & g_n(z) := \widehat{\Theta}_n z, \qquad r_n := \frac{\beta_n}{\sqrt{\lambda_{\min}(V_n)}}. \end{aligned} \tag{22}$$

Self-normalized martingale concentration (Abbasi-Yadkori et al., 2011, Theorem 1) gives $\|\widehat{\Theta}_n - \Theta_\star\|_{\mathrm{op}} \le r_n$ simultaneously over the exploration stage with probability at least $1 - \delta_{\mathrm{conf}}$. This establishes the confidence requirement in (21). For the general nonlinear systems, we use weighted ridge regression with an explicit approximation error term (Takemori & Sato, 2021; Bogunovic & Krause, 2021) to implement `LearnCertify`. let f^{pub} be a known component of the dynamics. For each $m \ge 1$, choose a deterministic sieve map $\psi_m : \mathbb{R}^d \to \mathbb{R}^m$ and assume f has the following form:

$$\begin{aligned} &f(z) = f^{\mathrm{pub}}(z) + \Theta_m^\star \psi_m(z) + h_m(z), \\ &\|h_m(z)\|_2 \le a_m \|z\|_2, \qquad \|\psi_m(z)\|_2 \le C_m \|z\|_2, \qquad \|\Theta_m^\star\|_{\mathrm{F}} \le S_m, \qquad \forall z \in \mathbb{R}^d, \end{aligned} \tag{23}$$

where $a_m \ge 0$, $C_m > 0$, and $S_m \ge 0$ are known. Normalize each observation by setting

$$\rho_i := \sqrt{1 + \|\widetilde{z}_i\|_2^2}, \qquad \varphi_{m,i} := \frac{\psi_m(\widetilde{z}_i)}{\rho_i}, \qquad y_i := \frac{\widetilde{x}_{i+1} - f^{\mathrm{pub}}(\widetilde{z}_i)}{\rho_i}, \qquad i = 0, \dots, n-1.$$

Choose $\delta_m = \delta_{\mathrm{conf}}/[2m(m+1)]$, so that $\sum_{m \ge 1} \delta_m = \delta_{\mathrm{conf}}/2$, and compute

$$\begin{aligned} &V_{m,n} := \lambda I_m + \sum_{i=0}^{n-1} \varphi_{m,i} \varphi_{m,i}^\top, \qquad \widehat{\Theta}_{m,n} := \left(\sum_{i=0}^{n-1} y_i \varphi_{m,i}^\top \right) V_{m,n}^{-1}, \\ &\beta_{m,n} := \bar{\sigma} \sqrt{2 d_x \log \left(\frac{d_x \det(V_{m,n})^{1/2}}{\delta_m \lambda^{m/2}} \right)} + \sqrt{\lambda} S_m, \\ &g_{m,n}(z) := f^{\mathrm{pub}}(z) + \widehat{\Theta}_{m,n} \psi_m(z), \qquad r_{m,n} := a_m + \frac{C_m(\beta_{m,n} + a_m \sqrt{n})}{\sqrt{\lambda_{\min}(V_{m,n})}}. \end{aligned} \tag{24}$$

Self-normalized concentration and a union bound over m yield

$$\begin{aligned} &\mathcal{E}_{\mathrm{ridge}} := \left\{ \|g_{m,n}(z) - f(z)\|_2 \le r_{m,n} \|z\|_2, \quad \forall z,\ m \ge 1,\ 0 \le n \le \widetilde{N} \right\}, \\ &\mathbb{P}(\mathcal{E}_{\mathrm{ridge}}) \ge 1 - \delta_{\mathrm{conf}}/2. \end{aligned}$$

We now give an initial excitation condition and a condition of $\widetilde{N}$ to guarantee that $\{(g_{m,n}, r_n, \widetilde{\epsilon}_n)\}_{n=0}^{\widetilde{N}}$ (the choices of r_n, $\widetilde{N}$, and $\widetilde{\epsilon}_n$ are shown in (27)) satisfy (21). Fix the startup dimension m and explore with independent inputs $\widetilde{u}_i \sim \mathcal{N}(0, \sigma_{\mathrm{exp}}^2 I_{d_u})$, where $\sigma_{\mathrm{exp}} > 0$. Let $\mathscr{A}_i = \sigma(\widetilde{x}_0, \widetilde{u}_0, \dots, \widetilde{u}_{i-1}, \widetilde{x}_i)$. Suppose there are known $\ell \in \mathbb{N}_+$ and $\gamma_m > 0$ satisfying the initial excitation condition

$$\mathbb{E}\left[\sum_{i=j\ell}^{(j+1)\ell - 1} \varphi_{m,i} \varphi_{m,i}^\top \,\middle|\, \mathscr{A}_{j\ell} \right] \succeq \ell \gamma_m I_m, \qquad j \ge 0. \tag{25}$$

Choose $\bar{\epsilon} \in (0, q_2)$ and define

$$\begin{aligned} &\overline{\beta}_m(n) := \bar{\sigma} \sqrt{2 d_x \left[\log \frac{d_x}{\delta_m} + \frac{m}{2} \log \left(1 + \frac{n C_m^2}{m\lambda} \right) \right]} + \sqrt{\lambda} S_m, \\ &b_m := a_m \left(1 + C_m \sqrt{\frac{2}{\gamma_m}} \right), \qquad R_m(n) := b_m + C_m \sqrt{\frac{2}{n\gamma_m}}\, \overline{\beta}_m(n), \qquad n \ge 1. \end{aligned}$$

Select m with a positive accuracy margin $\Delta_m := (q_2 - \bar{\epsilon})/(2\widetilde{C}) - b_m > 0$. Define

$$\begin{aligned} &A_m := \frac{4 C_m^2}{\gamma_m \Delta_m^2} \left[\bar{\sigma}^2 d_x \left(m + 2 \log \frac{d_x}{\delta_m} \right) + \lambda S_m^2 \right], \\ &N_{\mathrm{Gram}} := \frac{8 \ell C_m^4}{\gamma_m^2} \left(m \log 9 + \log \frac{4}{\delta_{\mathrm{conf}}} \right), \qquad N_{\mathrm{stat}} := 2 A_m \left[1 + \log \left(1 + \frac{2 C_m^2 A_m}{m\lambda} \right) \right]. \end{aligned}$$

The condition of $\widetilde{N}$ is shown as follow

$$\widetilde{N} \in \ell \mathbb{N}_+, \qquad \widetilde{N} \ge \max\{N_{\mathrm{Gram}}, N_{\mathrm{stat}}\}. \tag{26}$$

Finally, we set $g_n = g_{m,n}$ with $0 \le n \le \widetilde{N}$ and

$$r_n = \begin{cases} r_{m,n}, & 0 \le n < \widetilde{N}, \\ R_m(\widetilde{N}), & n = \widetilde{N}, \end{cases} \qquad \widetilde{\epsilon}_{\widetilde{N}} \le \bar{\epsilon}. \tag{27}$$

This implements `LearnCertify` for the general nonlinear system in (23).

Proposition 2. *Under the nonlinear sieve and noise bounds above,* (25)*, and* $\Delta_m > 0$*, the preceding* `LearnCertify` *implementation satisfies* (21) *for every budget obeying* (26)*.*

Proof. Write $n = \ell J$, $Y_j = \sum_{i=j\ell}^{(j+1)\ell-1} \varphi_{m,i}\varphi_{m,i}^\top$, and $D_j = Y_j - \mathbb{E}[Y_j \mid \mathscr{A}_{j\ell}]$. For each unit vector v, $0 \le v^\top Y_j v \le \ell C_m^2$. Conditional Hoeffding concentration gives $\mathbb{P}(|\sum_{j<J} v^\top D_j v| \ge n\gamma_m/4) \le 2\exp[-J\gamma_m^2/(8C_m^4)]$. A $1/4$-net of the unit sphere has at most 9^m points and controls the norm of a symmetric matrix by twice its largest absolute quadratic form on the net. Hence (25) implies

$$\mathbb{P}\left(\lambda_{\min}(V_{m,n}) < \lambda + n\gamma_m/2\right) \le 2\,9^m \exp\left(-\frac{J\gamma_m^2}{8C_m^4}\right).$$

At $n = \widetilde{N} \ge N_{\mathrm{Gram}}$, this probability is at most $\delta_{\mathrm{conf}}/2$. Moreover, $\mathrm{tr}(V_{m,n}) \le m\lambda + nC_m^2$ gives $\det(V_{m,n})/\lambda^m \le (1 + nC_m^2/(m\lambda))^m$ and hence $\beta_{m,n} \le \overline{\beta}_m(n)$. Note that

$$r_{m,\widetilde{N}} \le a_m + C_m\sqrt{\frac{2}{\gamma_m}}\left(a_m + \frac{\overline{\beta}_m(\widetilde{N})}{\sqrt{\widetilde{N}}}\right) = R_m(\widetilde{N}).$$

Define

$$\mathcal{E}_{\mathrm{Gram}} := \left\{\lambda_{\min}(V_{m,\widetilde{N}}) \ge \lambda + \frac{\widetilde{N}\gamma_m}{2}\right\}.$$

The preceding bound gives $\mathbb{P}(\mathcal{E}_{\mathrm{Gram}}^c) \le \delta_{\mathrm{conf}}/2$. On $\mathcal{E}_{\mathrm{ridge}} \cap \mathcal{E}_{\mathrm{Gram}}$, for every $z \in \mathbb{R}^d$,

$$\begin{aligned} &\|g_n(z) - f(z)\|_2 \le r_{m,n}\|z\|_2 = r_n\|z\|_2, && 0 \le n < \widetilde{N}, \\ &\|g_{\widetilde{N}}(z) - f(z)\|_2 \le r_{m,\widetilde{N}}\|z\|_2 \le R_m(\widetilde{N})\|z\|_2 = r_{\widetilde{N}}\|z\|_2. \end{aligned}$$

Consequently, it holds that

$$\mathcal{E}_{\mathrm{ridge}} \cap \mathcal{E}_{\mathrm{Gram}} \subseteq \mathcal{E}_{\mathrm{conf}}, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{conf}}) \ge 1 - \frac{\delta_{\mathrm{conf}}}{2} - \frac{\delta_{\mathrm{conf}}}{2} = 1 - \delta_{\mathrm{conf}}.$$

To verify the second inequality of (21), set $c_m = C_m^2/(m\lambda)$. Using $(a+b)^2 \le 2a^2 + 2b^2$ in $\overline{\beta}_m(n)^2$ yields

$$\frac{2C_m^2\overline{\beta}_m(n)^2}{\gamma_m\Delta_m^2} \le A_m[1 + \log(1 + c_m n)].$$

If $A_m > 0$ and $n \ge N_{\mathrm{stat}}$, then $t = n/(2A_m) \ge 1 + L$, where $L = \log(1 + 2c_m A_m)$. Since $t \ge 1$ and $\log t \le t - 1$, we have

$$1 + \log(1 + c_m n) \le 1 + L + \log t \le L + t \le 2t.$$

Consequently $A_m[1 + \log(1 + c_m n)] \le n$. If $A_m = 0$, then $\overline{\beta}_m(n) = 0$ and this conclusion holds directly. At $n = \widetilde{N}$, we obtain $R_m(\widetilde{N}) \le b_m + \Delta_m = (q_2 - \bar{\epsilon})/(2\widetilde{C})$. The terminal return rule and $\widetilde{\epsilon}_{\widetilde{N}} \le \bar{\epsilon}$ now give the second inequality in (21). □

B.1.2 CertifyGain step

For $(K, P) \in \mathfrak{S}$ and $0 \le n \le \widetilde{N} \wedge \tau$, define

$$\Phi_n(K, P) := \sup_{x \ne 0} \frac{\|g_n(x, Kx)\|_P}{\|x\|_P} + r_n\sqrt{\frac{\overline{v}(1 + \|K\|_{\mathrm{op}}^2)}{\underline{v}}}.$$

For a prespecified numerical tolerance $\widetilde{\epsilon}_n > 0$, `CertifyGain` returns a pair $(K_n, P_n) \in \mathfrak{S}$ and an upper bound U_n satisfying

$$\Phi_n(K_n, P_n) \le U_n \le \inf_{(K,P)\in\mathfrak{S}} \Phi_n(K, P) + \widetilde{\epsilon}_n. \tag{28}$$

The `CertifyGain` step can be implemented by a finite covering search or a branch-and-bound method with certified numerical bounds (Balakrishnan et al., 1991; Neumaier, 2004). For linear fitted models $g_n(x, u) = \widehat{A}_n x + \widehat{B}_n u$, the inner supremum has the explicit expression

$$\sup_{x\neq 0} \frac{\|g_n(x, Kx)\|_P}{\|x\|_P} = \left\| P^{1/2}(\widehat{A}_n + \widehat{B}_n K)P^{-1/2} \right\|_{\mathrm{op}}. \tag{29}$$

Thus, evaluating $\Phi_n(K, P)$ reduces to computing matrix norms with certified numerical error bounds. For nonlinear fitted models, consider

$$g_n(x, u) = \widehat{A}_n x + \widehat{B}_n u + h_n(x, u),$$

where $h_n(0, 0) = 0$ and h_n has known global bounds on its magnitude and its first two derivatives (e.g., finite Fourier sieves satisfy these conditions). Suppose that $\|h_n(z)\|_2 \le B_n$ and $\|D^2 h_n(z)\|_{\mathrm{op}} \le H_n$ for all z, and write $\kappa = \overline{v}/\underline{v}$. For small states, Taylor expansion approximates $g_n(x, Kx)$ by $J_n(K)x$, where $J_n(K) = D_x g_n(0, 0) + D_u g_n(0, 0)K$. For large states, the bounded nonlinear term can be controlled relative to the linear part. Specifically, the errors in $\frac{\|g_n(x,Kx)\|_P}{\|x\|_P}$ are bounded by

$$\begin{aligned} &\frac{\sqrt{\kappa} H_n(1 + \mathscr{K}^2)}{2}\, a, \qquad 0 < \|x\|_2 \le a, \\ &\frac{\sqrt{\kappa} B_n}{R}, \qquad \|x\|_2 \ge R, \end{aligned} \tag{30}$$

respectively. The remaining maximization over $a \le \|x\|_2 \le R$ can be solved using interval branch-and-bound (Neumaier, 2004). Decreasing a, increasing R, and refining the interval partition therefore yield an upper bound on $\Phi_n(K, P)$ with any prespecified positive accuracy. Using these bounds, we can also compute a Lipschitz constant $L_{\Phi,n}$ for Φ_n on $\mathfrak{S}$. For this purpose, define

$$G_n := \sup_{z\in\mathbb{R}^{d_x+d_u}} \|Dg_n(z)\|_{\mathrm{op}}, \qquad G_{u,n} := \sup_{z\in\mathbb{R}^{d_x+d_u}} \|D_u g_n(z)\|_{\mathrm{op}}.$$

We obviously have

$$G_n \le \left\| \begin{bmatrix} \widehat{A}_n & \widehat{B}_n \end{bmatrix} \right\|_{\mathrm{op}} + \sup_{z\in\mathbb{R}^{d_x+d_u}} \|Dh_n(z)\|_{\mathrm{op}}, \quad G_{u,n} \le \|\widehat{B}_n\|_{\mathrm{op}} + \sup_{z\in\mathbb{R}^{d_x+d_u}} \|D_u h_n(z)\|_{\mathrm{op}} \tag{31}$$

with

$$\sup_{z\in\mathbb{R}^{d_x+d_u}} \|Dh_n(z)\|_{\mathrm{op}} \le \inf_{a>0} \left(\frac{B_n}{a} + \frac{H_n a}{2} \right) = \sqrt{2B_n H_n}.$$

For $x \neq 0$, define $q_{n,K,P}(x) := \frac{\|g_n(x,Kx)\|_P}{\|x\|_P}$. Since $g_n(0, 0) = 0$ and $\|K\|_{\mathrm{op}} \le \mathscr{K}$ on $\mathfrak{S}$, we have $\frac{\|g_n(x,Kx)\|_2}{\|x\|_2} \le G_n\sqrt{1 + \mathscr{K}^2}$. Define $\mathfrak{S}_P := \{P \in \mathbb{R}^{d_x\times d_x} : \exists K \in \mathbb{R}^{d_u\times d_x}, \text{s.t. } (K, P) \in \mathfrak{S}\}$ and $\mathfrak{S}_K := \{K \in \mathbb{R}^{d_x\times d_u} : \exists P \in \mathbb{R}^{d_x\times d_x}, \text{s.t. } (K, P) \in \mathfrak{S}\}$. For fixed $P \in \mathfrak{S}_P$, any $K, K' \in \mathbb{R}^{d_u\times d_x}$, and $x \in \mathbb{R}^{d_x}$, we have

$$|q_{n,K,P}(x) - q_{n,K',P}(x)| \le \sqrt{\kappa}\, G_{u,n} \|K - K'\|_{\mathrm{F}}.$$

For fixed $K \in \mathfrak{S}_K$, any $P, P' \in \mathbb{R}^{d_x d_x}$, and $x \in \mathbb{R}^{d_x}$, we have

$$|q_{n,K,P}(x) - q_{n,K,P'}(x)| \le \frac{1 + \sqrt{\kappa}}{2\underline{v}}\, G_n\sqrt{1 + \mathscr{K}^2}\, \|P - P'\|_{\mathrm{F}}.$$

Moreover, it holds that for any $K, K' \in \mathbb{R}^{d_u\times d_x}$,

$$r_n\sqrt{\kappa} \left| \sqrt{1 + \|K\|_{\mathrm{op}}^2} - \sqrt{1 + \|K'\|_{\mathrm{op}}^2} \right| d \le r_n\sqrt{\kappa}\, \|K - K'\|_{\mathrm{F}}.$$

Taking the supremum of $q_{n,K,P}(x)$ over $x \neq 0$ and using the fact that $|\sup_{x\in\mathbb{R}^{d_x}} a(x) - \sup_{x\in\mathbb{R}^{d_x}} b(x)| \leq \sup_{x\in\mathbb{R}^{d_x}} |a(x) - b(x)|$ holds for any functions $a, b : \mathbb{R}^{d_x} \mapsto \mathbb{R}$, we have

$$L_{\Phi,n} := \max\left\{\sqrt{\kappa}\,(G_{u,n} + r_n),\ \frac{1+\sqrt{\kappa}}{2\underline{v}}\,G_n\sqrt{1+\mathscr{K}^2}\right\}, \tag{32}$$

such that

$$|\Phi_n(K,P) - \Phi_n(K',P')| \leq L_{\Phi,n}\big(\|K-K'\|_{\mathrm{F}} + \|P-P'\|_{\mathrm{F}}\big).$$

Choose a finite h_n-net $\{(K_j,P_j)\}_{j=1}^{J_n} \subseteq \mathfrak{S}$, and compute certified upper bounds $u_{n,j}$ satisfying

$$\Phi_n(K_j,P_j) \leq u_{n,j} \leq \Phi_n(K_j,P_j) + \eta_n.$$

Return the candidate with the smallest upper bound:

$$j_n \in \operatorname*{arg\,min}_{1\leq j\leq J_n} u_{n,j}, \qquad (K_n,P_n) = (K_{j_n},P_{j_n}), \qquad U_n = u_{n,j_n}, \qquad j = 1,\ldots,J_n. \tag{33}$$

Then

$$\Phi_n(K_n,P_n) \leq U_n \leq \inf_{(K,P)\in\mathfrak{S}} \Phi_n(K,P) + L_{\Phi,n}h_n + \eta_n.$$

Consequently, choosing $L_{\Phi,n}h_n + \eta_n \leq \widetilde{\epsilon}_n$ ensures (28).

B.1.3 FiniteHorizonSolve Step

Appendix E establishes convergence to the optimal objective value for the continuous-time policy gradient flow. Fix a model-update epoch s and consider $\mathcal{J}_s^{\mathrm{mod}}(\mu)$ in (20) with f_{θ_s} and $\widehat{\Sigma}_{w,s} \succ 0$ fixed. We use deterministic Markov residual policies $\mu = \{\mu^{(i)}\}_{i=0}^{H-1}$ satisfying $\mu^{(i)}(x) \in \mathbb{B}^{d_u}(0, B_{\mathrm{ctl}})$. Let $r \geq 0$ denote the optimization time and consider

$$\frac{\mathrm{d}}{\mathrm{d}r}\mu_r^{(i)}(x) = -\frac{\delta\mathcal{J}_s^{\mathrm{mod}}}{\delta\mu^{(i)}}(\mu_r)(x), \qquad i = 0,\ldots,H-1. \tag{34}$$

Theorem 4 in Appendix E shows that, under Assumptions 8, it holds that

$$\lim_{r\to\infty} \mathcal{J}_s^{\mathrm{mod}}(\mu_r) = \mathcal{J}_s^{\mathrm{mod}}(\mu^*_{\theta_s}),$$

where $\mu^*_{\theta_s}$ denotes a global minimizer for (20). Hence, results in Appendix E motivate a numerical implementation of `FiniteHorizonSolve` by Euler discretization and DNN fitting. Specifically, we first discretize the policy gradient flow (34). Letting μ_k denote the current policy, we write Q_{μ_k} and $\rho_i^{\mu_k}$ for the action-value function and state density in Appendix E with $v(i,x) = \mu_k^{(i)}(x)$. By Lemma 4, the forward Euler update is

$$\begin{aligned}\mu_{k+1}^{(i)}(x) &= \mu_k^{(i)}(x) - \alpha_k \frac{\delta\mathcal{J}_s^{\mathrm{mod}}}{\delta\mu^{(i)}}(\mu_k)(x)\\ &= \mu_k^{(i)}(x) - \alpha_k\rho_i^{\mu_k}(x)\nabla_a Q_{\mu_k}\big(i,x,\mu_k^{(i)}(x)\big)\end{aligned}$$

with stepsize $\alpha_k > 0$. Under the current policy, we generate independent N_{traj} trajectories from the fitted model:

$$\begin{aligned}&x_{i+1}^{(j)} = f_{\theta_s}\big(x_i^{(j)}, \widehat{\mathbf{K}}x_i^{(j)} + \mu_k^{(i)}(x_i^{(j)})\big) + \widehat{w}_i^{(j)},\\ &x_0^{(j)} \sim \nu_0, \qquad \widehat{w}_i^{(j)} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,\widehat{\Sigma}_{w,s}), \qquad i = 0,\ldots,H-1, \qquad j = 1,\ldots,N_{\mathrm{traj}}.\end{aligned}$$

These trajectories provide the sample set

$$\mathcal{S}_k := \big\{(i, x_i^{(j)}) : i = 0,\ldots,H-1,\ j = 1,\ldots,N_{\mathrm{traj}}\big\}.$$

We parameterize the admissible policies by neural networks μ_ϑ with $\mu_k = \mu_{\vartheta_k}$. The stochastic update is then approximated through the least-squares problem

$$\vartheta_{k+1} \in \operatorname*{arg\,min}_{\vartheta} \sum_{(i,x)\in\mathcal{S}_k} \Big\|\mu_\vartheta^{(i)}(x) - \mu_{\vartheta_k}^{(i)}(x) + \alpha_k\nabla_a Q_{\mu_k}\big(i,x,\mu_k^{(i)}(x)\big)\Big\|_2^2,$$

and set $\mu_{k+1} = \mu_{\vartheta_{k+1}}$. In fact, using neural networks to approximate solutions of ODEs and PDEs is an established approach; see (Lagaris et al., 1998; Han et al., 2017; Bruna et al., 2024).

B.2 LITERATURE COMPARISON

The following table shows the differences between our work and the existing works.

Table 1: Comparison of dynamics, noise information, assumptions, and average regret.

Work	Model	Bounded noise/ Noise law	Given/ Learn stabilizer	Growing sieve	Assumptions	Avg. regret
Cohen et al. (2019)	$f(x,u)=\theta\phi(x,u)$; $\phi=(x^\top,u^\top)^\top$ is known; $\theta=(A,B)$ is unknown; the system runs H steps; the stage cost function is $c(x,u)=x^\top Qx+u^\top Ru$; the terminal cost function is $h(x)\equiv 0$.	N/U	Y/N	N	(A,B) is controllable; there exist known positive constants α_0, α_1, σ^2, ϑ, v, such that $\alpha_0 I\preceq Q\preceq\alpha_1 I$, $\alpha_0 I\preceq R\preceq\alpha_1 I$, $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2 I)$, $t=0,\dots,H-1$, $\|(A,B)\|_2\le\vartheta$, and $\mathcal{J}^{H,*}\le v$; the given gain K_0 is (κ_0,γ_0)-strongly stable with known κ_0,γ_0 (see Definition 3).	$\widetilde{O}(H^{-1/2})$
Simchowitz & Foster (2020)	$f(x,u)=\theta\phi(x,u)$; $\phi=(x^\top,u^\top)^\top$ is known; $\theta=(A,B)$ is unknown; the system runs H steps; the stage cost function is $c(x,u)=x^\top Qx+u^\top Ru$; the terminal cost function is $h(x)\equiv 0$.	N/U	Y/N	N	$R=I$, $Q\succeq I$, and $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,I)$ with $t=0,\dots,T-1$.	$\widetilde{O}(H^{-1/2})$
Lu & Mo (2025)	$f(x,u)=\theta\phi(x,u)$; $\phi=(x^\top,u^\top)^\top$ is known; $\theta=(A,B)$ is unknown; the system runs H steps; the stage cost function is $c(x,u)=x^\top Qx+u^\top Ru$; the terminal cost function is $h(x)\equiv 0$.	N/U	Y/N	N	(A,B) is controllable; $Q,R\succ 0$; $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,I)$ with $t=0,\dots,T-1$.	$\widetilde{O}(H^{-1/2})$
Boffi et al. (2021)	$f(x,u,t)=g(x,t)+B(x,t)[u-Y(x,t)\theta]$; g,B,Y are known; θ is unknown; the system runs H steps; the stage cost function is $c(x,u)=\|x\|_2^2$; the terminal cost function is $h(x)\equiv 0$.	Y/U	N/N	N	There exists a constant W, such that for every deterministic sequence $\{q_t\}_{t\ge0}$ satisfying $\sup_{t\ge0}\|q_t\|_2\le W$, the perturbed dynamics $x_{t+1}=g(x_t,t)+q_t$ are exponentially incrementally input-to-state stable (see Definition 1); there exists a constant M such that $\sup_{x,t}\|B(x,t)\|_2\le M$ and $\sup_{x,t}\|Y(x,t)\|_2\le M$; the unknown parameter satisfies $\|\theta\|_2\le D$ with known D; $\{w_t\}_{t\ge0}$ are mutually independent with $\mathbb{E}w_t=0$.	$\widetilde{O}(H^{-1/2})$

Continued on the next page

Table 1 continued

Work	Model	Bounded noise/ Noise law	Given/ Learn stabilizer	Growing sieve	Assumptions	Avg. regret
Kakade et al. (2020)	$f(x,u)=\theta\phi(x,u)$; ϕ is known; θ is unknown; the system runs for H episodes and each episode consists of T time steps; In each episode, the stage cost functions $c(x,u)$ with $t=0,\dots,T-1$ are not assumed to have specific forms; the terminal cost function is $h(x)\equiv 0$.	N/K	N/N	N	ϕ is uniformly bounded over $\mathrm{col}(x,u)\in\mathbb{R}^{d_x+d_u}$ there exists an oracle that can exactly solve the optimistic planning problem; for each episode, the second moment of the cumulative cost $\sum_{t=0}^{T-1}c(x_t,u_t)$ is uniformly bounded for all candidate policies and all dynamics parameters ϑ in a neighborhood of the true parameter θ; $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2 I),\ t=0,\dots,T-1$.	$\widetilde{O}(H^{-1/2})$
Wang et al. (2025)	$f=f_\theta$ belongs to the given set $\{f_\vartheta\}_{\vartheta\in\Theta}$; θ is unknown; the system runs for H episodes and each episodes consists of T time steps; In each episode, the stage cost functions $c_t(x,u)$ with $t=0,\dots,T-1$ are time-varying and are not assumed to have specific forms; the terminal cost function is $h(x)\equiv 0$.	N/K	N/N	N	f is four times differentiable with respect to u and θ; an initial identifying policy satisfying Lojasiewicz condition (see Definition 2) is available; the Fisher information matrix under the optimal policy is positive definite. Θ is compact and $\{f_\vartheta\}_{\vartheta\in\Theta}$ is uniformly bounded over $\mathrm{col}(x,u)\in\mathbb{R}^{d_x}$; $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2 I)$ with $t=0,\dots,T-1$.	$\widetilde{O}(H^{-1})$
Muehlebach et al. (2026)	$f=f_\theta$ belongs to the given set $\{f_\vartheta\}_{\vartheta\in\Theta}$; θ is unknown; the system runs H steps; the stage cost function $c(x,u)\geq 0$ is globally Lipschitz smooth and is not assumed to have a specific form; the terminal cost function is $h(x)\equiv 0$.	N/K	N/N	N	Θ is compact; a local Bellman inequality holds (see (36)); a variant of the PE condition holds (see (35)); The dynamics $f_\vartheta(x,u)$ and the corresponding feedback policies $\mu_\vartheta(x)$ are globally Lipschitz over $\mathrm{col}(x,u)\in\mathbb{R}^{d_x+d_u}$ and $x\in\mathbb{R}^{d_x}$, respectively, with constants independent of $\vartheta\in\Theta$; $w_t\overset{\text{iid}}{\sim}\mathcal{N}(0,\sigma^2 I_{d_x})$.	$\widetilde{O}(H^{-1/2})$
Vinod et al. (2022)	$x_{t+1}=f(x_t,u_t)$; f is unknown; the system runs H steps; the stage cost function $c(x,u)$ is not assumed to have a specific form.	Y/K	N/N	N	There exist known state and action sets $\mathcal{X}\subset\mathbb{R}^{d_x}$ and $\mathcal{U}\subset\mathbb{R}^{d_u}$ which are convex and compact with $f:\mathcal{X}\times\mathcal{U}\to\mathcal{X}$; $w_t\equiv 0$ for $t=0,\dots,H-1$; ∇c is Lipschitz and uniformly bounded.	$O(H^{-2/d})$ $d=d_x+d_u$

Continued on the next page

Table 1 continued

Work	Model	Bounded noise/ Noise law	Given/ Learn stabilizer	Growing sieve	Assumptions	Avg. regret
This work (Appendix D.1)	$f(x,u)=\theta\phi(x,u)$; $\phi=(x^\top,u^\top)^\top$ is known; $\theta=(A,B)$ is unknown; the stage cost function is $c(x,u)=$ $x^\top Qx+u^\top Ru$; the terminal cost function is $h(x)\equiv 0$.	N/U	N/Y	N	Assumption 5	$\widetilde{O}(H^{-1/2})$
This work (Appendix D.2)	$f(x,u)=$ $ax+g_\star(x)+bu$; a is known and $g_\star, b$ are unknown; $g_\star(x)=$ $\sum_{\ell\geq 1}\alpha_\ell\sin(\ell x)$ with $\lvert\alpha_\ell\rvert\leq A_{\mathrm{F}}e^{-\nu_{\mathrm{F}}\ell}$; see (76) for the formal definitions of $c(x,u)$ and $h(x)$.	N/U	N/Y	Y	Assumption 6	$\widetilde{O}(H^{-1/2})$
This work (Appendix D.3)	$x=(y,z)^\top$; $f(x,u)=$ $\big(y-u,\rho z+\lambda\widetilde{q}_\nu(u)\big)^\top$; $\widetilde{q}_\nu$ is known; ρ,λ are unknown; see (87) for the formal definitions of $c(x,u)$ and $h(x)$.	N/U	N/Y	N	Assumption 7	$\widetilde{O}(H^{-1/2})$
This work (Main results)	f is unknown and is not assumed to belong to a known finite-dimensional parametric family; both the stage cost function $c(x,u)$ and the terminal cost function $h(x)$ are not assumed to have specific forms.	N/U	N/Y	Y	Assumptions 1–3	$\mathrm{Err}_{\mathrm{exp}}(H)$ Theorem 3

Notation. K/U: known/unknown; Y/N: yes/no.

The column of Model shows the transition function f and the cost functions c, h studied in these works. The column of Bounded noise/Noise law shows whether each work assumes that the noise is bounded and whether each work assumes that the noise distribution is known. The column of Given/ Learn stabilizer shows whether each work assumes that there exists a known initial stabilizing controller and whether each work uses an online data-driven method to learn an initial stabilizing controller. The column of Growing sieve shows whether the dimension of $\phi^{(t)}$ in each work grows with the horizon. The column of Assumptions shows some critical assumptions in these works. The column of Avg. regret shows the corresponding regret bounds in these works.

Based on Table 1, we can clearly compare the differences and relations between our work and the prior works. First, many of the works in the table require estimating the unknown dynamics f. However, according to the Model column, we can find that nearly all of the prior works in the table assumed that f belongs to some special parameterized function sets (see Cohen et al. (2019); Simchowitz & Foster (2020); Lu & Mo (2025); Boffi et al. (2021); Kakade et al. (2020); Wang et al. (2025); Muehlebach et al. (2026)). Under such assumptions, these works transformed the estimation

of f into a finite dimensional parameter estimation problem. In contrast, to study a more general situation, we do not assume that f belongs to any parameterized function set. Instead, we use a nonparametric estimation method to estimate f and obtain the error bound of f in Proposition 1. Second, many existing works in the table (such as Cohen et al. (2019); Simchowitz & Foster (2020)) assume that there exists a known initial stabilizing controller, and some other works assume that the system is self-stable (such as Boffi et al. (2021)) or that the state-action trajectories remain confined to compact sets (such as Muehlebach et al. (2026); Kakade et al. (2020); Vinod et al. (2022)). These works use such assumptions to exclude the issue of stability of the systems and purely focus on the analysis of the regret bound. However, some works have shown that there are many nonlinear systems and high-dimensional systems for which a stabilizing controller is hard to be designed without any model information (see (Zhang & Guo, 2002; Tsiamis et al., 2022)), making the above assumptions unavailable. To bridge this gap, we do not assume the aforementioned assumptions but instead design an online learning step in our algorithm to learn a stabilizing controller. Specifically, we prove that our algorithm can find a stabilizing gain $\widehat{\mathbf{K}}$ in a finite number of steps with high probability (see Lemma 1). Then we show that under such a stabilizing gain, the bound $\|x_t\| + \|u_t\| = O(\log t)$ holds with high probability (see Lemma 3), thereby guaranteeing that the final regret bound in Theorem 3 holds. This bound is used to guarantee (8) in Theorem 1 and (10) in Theorem 2 hold. Furthermore, compared with our work, Theorem 2 of Vinod et al. (2022) and Theorem 10 of Brown et al. (2024) establish regret guarantees under prespecified compact state and action spaces. However, as we have proved in this paper, the state may not remain in any fixed compact set almost surely when there exist Gaussian noises in the systems. This restricts their applicability to systems with unbounded process noise. Third, many works require the true noise covariance matrix for designing a controller (such as Kakade et al. (2020); Muehlebach et al. (2026); Wang et al. (2025)). However, in many practical problems, we do not know the noise covariance matrix Odelson et al. (2006); Zhang et al. (2020); Kong (2023). To bridge this gap, we use the estimation of the noise covariance Σ_w as the alternative to design the controller and analyze the estimation error bound of Σ_w in Theorem 2. Fourth, our Assumption 2 is a variant of the classical PE condition (see Definition 4.3.1 of Ioannou & Sun 2012). There are many works in the table requiring similar variants of the PE condition; see Assumption 6 of Muehlebach et al. (2026), Theorem 20 of Cohen et al. (2019), Lemma E.4 of Simchowitz & Foster (2020), and Lemma 18 of Lu & Mo (2025). Furthermore, our assumption is a relatively weaker condition in the literature. For example, Assumption 6 of Muehlebach et al. (2026) is also a variant of the PE condition and can be written as

$$\frac{1}{M}\sum_{i=t}^{t+M-1}\mathbb{E}\left[\frac{|f_\theta(x_i,u_i)-f(x_i,u_i)|^2}{1+(\|x_i\|_2^2+\|u_i\|_2^2)/b^2}\mid \mathscr{H}_{t-1}\right]\geq c_{\mathrm{e}}\sigma_{\mathrm{id}}^2\|\theta-\theta_t^*\|_{\mathrm{F}}^2,\quad \forall\theta\in\Theta \tag{35}$$

for any update time t with $b, c_{\mathrm{e}} > 0$. In contrast, under BMSB condition (6) of Assumption 2, for every $0 \leq j \leq n_t - k_t$, we have

$$\frac{1}{k_t}\sum_{i=j}^{j+k_t-1}\mathbb{E}\big[\|f_\theta(x_i,u_i)-f_{\theta_t^*}(x_i,u_i)\|_2^2\mid\mathscr{H}_{j-1}\big]\geq\frac{\lambda_{\min}(\Gamma_{\mathrm{PE},t})}{4\kappa_{\mathrm{PE},t}}\|\theta-\theta_t^*\|_{\mathrm{F}}^2,\quad\forall\theta\in\Theta_t.$$

Here, $f_{\theta_t^*}$ belongs to $\{f_\vartheta\}_{\vartheta\in\Theta_t}$ satisfying (5). It is easy to note that our formulation is similar to (35) but only requires the lower bound of the distance between f_θ and $f_{\theta_t^*}$ not the one between f_θ and f. Moreover, we allow Θ_t, block length k_t, and excitation matrix $\Gamma_{\mathrm{PE},t}$ to vary with t, whereas Assumption 6 of Muehlebach et al. (2026) does not. Fifth, to derive the regret bound in Theorem 3, we assume that Bellman-type inequality (13) holds. In fact, a similar inequality has been proved in the linear systems; see Lemma 7 of Cohen et al. (2019), Theorem 2 of Mania et al. (2019), Theorem 3 of Simchowitz & Foster (2020), and Lemma 22 of Lu & Mo (2025). For nonlinear systems, such an inequality holds when the system is dispassitive (see Proposition D.1 of Muehlebach et al. (2026), Theorem 8 of Boffi et al. (2021), and Chapter 5 of Khalil & Grizzle (2002)). In Appendix D, we also show that inequality (13) holds for three different systems. Furthermore, our condition can be seen as a general version of the Bellman condition of Muehlebach et al. (2026). Concretely, let V and γ denote a cost-to-go function and average cost in Muehlebach et al. (2026). Define

$$\mathcal{D}_{t;V,\gamma}^H(x,u):=\frac{c(x,u)+\mathbb{E}_{w_t}[V(f(x,u)+w_t)]-V(x)-\gamma}{H},\ \forall(x,u)\in\mathbb{R}^{d_x+d_u},\ 0\leq t<H.$$

For a fixed sufficiently small radius $a_s > 0$, a corresponding function V, and $\theta_s \in \Theta$ satisfying $\|\theta_s - \theta_s^*\|_{\mathrm{F}} < a_s$, Assumption 7 of Muehlebach et al. (2026) can be written as

$$\mathbb{E}_{e_t^{\mathrm{id}}}\big[\mathcal{D}^H_{t;V,\gamma}(x, \bar{\mu}_{\theta_s}(x) + e_t^{\mathrm{id}})\big] \le \frac{L_{\mathrm{B}} a_s^2}{H} + \frac{L_{\mathrm{e}}}{H}\mathbb{E}\|e_t^{\mathrm{id}}\|_2^2, \qquad x \in \mathbb{R}^{d_x}. \tag{36}$$

Here, μ_{θ_s} denotes the controller associated with θ_s and $L_{\mathrm{B}}, L_{\mathrm{e}} > 0$ are the constants in that assumption. In contrast, by introducing the event $\mathcal{G}_s$ and variable c_s, our Bellman inequality (13) additionally accounts for the influence of the probability that the state-action trajectory is unbounded and the error from covariance estimation.

B.3 Online identification and certification of the initial gain

We now study the probability guarantee of the stability of system (2) under Algorithm 1. Define

$$\mathcal{E}_{\mathrm{cert}} := \left\{ \tau \le N_u, \quad \sup_{x \neq 0} \frac{\|f(x, K_\tau x)\|_{P_\tau}}{\|x\|_{P_\tau}} \le q_1 \right\}.$$

Lemma 1. *Under Assumption 1 and* (28)*,* $\mathbb{P}(\mathcal{E}_{\mathrm{cert}}) \ge 1 - \delta_{\mathrm{conf}}$*. On* $\mathcal{E}_{\mathrm{cert}}$*, set* $\widehat{\mathbf{K}} = K_\tau$*,* $\widehat{P} = P_\tau$*, and define* $V_{\mathrm{sta}}(x) := x^\top \widehat{P} x$*. For every* $x, w \in \mathbb{R}^{d_x}$ *and* $v \in \mathbb{R}^{d_u}$*,*

$$\underline{v}\|x\|_2^2 \le V_{\mathrm{sta}}(x) \le \overline{v}\|x\|_2^2, \tag{37}$$

$$V_{\mathrm{sta}}\big(f(x, \widehat{\mathbf{K}}x + v) + w\big) \le \rho_{\mathrm{sta}} V_{\mathrm{sta}}(x) + b_u\|v\|_2^2 + b_w\|w\|_2^2, \tag{38}$$

where

$$\rho_{\mathrm{sta}} := \frac{1 + q_1^2}{2}, \qquad b_w := \frac{2\overline{v}(1 + q_1^2)}{1 - q_1^2}, \qquad b_u := \mathscr{L}^2 b_w.$$

Proof. On $\mathcal{E}_{\mathrm{conf}}$, for every $1 \le n \le \widetilde{N}$, $(K, P) \in \mathfrak{S}$, and $x \in \mathbb{R}^{d_x}$, it holds that

$$\|g_n(x, Kx) - f(x, Kx)\|_P \le r_n \sqrt{\frac{\overline{v}(1 + \|K\|_{\mathrm{op}}^2)}{\underline{v}}}\|x\|_P \le \widetilde{C} r_n \|x\|_P.$$

Consequently, every accepted pair (K_n, P_n) in step 7 of Algorithm 1 satisfies

$$\sup_{x \neq 0} \frac{\|f(x, K_n x)\|_{P_n}}{\|x\|_{P_n}} \le \Phi_n(K_n, P_n) \le U_n \le q_1.$$

For the pair $(K_\star, P_\star)$ in Assumption 1(b), it holds that $\Phi_n(K_\star, P_\star) \le q_1 - q_2 + 2\widetilde{C} r_n$. If no earlier test succeeds, the actual data through $\widetilde{N}$ are the exploration data. Therefore,

$$U_{\widetilde{N}} \le \Phi_{\widetilde{N}}(K_\star, P_\star) + \widetilde{\epsilon}_{\widetilde{N}} \le q_1 - q_2 + 2\widetilde{C} r_{\widetilde{N}} + \widetilde{\epsilon}_{\widetilde{N}} \le q_1.$$

Hence, we have $\mathcal{E}_{\mathrm{conf}} \subseteq \mathcal{E}_{\mathrm{cert}}$, $\mathbb{P}(\mathcal{E}_{\mathrm{cert}}) \ge \mathbb{P}(\mathcal{E}_{\mathrm{conf}}) \ge 1 - \delta_{\mathrm{conf}}$. On $\mathcal{E}_{\mathrm{cert}}$, (37) follows from $(K_\tau, P_\tau) \in \mathfrak{S}$. The input Lipschitz condition gives

$$\|f(x, \widehat{\mathbf{K}}x + v) + w\|_{\widehat{P}} \le q_1\|x\|_{\widehat{P}} + \sqrt{\overline{v}}\big(\mathscr{L}\|v\|_2 + \|w\|_2\big).$$

Since $\rho_{\mathrm{sta}} > q_1^2$, Young's inequality yields

$$\begin{aligned} V_{\mathrm{sta}}\big(f(x, \widehat{\mathbf{K}}x + v) + w\big) &\le \rho_{\mathrm{sta}}\|x\|_{\widehat{P}}^2 + \frac{\overline{v}\rho_{\mathrm{sta}}}{\rho_{\mathrm{sta}} - q_1^2}\big(\mathscr{L}\|v\|_2 + \|w\|_2\big)^2 \\ &\le \rho_{\mathrm{sta}} V_{\mathrm{sta}}(x) + \frac{2\overline{v}(1 + q_1^2)}{1 - q_1^2}\big(\mathscr{L}^2\|v\|_2^2 + \|w\|_2^2\big), \end{aligned}$$

which proves (38). □

B.4 SMALL-BALL TOOLS FOR THE STATISTICAL PROOFS

Definition 5. Let $\{\varphi_i\}_{i=0}^{n-1}$ be a sequence of d-dimensional random vectors adapted to a filtration $(\mathscr{H}_i)_{i\geq -1}$. For $k \in [n]$, $p \in (0,1]$, and $\Gamma_{\mathrm{sb}} \succ 0$, the sequence $\{\varphi_i\}_{i=0}^{n-1}$ satisfies the $(k, \Gamma_{\mathrm{sb}}, p)$-BMSB condition if, for every fixed $v \in \mathbb{S}^{d-1}$ and every $0 \leq j \leq n-k$, the inequality

$$\frac{1}{k}\sum_{r=1}^{k} \mathbb{P}\Big(|v^\top \varphi_{j+r-1}| \geq \sqrt{v^\top \Gamma_{\mathrm{sb}} v} \,\Big|\, \mathscr{H}_{j-1}\Big) \geq p \quad \text{a.s.}$$

holds.

Lemma 2. *Fix $n, d \in \mathbb{N}_+$, $k \in [n]$, $p \in (0,1]$, $\delta_{\mathrm{sb}} \in (0,1)$, and $\delta_{\mathrm{up}} \in [0,1)$. Let $\Gamma_{\mathrm{sb}}, \Gamma \in \mathbb{S}^d_{++}$ be deterministic, define $G_n := \sum_{i=0}^{n-1} \varphi_i \varphi_i^\top$, and suppose $\{\varphi_i\}_{i=0}^{n-1}$ satisfies the $(k, \Gamma_{\mathrm{sb}}, p)$-BMSB condition, $\Gamma_{\mathrm{sb}} \preceq \Gamma$, and $\mathbb{P}(G_n \not\preceq n\Gamma) \leq \delta_{\mathrm{up}}$. If*

$$n \geq \frac{10k}{p^2}\left\{\log\frac{1}{\delta_{\mathrm{sb}}} + 2d\log\frac{10}{p} + \log\det(\Gamma_{\mathrm{sb}}^{-1/2}\Gamma\Gamma_{\mathrm{sb}}^{-1/2})\right\}, \tag{39}$$

then

$$\mathbb{P}\left(\left\{G_n \succeq \frac{p^2}{16}k\left\lfloor\frac{n}{k}\right\rfloor \Gamma_{\mathrm{sb}}\right\} \cap \{G_n \preceq n\Gamma\}\right) \geq 1 - \delta_{\mathrm{up}} - \delta_{\mathrm{sb}}. \tag{40}$$

Proof. For a fixed $v \in \mathbb{S}^{d-1}$, by using Proposition 2.5 of Simchowitz et al. (2018), we get

$$\mathbb{P}\left(\sum_{i=0}^{n-1}(v^\top\varphi_i)^2 \leq \frac{p^2}{8}k\left\lfloor\frac{n}{k}\right\rfloor v^\top\Gamma_{\mathrm{sb}}v\right) \leq \exp\left\{-\frac{p^2}{8}\left\lfloor\frac{n}{k}\right\rfloor\right\}. \tag{41}$$

Set

$$\Gamma_{\min} := \frac{p^2}{8}k\left\lfloor\frac{n}{k}\right\rfloor\Gamma_{\mathrm{sb}}, \qquad \Gamma_{\max} := n\Gamma.$$

Condition (39) implies $n \geq 10k/p^2 \geq 10k$ since $p \leq 1$ and $2d\log(10/p) > 1$. Hence $\Gamma_{\min} \preceq (p^2n/8)\Gamma_{\mathrm{sb}} \preceq n\Gamma = \Gamma_{\max}$, verifying the condition required in Lemma 4.1 of Simchowitz et al. (2018). $\mathcal{T}$ is a $1/4$-net of $\mathbb{S}_{\Gamma_{\min}} := \Gamma_{\min}^{-1/2}\mathbb{S}^{d-1} = \{v : \|\Gamma_{\min}^{1/2}v\|_2 = 1\}$, measured in the $\|\Gamma_{\max}^{1/2}(\cdot)\|_2$ metric. If for any $v \in \mathcal{T}$, it holds that

$$\sum_{i=0}^{n-1}(v^\top\varphi_i)^2 \leq \frac{p^2}{8}k\left\lfloor\frac{n}{k}\right\rfloor v^\top\Gamma_{\mathrm{sb}}v,$$

we can show that $G_n \succeq \Gamma_{\min}/2$ holds on the set $\{G_n \preceq \Gamma_{\max}\}$ via Lemma 4.1 of Simchowitz et al. (2018). Meanwhile, note that $n \geq 10k/p^2$ gives

$$k\left\lfloor\frac{n}{k}\right\rfloor \geq n - k \geq 0.9n.$$

By using Lemma D.1 of Simchowitz et al. (2018) and $\Gamma_{\mathrm{sb}} \preceq \Gamma$, it holds that

$$\begin{aligned}\log|\mathcal{T}| \leq& d\log 9 + d\log\left(\frac{8n}{p^2k\lfloor n/k\rfloor}\right) + \log\det(\Gamma_{\mathrm{sb}}^{-1/2}\Gamma\Gamma_{\mathrm{sb}}^{-1/2})\\ \leq& 2d\log(10/p) + \log\det(\Gamma_{\mathrm{sb}}^{-1/2}\Gamma\Gamma_{\mathrm{sb}}^{-1/2}).\end{aligned}$$

Applying (39), we have

$$\frac{p^2}{8}\left\lfloor\frac{n}{k}\right\rfloor \geq \frac{np^2}{10k} \geq \log(1/\delta_{\mathrm{sb}}) + \log|\mathcal{T}|, \tag{42}$$

where the last inequality holds by (39). Finally, by combining (41), (42), $\mathbb{P}(G_n \not\preceq n\Gamma) \leq \delta_{\mathrm{up}}$, and the fact that $G_n \succeq \Gamma_{\min}/2$ if $G_n \succeq \Gamma_{\max}$ together, we get (40). □

C MATHEMATICAL PROOF

C.1 PROOFS OF THEOREM 1

Lemma 3 (Uniform state–action envelope). *Let B_{ctl} and $(\sigma_{\mathrm{id},i})_{i\ge 0}$ be as in Algorithm 1, and set*

$$\overline{\sigma}_{\mathrm{id},t} := \max_{0\le i\le t} \sigma_{\mathrm{id},i}, \qquad \sigma_\infty^2 := \max_{\ell\in[d_x]} [\Sigma_w]_{\ell\ell}.$$

For $a \in (0,1)$ and $t \ge 0$, define

$$\begin{aligned}
&q_0(a) := \inf\{r \ge 0 : \mathbb{P}_\tau(V_{\mathrm{sta}}(x_0) > r) \le a/2\},\\
&\ell_t(a) := \log\left(\frac{2\pi^2(d_x+d_u)(t+1)^2}{3a}\right),\\
&C_{v,t}(a) := 2B_{\mathrm{ctl}}^2 + 4d_u\overline{\sigma}_{\mathrm{id},t}^2\ell_t(a),\\
&R_{x,t}^2(a) := \frac{1}{\underline{v}}\left[q_0(a) + \frac{b_u C_{v,t}(a) + 2b_w d_x\sigma_\infty^2\ell_t(a)}{1-\rho_{\mathrm{sta}}}\right],\\
&R_{u,t}(a) := \|\widehat{\mathbf{K}}\|_\infty R_{x,t}(a) + B_{\mathrm{ctl}} + \overline{\sigma}_{\mathrm{id},t}\sqrt{2\ell_t(a)}.
\end{aligned}$$

Then, on the event $\mathcal{E}_{\mathrm{cert}}$, by using Lemma 1, with probability at least $1-a$, the following inequalities hold simultaneously at every time $0 \le t \le H-1$:

$$\max_{0\le i\le t}\|x_i\|_\infty \le R_{x,t}(a), \qquad \max_{0\le i\le t}\|u_i\|_\infty \le R_{u,t}(a). \tag{43}$$

The event in (43) *for a fixed t is denoted by $\mathcal{E}_{\mathrm{app},t}(a)$.*

Proof. Since $V_{\mathrm{sta}}(x_0) < \infty$ almost surely, it holds $q_0(a) < \infty$. For $i \ge 0$, let

$$\Omega_a := \bigcap_{i=0}^{\infty}\left[\bigcap_{\ell=1}^{d_x}\left\{|[w_i]_\ell| \le \sigma_\infty\sqrt{2\ell_i(a)}\right\} \cap \bigcap_{r=1}^{d_u}\left\{|[\zeta_i^{\mathrm{id}}]_r| \le \sqrt{2\ell_i(a)}\right\}\right].$$

The one-dimensional Gaussian tail bound implies

$$\mathbb{P}_\tau(\Omega_a^c) \le \sum_{i=0}^{\infty} 2(d_x+d_u)\exp\{-\ell_i(a)\} = \frac{3a}{\pi^2}\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{a}{2}.$$

On $\Omega_a \cap \{V_{\mathrm{sta}}(x_0) \le q_0(a)\}$, whose probability is at least $1-a$, for every $i \ge 0$, we have

$$\|w_i\|_2^2 \le 2d_x\sigma_\infty^2\ell_i(a), \qquad \|u_{\mathrm{opt},i} + e_i^{\mathrm{id}}\|_2^2 \le C_{v,i}(a).$$

Applying (38) with $v = u_{\mathrm{opt},i} + e_i^{\mathrm{id}}$ and iterating, for $0 \le j \le t$ we obtain

$$\begin{aligned}
V_{\mathrm{sta}}(x_j) &\le \rho_{\mathrm{sta}}^j V_{\mathrm{sta}}(x_0) + \sum_{i=0}^{j-1}\rho_{\mathrm{sta}}^{j-1-i}\left\{b_u C_{v,i}(a) + 2b_w d_x\sigma_\infty^2\ell_i(a)\right\}\\
&\le q_0(a) + \frac{b_u C_{v,t}(a) + 2b_w d_x\sigma_\infty^2\ell_t(a)}{1-\rho_{\mathrm{sta}}}.
\end{aligned}$$

The lower bound in (37) yields the state bound in (43). Finally, $u_i = \widehat{\mathbf{K}}x_i + u_{\mathrm{opt},i} + e_i^{\mathrm{id}}$ and the coordinate bounds on e_i^{id} yield the corresponding action bound. Since the event $\Omega_a \cap \{V_{\mathrm{sta}}(x_0) \le q_0(a)\}$ is independent of t and supports these estimates for every t, both bounds hold simultaneously over the entire horizon. □

Proof of Theorem 1. Step 1: establish two-sided empirical curvature. Fix an update time t, suppress the superscript t on the feature matrices, and write $n := n_t$, $d := d_t$, and $D := D_t = d_x d$. Define $G_\phi := \sum_{i=0}^{n-1}\phi_i\phi_i^\top$, and $G := \sum_{i=0}^{n-1}\mathbf{\Phi}_i^\top\mathbf{\Phi}_i$. By $\mathbf{\Phi}_i = \phi_i^\top \otimes I_{d_x}$, we have

$$\mathbf{\Phi}_i^\top\mathbf{\Phi}_i = (\phi_i \otimes I_{d_x})(\phi_i^\top \otimes I_{d_x}) = (\phi_i\phi_i^\top)\otimes I_{d_x};$$

thus, $G = G_\phi \otimes I_{d_x}$. We first derive the two-sided empirical-curvature event. Denoting $Y := \Gamma_{\mathrm{up},t}^{-1/2} G_\phi \Gamma_{\mathrm{up},t}^{-1/2} \succeq 0$, Assumption 2 implies $\mathbb{E}[Y \mid \mathscr{H}_{-1}] \preceq nI_d$ and, almost surely, $\mathbb{E}[\mathrm{tr}(Y) \mid \mathscr{H}_{-1}] = \mathrm{tr}(\mathbb{E}[Y \mid \mathscr{H}_{-1}]) \le nd$. Because $\overline{\Gamma}_t := \frac{1}{2}\Gamma_{\mathrm{PE},t} + \frac{d_t}{\alpha_{\mathrm{up},t}}\Gamma_{\mathrm{up},t} \succeq (d/\alpha_{\mathrm{up},t})\Gamma_{\mathrm{up},t}$, one has

$$\{G_\phi \not\preceq n\overline{\Gamma}_t\} \subseteq \left\{G_\phi \not\preceq \frac{nd}{\alpha_{\mathrm{up},t}}\Gamma_{\mathrm{up},t}\right\} = \left\{\lambda_{\max}(Y) > \frac{nd}{\alpha_{\mathrm{up},t}}\right\} \subseteq \left\{\mathrm{tr}(Y) > \frac{nd}{\alpha_{\mathrm{up},t}}\right\}.$$

Applying conditional Markov inequality to $\mathrm{tr}(Y)$ yields, almost surely,

$$\mathbb{P}\big(G_\phi \not\preceq n\overline{\Gamma}_t \,\big|\, \mathscr{H}_{-1}\big) \le \mathbb{P}\left(\mathrm{tr}(Y) > \frac{nd}{\alpha_{\mathrm{up},t}} \,\middle|\, \mathscr{H}_{-1}\right) \le \frac{\alpha_{\mathrm{up},t}}{nd}\mathbb{E}[\mathrm{tr}(Y) \mid \mathscr{H}_{-1}] \le \alpha_{\mathrm{up},t},$$

which gives $\mathbb{P}_\tau(G_\phi \not\preceq n\overline{\Gamma}_t) \le \alpha_{\mathrm{up},t}$. Fix $v \in \mathbb{S}^{d-1}$ and $0 \le j \le n - k_t$. Conditionally on $\mathscr{H}_{j-1}$, let R be uniform on $\{1, \dots, k_t\}$, independently of the trajectory, and set $Z := |v^\top \phi_{j+R-1}|^2$. Then $\mathbb{E}[Z \mid \mathscr{H}_{j-1}] = m_{2,j,t}(v)$ and $\mathbb{E}[Z^2 \mid \mathscr{H}_{j-1}] = m_{4,j,t}(v)$. Conditional Paley–Zygmund inequality, followed by Assumption 2, gives

$$\begin{aligned}
&\frac{1}{k_t}\sum_{r=1}^{k_t} \mathbb{P}\left(|v^\top \phi_{j+r-1}| \ge \sqrt{m_{2,j,t}(v)/2} \,\middle|\, \mathscr{H}_{j-1}\right) \\
&= \mathbb{P}\left(Z \ge \frac{1}{2}\mathbb{E}[Z \mid \mathscr{H}_{j-1}] \,\middle|\, \mathscr{H}_{j-1}\right) \ge \frac{m_{2,j,t}(v)^2}{4m_{4,j,t}(v)} \ge \frac{1}{4\kappa_{\mathrm{PE},t}} =: p_t^{\mathrm{sb}}.
\end{aligned}$$

Moreover, by $\sqrt{m_{2,j,t}(v)/2} \ge \sqrt{v^\top \Gamma_{\mathrm{PE},t} v/2}$, one has

$$\frac{1}{k_t}\sum_{r=1}^{k_t} \mathbb{P}\left(|v^\top \phi_{j+r-1}| \ge \sqrt{v^\top \Gamma_{\mathrm{PE},t} v/2} \,\middle|\, \mathscr{H}_{j-1}\right) \ge p_t^{\mathrm{sb}}.$$

This is precisely the $(k_t, \frac{1}{2}\Gamma_{\mathrm{PE},t}, p_t^{\mathrm{sb}})$-BMSB condition in Definition 5. The remaining hypotheses of Lemma 2 hold because $0 \prec \frac{1}{2}\Gamma_{\mathrm{PE},t} \preceq \overline{\Gamma}_t$ and $\mathbb{P}_\tau(G_\phi \not\preceq n\overline{\Gamma}_t) \le \alpha_{\mathrm{up},t}$. Under the substitution

$$(k, p, d, \Gamma_{\mathrm{sb}}, \Gamma, \delta_{\mathrm{up}}, \delta_{\mathrm{sb}}) = (k_t, p_t^{\mathrm{sb}}, d_t, \tfrac{1}{2}\Gamma_{\mathrm{PE},t}, \overline{\Gamma}_t, \alpha_{\mathrm{up},t}, \delta_{\mathrm{sb},t}),$$

the sample-size hypothesis of Lemma 2 is exactly (7). Hence the event

$$\mathcal{E}_{G,t} := \left\{G_\phi \succeq \frac{(p_t^{\mathrm{sb}})^2}{32} k_t \left\lfloor \frac{n}{k_t} \right\rfloor \Gamma_{\mathrm{PE},t}\right\} \cap \{G_\phi \preceq n\overline{\Gamma}_t\}$$

satisfies $\mathbb{P}_\tau(\mathcal{E}_{G,t}^c) \le \alpha_{\mathrm{up},t} + \delta_{\mathrm{sb},t}$. On $\mathcal{E}_{G,t}$, it is obvious that

$$\begin{aligned}
G = G_\phi \otimes I_{d_x} &\succeq \frac{(p_t^{\mathrm{sb}})^2}{32} k_t \left\lfloor \frac{n}{k_t} \right\rfloor (\Gamma_{\mathrm{PE},t} \otimes I_{d_x}) \\
&\succeq \frac{(p_t^{\mathrm{sb}})^2}{32} k_t \left\lfloor \frac{n}{k_t} \right\rfloor \lambda_{\min}(\Gamma_{\mathrm{PE},t}) I_D =: g_{-,t} I_D.
\end{aligned}$$

By the same token, $G \preceq n(\overline{\Gamma}_t \otimes I_{d_x})$.

Step 2: control the Gaussian noise score. Denote $a_i := f(x_i, u_i) - f_{\theta_t^*}(x_i, u_i)$, $S_a := \sum_{i=0}^{n-1} \mathbf{\Phi}_i^\top a_i$, $S_w := \sum_{i=0}^{n-1} \mathbf{\Phi}_i^\top w_i$, and $S := S_a + S_w$. For every $v \in \mathbb{R}^{d_x}$, the independence of w_i from $\mathscr{H}_i$ yields

$$\mathbb{E}[\exp(v^\top w_i) \mid \mathscr{H}_i] = \exp\left(\tfrac{1}{2} v^\top \Sigma_w v\right) \le \exp\left(\tfrac{\|\Sigma_w\|_{\mathrm{op}}}{2}\|v\|_2^2\right).$$

For fixed $q \in \mathbb{R}^D$, writing

$$L_r(q) := \exp\left\{q^\top \sum_{i=0}^{r-1} \mathbf{\Phi}_i^\top w_i - \frac{\|\Sigma_w\|_{\mathrm{op}}}{2} q^\top \sum_{i=0}^{r-1} \mathbf{\Phi}_i^\top \mathbf{\Phi}_i q\right\}, \qquad 0 \le r \le n,$$

we have $\mathbb{E}[L_{r+1}(q) \mid \mathscr{H}_r] \le L_r(q)$; hence $(L_r(q))_{r=0}^n$ is a nonnegative supermartingale with $L_0(q) = 1$, and therefore $\mathbb{E}L_n(q) \le 1$. Let $V := G + \lambda_t I_D$. If $\|\Sigma_w\|_{\mathrm{op}} > 0$, Let $\nu_t :=$

$\mathcal{N}\big(0, (\sigma_w^2\lambda_t)^{-1} I_D\big)$ and define the Gaussian mixture $\overline{L}_n := \int_{\mathbb{R}^D} L_n(q)\, \nu_t(\mathrm{d}q)$. TBecause the integrand is nonnegative, Tonelli's theorem and $\mathbb{E}_\tau L_n(q) \le 1$ give $\mathbb{E}_\tau \overline{L}_n = \int_{\mathbb{R}^D} \mathbb{E}_\tau L_n(q)\, \nu_t(\mathrm{d}q) \le 1$. Moreover, it follows that

$$\overline{L}_n = \left\{ \frac{\det(\lambda_t I_D)}{\det V} \right\}^{1/2} \exp\left\{ \frac{\|S_w\|_{V^{-1}}^2}{2\|\Sigma_w\|_{\mathrm{op}}} \right\}.$$

Markov's inequality therefore shows that, with probability at least $1 - \delta_{\mathrm{mg},t}$,

$$\|S_w\|_{V^{-1}}^2 \le \|\Sigma_w\|_{\mathrm{op}} \left\{ \log \frac{\det V}{\det(\lambda_t I_D)} + 2 \log \frac{1}{\delta_{\mathrm{mg},t}} \right\}. \tag{44}$$

If $\|\Sigma_w\|_{\mathrm{op}} = 0$, then $\Sigma_w = 0$ and $w_i = 0$ almost surely, so (44) holds surely. Let $\mathcal{E}_{\mathrm{mg},t}$ denote this event. In either case, $\mathbb{P}_\tau(\mathcal{E}_{\mathrm{mg},t}^c) \le \delta_{\mathrm{mg},t}$.

On $\mathcal{E}_{G,t}$, $V = (G_\phi + \lambda_t I_d) \otimes I_{d_x}$. Using $\det(A \otimes I_{d_x}) = \det(A)^{d_x}$ and $D = d_x d$, we obtain

$$\log \frac{\det V}{\det(\lambda_t I_D)} = d_x \log\det\left(I_d + \frac{G_\phi}{\lambda_t} \right) \le d_x \log\det\left(I_d + \frac{n\overline{\Gamma}_t}{\lambda_t} \right).$$

The last inequality follows from $G_\phi \preceq n\overline{\Gamma}_t$ and monotonicity of $A \mapsto \log\det(A)$ on the positive-definite cone. The lower-Gram bound $G \succeq g_{-,t} I_D$ also gives $\lambda_t I_D \preceq \frac{\lambda_t}{g_{-,t}} G$ and $V \preceq \left(1 + \frac{\lambda_t}{g_{-,t}}\right) G$. Both G and V are positive definite on $\mathcal{E}_{G,t}$; hence $G^{-1} \preceq (1 + \lambda_t / g_{-,t}) V^{-1}$. It follows that, on $\mathcal{E}_{G,t} \cap \mathcal{E}_{\mathrm{mg},t}$,

$$\begin{aligned} \|S_w\|_{G^{-1}}^2 &\le \left(1 + \frac{\lambda_t}{g_{-,t}}\right) \|S_w\|_{V^{-1}}^2 \\ &\le \|\Sigma_w\|_{\mathrm{op}} \left(1 + \frac{\lambda_t}{g_{-,t}}\right) \left\{ d_x \log\det\left(I_d + \frac{n\overline{\Gamma}_t}{\lambda_t} \right) + 2 \log \frac{1}{\delta_{\mathrm{mg},t}} \right\}. \end{aligned}$$

Step 3: control the approximation score and assemble the good event. We now control the approximation score. Let $\mathsf{X} \in \mathbb{R}^{nd_x \times D}$ be the block-row matrix whose ith block row is $\mathbf{\Phi}_i$, and let $\mathsf{a} := (a_0^\top, \dots, a_{n-1}^\top)^\top \in \mathbb{R}^{nd_x}$. Then $S_a = \mathsf{X}^\top \mathsf{a}$ and $G = \mathsf{X}^\top \mathsf{X}$. On $\mathcal{E}_{G,t}$, $G \succeq g_{-,t} I_D \succ 0$, and hence X has full column rank. Therefore, $\|S_a\|_{G^{-1}}^2 = \mathsf{a}^\top \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top \mathsf{a}$. The matrix $P_{\mathsf{X}} := \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top$ is symmetric and idempotent, and hence is the orthogonal projector onto $\mathrm{col}(\mathsf{X})$. Thus $0 \preceq P_{\mathsf{X}} \preceq I_{nd_x}$, and

$$\|S_a\|_{G^{-1}}^2 = \mathsf{a}^\top P_{\mathsf{X}} \mathsf{a} \le \|\mathsf{a}\|_2^2 = \sum_{i=0}^{n-1} \|a_i\|_2^2.$$

Fix $\alpha_{\mathrm{app},t} \in (0,1)$, denote $T_t := (t + 2M - 1) \wedge (H - 1)$, and define

$$\begin{aligned} &\mathcal{K}_t(\alpha_{\mathrm{app},t}) := \{(x,u) : \|x\|_\infty \le R_{x,T_t}(\alpha_{\mathrm{app},t}),\ \|u\|_\infty \le R_{u,T_t}(\alpha_{\mathrm{app},t})\}, \\ &\Delta_t := \sup_{(x,u) \in \mathcal{K}_t(\alpha_{\mathrm{app},t})} \|f(x,u) - f_{\theta_t^*}(x,u)\|_2. \end{aligned}$$

Hence, Lemma 3 gives an event $\mathcal{E}_{\mathrm{app},t} := \mathcal{E}_{\mathrm{app},T_t}(\alpha_{\mathrm{app},t})$ of probability at least $1 - \alpha_{\mathrm{app},t}$ on which $(x_i, u_i) \in \mathcal{K}_t(\alpha_{\mathrm{app},t})$ for every $i \le T_t$. Since $n - 1 = t - M - 2 \le T_t$, this event implies

$$\frac{1}{n} \sum_{i=0}^{n-1} \|f(x_i,u_i) - f_{\theta_t^*}(x_i,u_i)\|_2^2 \le \Delta_t^2.$$

The event on the left below is contained in $\mathcal{E}_{\mathrm{app},T_t}(\alpha_{\mathrm{app},t})^c$; hence

$$\mathbb{P}_\tau\left(\frac{1}{n} \sum_{i=0}^{n-1} \|f(x_i,u_i) - f_{\theta_t^*}(x_i,u_i)\|_2^2 > \Delta_t^2 \right) \le \mathbb{P}_\tau\left(\mathcal{E}_{\mathrm{app},T_t}(\alpha_{\mathrm{app},t})^c \right) \le \alpha_{\mathrm{app},t}. \tag{45}$$

Hence, it holds that $\mathbb{P}_\tau(\mathcal{E}_{\mathrm{app},t}^c) \le \alpha_{\mathrm{app},t}$; on $\mathcal{E}_{G,t} \cap \mathcal{E}_{\mathrm{app},t}$, $\|S_a\|_{G^{-1}} \le \sqrt{n}\, \Delta_t$. In addition, it holds that

$$\|S\|_{G^{-1}} \le \|S_a\|_{G^{-1}} + \|S_w\|_{G^{-1}}$$

$$\leq \sqrt{n_t}\Delta_t + \|\Sigma_w\|_{\mathrm{op}}^{1/2}\sqrt{1+\frac{\lambda_t}{g_{-,t}}}\left\{d_x \log\det\left(I_{d_t}+\frac{n_t\overline{\Gamma}_t}{\lambda_t}\right)+2\log\frac{1}{\delta_{\mathrm{mg},t}}\right\}^{1/2} =: z_t$$

on the event $\mathcal{E}_t := \mathcal{E}_{G,t}\cap\mathcal{E}_{\mathrm{mg},t}\cap\mathcal{E}_{\mathrm{app},t}$. The union bound applies to these possibly dependent events and gives

$$\mathbb{P}_\tau(\mathcal{E}_t^c) \leq \mathbb{P}_\tau(\mathcal{E}_{G,t}^c) + \mathbb{P}_\tau(\mathcal{E}_{\mathrm{mg},t}^c) + \mathbb{P}_\tau(\mathcal{E}_{\mathrm{app},t}^c) \leq \alpha_{\mathrm{up},t} + \delta_{\mathrm{sb},t} + \delta_{\mathrm{mg},t} + \alpha_{\mathrm{app},t}.$$

Step 4: convert the score bound into a uniform loss gap. Fixing arbitary $\theta \in \Theta_t$ and setting $b := \mathrm{vec}(\theta - \theta_t^*) \in \mathbb{R}^D$, one has

$$\begin{aligned} s_t(\theta) - s_t(\theta_t^*) &= \sum_{i=0}^{n-1}\left\{\|a_i + w_i - \mathbf{\Phi}_i b\|_2^2 - \|a_i + w_i\|_2^2\right\} \\ &= \sum_{i=0}^{n-1}\left\{b^\top \mathbf{\Phi}_i^\top \mathbf{\Phi}_i b - 2b^\top \mathbf{\Phi}_i^\top (a_i + w_i)\right\} \\ &= b^\top G b - 2b^\top S. \end{aligned}$$

On $\mathcal{E}_t$, Cauchy–Schwarz inequality yields

$$|b^\top S| = |(G^{1/2}b)^\top (G^{-1/2}S)| \leq \|b\|_G \|S\|_{G^{-1}} \leq z_t \|b\|_G.$$

Writing $q := \|b\|_G$, we obtain

$$s_t(\theta) - s_t(\theta_t^*) \geq q^2 - 2z_t q \geq \frac{1}{2}q^2 - 2z_t^2,$$

where the last inequality follows from $2z_t q \leq q^2/2 + 2z_t^2$. If $\|\theta - \theta_t^*\|_{\mathrm{F}} \geq \epsilon_t$, then $\|b\|_2 = \|\theta - \theta_t^*\|_{\mathrm{F}} \geq \epsilon_t$, and $q^2 = b^\top G b \geq g_{-,t}\|b\|_2^2 \geq g_{-,t}\epsilon_t^2$. It follows that, simultaneously for every $\theta \in \Theta_t$ satisfying $\|\theta - \theta_t^*\|_{\mathrm{F}} \geq \epsilon_t$,

$$s_t(\theta) - s_t(\theta_t^*) \geq \frac{1}{2}g_{-,t}\epsilon_t^2 - 2z_t^2 =: \mathfrak{G}_t. \tag{46}$$

Step 5: bound the Gibbs mass outside the contraction ball and conclude contraction. Defining $\mathcal{O}_t(\epsilon_t) := \{\theta \in \Theta_t : \|\theta - \theta_t^*\|_{\mathrm{F}} \geq \epsilon_t\}$, we have

$$\begin{aligned} p_t(\mathcal{O}_t(\epsilon_t)) &= \frac{\displaystyle\sum_{\theta\in\mathcal{C}_t\cap\mathcal{O}_t(\epsilon_t)} \exp\{-\eta_t s_t(\theta)\}}{\displaystyle\sum_{\theta\in\mathcal{C}_t}\exp\{-\eta_t s_t(\theta)\}} \\ &\leq \frac{|\mathcal{C}_t \cap \mathcal{O}_t(\epsilon_t)|\exp\{-\eta_t[s_t(\theta_t^*) + \mathfrak{G}_t]\}}{\exp\{-\eta_t s_t(\widehat{\theta}_t)\}} \\ &\leq \mathcal{M}_t \exp(-\eta_t \mathfrak{G}_t)\exp\{-\eta_t[s_t(\theta_t^*) - s_t(\widehat{\theta}_t)]\} \\ &\leq \mathcal{M}_t \exp(-\eta_t \mathfrak{G}_t), \end{aligned}$$

where last inequality holds due to $s_t(\theta_t^*) - s_t(\widehat{\theta}_t) \geq 0$. It remains to remove the conditioning on the update data. Recall that $\mathscr{D}_t$ be the sigma-field generated by the data entering p_t, excluding θ_t. For fixed update data, s_t is continuous on the compact set Θ_t. The measurable maximum theorem yeilds the prespecified lexicographic minimizer $\widehat{\theta}_t$ measurable. The measurable construction in Definition 4 then makes $\mathcal{C}_t$, and hence p_t, $\mathscr{D}_t$-measurable. The matrices and vectors defining $\mathcal{E}_t$ use only the training observations $(x_i, u_i, x_{i+1})_{i=0}^{n-1}$ and the deterministic maps f and $f_{\theta_t^*}$; therefore $\mathcal{E}_t \in \mathscr{D}_t$. By construction, U_t^{G} is independent of $\mathscr{D}_t$, and the inverse-CDF draw satisfies $\mathcal{L}(\theta_t \mid$

$\mathscr{D}_t) = p_t$. For $B_t := \{\|\theta_t - \theta_t^*\|_{\mathrm{F}} \geq \epsilon_t\}$, we can deduce that

$$\begin{aligned}
\mathbb{P}_\tau(B_t) &= \mathbb{E}_\tau[\mathbb{P}(B_t \mid \mathscr{D}_t)] \\
&= \mathbb{E}_\tau[p_t(\mathcal{O}_t(\epsilon_t))] \\
&= \mathbb{E}_\tau[\mathbf{1}_{\mathcal{E}_t} p_t(\mathcal{O}_t(\epsilon_t))] + \mathbb{E}_\tau\big[\mathbf{1}_{\mathcal{E}_t^c} p_t(\mathcal{O}_t(\epsilon_t))\big] \\
&\leq \mathcal{M}_t \exp(-\eta_t \mathfrak{G}_t)\mathbb{P}_\tau(\mathcal{E}_t) + \mathbb{P}_\tau(\mathcal{E}_t^c) \\
&\leq \mathcal{M}_t \exp(-\eta_t \mathfrak{G}_t) + \mathbb{P}_\tau(\mathcal{E}_t^c) \\
&\leq \alpha_{\mathrm{up},t} + \delta_{\mathrm{sb},t} + \alpha_{\mathrm{app},t} + \delta_{\mathrm{mg},t} + \mathcal{M}_t \exp(-\eta_t \mathfrak{G}_t) \\
&= q_t.
\end{aligned}$$

This proves (8). □

C.2 Proof of Proposition 1

We first give some notation used in proof of Proposition 1. Recalling the radii in Lemma 3, define

$$R_t := 1 \vee R_{x,t+2M-1}\left(\frac{1}{5(t+1)^2}\right) \vee R_{u,t+2M-1}\left(\frac{1}{5(t+1)^2}\right), \qquad \mathcal{K}_t := [-R_t, R_t]^{d_x+d_u}. \tag{47}$$

Furthermore, we choose

$$\begin{aligned}
&\alpha_{\mathrm{up},t} = \delta_{\mathrm{sb},t} = \alpha_{\mathrm{app},t} = \delta_{\mathrm{mg},t} = \frac{1}{5(t+1)^2}, \qquad \lambda_t = \eta_t = 1, \qquad \varepsilon_{\mathrm{cov},t} = n_t^{-1}, \\
&\epsilon_t^2 = \frac{4z_t^2 + 2\log\big(5(t+1)^2\mathcal{M}_t\big)}{g_{-,t}}.
\end{aligned} \tag{48}$$

For $m \in \mathbb{N}$ and $z = (x^\top, u^\top)^\top = (z_1, \ldots, z_{d_x+d_u})^\top \in \mathcal{K}_t$, define

$$\mathcal{P}_m(\mathcal{K}_t) := \Big\{v : \mathcal{K}_t \to \mathbb{R} : \quad v(z) = \sum_{\ell_1=0}^{m} \cdots \sum_{\ell_{d_x+d_u}=0}^{m} c_{\ell_1,\ldots,\ell_{d_x+d_u}} \prod_{j=1}^{d_x+d_u} z_j^{\ell_j}, c_{\ell_1,\ldots,\ell_{d_x+d_u}} \in \mathbb{R}\Big\},$$

$$[\mathcal{P}_m(\mathcal{K}_t)]^{d_x} := \big\{z \mapsto (v_1(z), \ldots, v_{d_x}(z))^\top : v_j \in \mathcal{P}_m(\mathcal{K}_t), \quad 1 \leq j \leq d_x\big\}.$$

We introduce the following assumption.

Assumption 4. For fixed $s > 0$, $a, b \geq 0$ and $L, C_{\dim}, C_\phi, C_0 \geq 1$, there exist $m_t \in \mathbb{N}_+$ such that

$$\begin{aligned}
&(f - f^{\mathrm{pub}})|_{\mathcal{K}_t} \in \mathcal{H}^s(\mathcal{K}_t, L(1+R_t)^a), \qquad \mathcal{P}_{m_t}(\mathcal{K}_t) \subseteq \mathrm{span}\{\phi_1^{(t)}, \ldots, \phi_{d_t}^{(t)}\}, \\
&d_t \leq C_{\dim}(m_t+1)^{d_x+d_u} \wedge n_t, \qquad \sup_{\mathcal{K}_t} \|\phi^{(t)}\|_2 \leq C_\phi\sqrt{d_t}, \\
&k_t \leq C_0, \quad 0 \prec \Gamma_{\mathrm{PE},t} \preceq C_0 I_{d_t}, \quad \frac{\kappa_{\mathrm{PE},t}^2}{\lambda_{\min}(\Gamma_{\mathrm{PE},t})} \leq C_0 d_t^b, \\
&\max\{1, \mathrm{diam}(\Theta_t), \|\Gamma_{\mathrm{up},t}\|_{\mathrm{op}}\} \leq C_0(t+1)^{C_0}, \quad \frac{d_t^{b+1}\log(e+t)}{t} = o(1).
\end{aligned} \tag{49}$$

The compact set Θ_t contains a coefficient matrix θ_t^* representing a best uniform approximation to $f - f^{\mathrm{pub}}$ from $[\mathcal{P}_{m_t}(\mathcal{K}_t)]^{d_x}$.

Based on Assumption 4 and Equation (3.7) of Belloni et al. (2015), we have

$$\sup_{(x,u)\in\mathcal{K}_t} \|f(x,u) - f_{\theta_t^*}(x,u)\|_2 \leq C_{\mathrm{J}} L(1+R_t)^a R_t^s d_t^{-s/(d_x+d_u)}, \qquad \theta_t^* \in \Theta_t, \tag{50}$$

where $C_{\mathrm{J}} > 0$ depends only on $d_x, d_u, s, C_{\dim}$. In the subsequent analysis, we set $\Delta_t = C_{\mathrm{J}} L(1+R_t)^a R_t^s d_t^{-s/(d_x+d_u)}$.

Proof of Proposition 1. By Lemma 3, we have

$$q_0\left(\frac{1}{5(t+1)^2}\right) = V_{\mathrm{sta}}(x_0),$$

$$\ell_{t+2M-1}\left(\frac{1}{5(t+1)^2}\right) = \log\left(\frac{10\pi^2(d_x+d_u)(t+2M)^2(t+1)^2}{3}\right) \le C\log(e+t),$$
$$R_t \le C\sqrt{\log(e+t)}, \qquad L(1+R_t)^a R_t^s \le C[\log(e+t)]^{(a+s)/2},$$
$$n_t - 1 < H, \qquad R_{x,n_t-1}(\alpha_{\mathrm{app},t}) \vee R_{u,n_t-1}(\alpha_{\mathrm{app},t}) \le R_t,$$
$$\mathbb{P}_\tau((x_i,u_i) \in \mathcal{K}_t \text{ for all } 0 \le i < n_t) \ge 1 - \frac{1}{5(t+1)^2},$$

where constant $C \ge 1$ depends only on $d_x, d_u, s, a, b, L, C_{\dim}, C_\phi, C_0$. Together with (50), we have

$$\mathbb{P}_\tau\left(\frac{1}{n_t}\sum_{i=0}^{n_t-1}\|f(x_i,u_i) - f_{\theta_t^*}(x_i,u_i)\|_2^2 > \Delta_t^2\right) \le \alpha_{\mathrm{app},t}.$$

We next verify the sample-size condition (7) required by Theorem 1. By using (49) and (48), we have

$$p_t^{\mathrm{sb}} = \frac{1}{4\kappa_{\mathrm{PE},t}}, \qquad \kappa_{\mathrm{PE},t}^2 \le C_0^2 d_t^b, \qquad \lambda_{\min}(\Gamma_{\mathrm{PE},t})^{-1} \le C_0 d_t^b,$$
$$\overline{\Gamma}_t = \tfrac{1}{2}\Gamma_{\mathrm{PE},t} + 5d_t(t+1)^2\Gamma_{\mathrm{up},t},$$
$$\log\det(2\overline{\Gamma}_t\Gamma_{\mathrm{PE},t}^{-1}) \le d_t\log\big(1 + 10C_0^2 d_t^{b+1}(t+1)^{C_0+2}\big) \le Cd_t\log(e+t).$$

Consequently, it holds that

$$\begin{aligned}
&\frac{10k_t}{(p_t^{\mathrm{sb}})^2}\left[\log\frac{1}{\delta_{\mathrm{sb},t}} + 2d_t\log\frac{10}{p_t^{\mathrm{sb}}} + \log\det(2\overline{\Gamma}_t\Gamma_{\mathrm{PE},t}^{-1})\right] \\
=&160k_t\kappa_{\mathrm{PE},t}^2\left[\log\big(5(t+1)^2\big) + 2d_t\log(40\kappa_{\mathrm{PE},t}) + \log\det(2\overline{\Gamma}_t\Gamma_{\mathrm{PE},t}^{-1})\right] \\
\le& Cd_t^{b+1}\log(e+t).
\end{aligned} \tag{51}$$

By the dimension-growth condition, we have

$$\frac{Cd_t^{b+1}\log(e+t)}{n_t} = \frac{t}{t-M-1}\frac{Cd_t^{b+1}\log(e+t)}{t} \longrightarrow 0.$$

Hence there exists $t_0 \ge 2(M+1)$, independent of H, such that, for every $t \in \mathcal{T}_H$ with $t \ge t_0$,

$$n_t = t - M - 1 \ge t/2, \qquad n_t \ge Cd_t^{b+1}\log(e+t).$$

Together with (51), this verifies (7). Consequently,

$$n_t \ge \frac{20k_td_t}{(p_t^{\mathrm{sb}})^2}\log\frac{10}{p_t^{\mathrm{sb}}} > 2k_t, \qquad k_t\lfloor n_t/k_t\rfloor \ge n_t - k_t \ge n_t/2.$$

Having verified (7) for $t \ge t_0$, we next bound the quantities in Theorem 1. By $k_t\lfloor n_t/k_t\rfloor \ge n_t/2$ and Assumption 4, we have

$$g_{-,t} = \frac{k_t\lfloor n_t/k_t\rfloor\lambda_{\min}(\Gamma_{\mathrm{PE},t})}{512\kappa_{\mathrm{PE},t}^2} \ge C^{-1}td_t^{-b},$$
$$\log\mathcal{M}_t \le d_xd_t\log(1 + 2n_t\operatorname{diam}(\Theta_t)) \le Cd_t\log(e+t),$$
$$\begin{aligned}
z_t^2 &\le 2n_t\Delta_t^2 + 2\|\Sigma_w\|_{\mathrm{op}}(1+g_{-,t}^{-1})\left[d_xd_t\log(1+n_t\|\overline{\Gamma}_t\|_{\mathrm{op}}) + 2\log\big(5(t+1)^2\big)\right] \\
&\le C\left[t\Delta_t^2 + d_t\log(e+t)\right].
\end{aligned}$$

Substituting these bounds into (48) gives

$$\epsilon_t^2 \le Cd_t^b\left[[\log(e+t)]^{a+s}d_t^{-2s/(d_x+d_u)} + \frac{d_t\log(e+t)}{t}\right]. \tag{52}$$

By using (48) and the definitions of $\mathfrak{G}_t$ and q_t, we have

$$\mathfrak{G}_t = \tfrac{1}{2}g_{-,t}\epsilon_t^2 - 2z_t^2 = \log\big(5(t+1)^2\mathcal{M}_t\big),$$

$$\mathcal{M}_t e^{-\eta_t \mathfrak{G}_t} = \frac{1}{5(t+1)^2}, \qquad q_t = (t+1)^{-2}.$$

Theorem 1 therefore gives $\mathbb{P}_\tau(\|\theta_t - \theta_t^*\|_{\mathrm{F}} \le \epsilon_t) \ge 1 - (t+1)^{-2}$. On the event $\{\|\theta_t - \theta_t^*\|_{\mathrm{F}} \le \epsilon_t\}$, by using (50) and Assumption 4, we get

$$\sup_{\mathcal{K}_t} \|f - f_{\theta_t}\|_2 \le \sup_{\mathcal{K}_t} \|f - f_{\theta_t^*}\|_2 + \|\theta_t - \theta_t^*\|_{\mathrm{F}} \sup_{\mathcal{K}_t} \|\phi^{(t)}\|_2 \le \Delta_t + C_\phi \sqrt{d_t}\, \epsilon_t.$$

Using (52), $L(1+R_t)^a R_t^s \le C[\log(e+t)]^{(a+s)/2}$, and $d_t^{(b+1)/2} \ge 1$, we obtain

$$\sup_{\mathcal{K}_t} \|f - f_{\theta_t}\|_2 \le C \left[[\log(e+t)]^{(a+s)/2} d_t^{(b+1)/2 - s/(d_x+d_u)} + d_t^{1+b/2} \sqrt{\frac{\log(e+t)}{t}} \right]$$

on the event $\{\|\theta_t - \theta_t^*\|_{\mathrm{F}} \le \epsilon_t\}$. This proves (9).

□

Remark 5. Suppose

$$s > \frac{(b+1)(d_x+d_u)}{2}, \qquad d_t \asymp t^{(d_x+d_u)/(2s+d_x+d_u)}. \tag{53}$$

These choices satisfy $d_t^{b+1} \log(e+t) = o(t)$. Substituting (53) into (9), for all sufficiently large $t \in \mathcal{T}_H$, we obtain

$$\mathbb{P}_\tau \left(\sup_{(x,u)\in\mathcal{K}_t} \|f(x,u) - f_{\theta_t}(x,u)\|_2 \le C \frac{[\log(e+t)]^{\max\{a+s,1\}/2}}{t^{\{s-(b+1)(d_x+d_u)/2\}/(2s+d_x+d_u)}} \right) \ge 1 - (t+1)^{-2}. \tag{54}$$

Both the error bound and the failure probability tend to zero.

C.3 Formal statement and proof of Theorem 2

Proof. Fix an update time $t < H$, write $n_{\mathrm{blk},t} := t/(2M)$, and denote $s_k := 2(k+1)M$, $r_k := s_k - M - 1$, $\mathcal{I}_k := \{r_k, \dots, s_k - 2\}$,and $k = 0, \dots, n_{\mathrm{blk},t} - 1$. For each update s fix a deterministic radius $\epsilon_s^{\mathrm{cov}} > 0$, while for $\alpha \in (0,1)$, using the deterministic radii in Lemma 3, define

$$\begin{aligned}
R_s &:= \sup_{\theta,\theta' \in \Theta_s} \|\theta - \theta'\|_{\mathrm{op}}, \quad \bar{d}_s(\alpha) := \sup_{\substack{\|x\|_\infty \le R_{x,s}(\alpha) \\ \|u\|_\infty \le R_{u,s}(\alpha)}} \|\phi^{(s)}(x,u)\|_2, \\
\varrho_s(\alpha) &:= \sup_{\substack{\|x\|_\infty \le R_{x,s}(\alpha) \\ \|u\|_\infty \le R_{u,s}(\alpha)}} \|f(x,u) - f_{\theta_s^*}(x,u)\|_2, \\
L_s(\alpha) &:= \varrho_s(\alpha) + \bar{d}_s(\alpha)\epsilon_s, \quad U_s(\alpha) := \varrho_s(\alpha) + \bar{d}_s(\alpha) R_s.
\end{aligned}$$

The covariance recursion unrolls exactly as

$$\widehat{\Sigma}_{w,t} = \frac{2}{t} \sum_{k=0}^{n_{\mathrm{blk},t}-1} \sum_{i \in \mathcal{I}_k} (w_i + d_i)(w_i + d_i)^\top \tag{55}$$

with $d_i := f(x_i,u_i) - f_{\theta_{s_k}}(x_i,u_i)$ for any $i \in \mathcal{I}_k$. Define the event

$$\mathcal{E}_t(\alpha) := \bigcap_{k=0}^{n_{\mathrm{blk},t}-1} \bigcap_{i \in \mathcal{I}_k} \{\|x_i\|_\infty \le R_{x,s_k}(\alpha),\ \|u_i\|_\infty \le R_{u,s_k}(\alpha)\}.$$

Since every $i \in \mathcal{I}_k$ satisfies $i < s_k$, Lemma 3 gives $\mathbb{P}_\tau(\mathcal{E}_t(\alpha)^c) \le \alpha$. Denote $J_i := \mathbf{1}\{\|x_i\|_\infty \le R_{x,s_k}(\alpha),\ \|u_i\|_\infty \le R_{u,s_k}(\alpha)\}$, $\mathcal{H}_k := \{\|\theta_{s_k} - \theta_{s_k}^*\|_{\mathrm{F}} \le \epsilon_{s_k}\}$, and $B_t^{\mathrm{bad}}(\alpha) := \frac{2M}{t} \sum_{k=0}^{n_{\mathrm{blk},t}-1} U_{s_k}(\alpha)^2 \mathbf{1}_{\mathcal{H}_k^c}$. Then, Markov's inequality applies to $\mathcal{H}_k$ yields

$$\mathbb{P}_\tau(B_t^{\mathrm{bad}}(\alpha) > \zeta) \le \frac{2M}{t\zeta} \sum_{k=0}^{n_{\mathrm{blk},t}-1} U_{s_k}(\alpha)^2 q_{s_k} =: \Lambda_t(\alpha,\zeta).$$

For $i \in \mathcal{I}_k$, set

$$\chi_k := \mathbf{1}\left\{\frac{2M}{t}\sum_{\ell=0}^{k} U_{s_\ell}(\alpha)^2 \mathbf{1}_{\mathcal{H}_\ell^c} \le \zeta\right\}, \qquad \widetilde{d}_i := J_i \chi_k d_i.$$

Indeed, the Gibbs sampling kernel p_{s_k} is $\mathscr{F}_{r_k}$-measurable; since θ_{s_k} is generated from p_{s_k}, both θ_{s_k} and $\mathcal{H}_k$ are measurable with respect to $\sigma(\mathscr{F}_{r_k} \cup \sigma(\theta_{s_k}))$. For $i \in \mathcal{I}_k$, define $\widetilde{\mathscr{F}}_i := \sigma(\mathscr{F}_i \cup \sigma(\theta_{s_0}, \dots, \theta_{s_k}))$. Then $\widetilde{d}_i$ is $\widetilde{\mathscr{F}}_i$-measurable and

$$\mathbb{E}[w_i \mid \widetilde{\mathscr{F}}_i] = 0, \qquad \mathbb{E}[\exp(\lambda w_i^{(m)}) \mid \widetilde{\mathscr{F}}_i] \le \exp\left(\frac{\lambda^2 \sigma_\infty^2}{2}\right) \tag{56}$$

with $\sigma_\infty^2 := \max_{\ell \in [d_x]} [\Sigma_w]_{\ell\ell}$.

On $\{J_i = 1\}$, compactness of Θ_{s_k} gives

$$\|d_i\|_2 \le \begin{cases} L_{s_k}(\alpha), & \mathcal{H}_k, \\ U_{s_k}(\alpha), & \mathcal{H}_k^c. \end{cases}$$

Because the partial sums defining χ_k are nondecreasing, it holds that

$$\begin{aligned} \sum_{k=0}^{n_{\mathrm{blk},t}-1} \sum_{i \in \mathcal{I}_k} \|\widetilde{d}_i\|_2^2 &\le M \sum_{k=0}^{n_{\mathrm{blk},t}-1} L_{s_k}(\alpha)^2 + M \sum_{k=0}^{n_{\mathrm{blk},t}-1} U_{s_k}(\alpha)^2 \mathbf{1}_{\mathcal{H}_k^c} \chi_k \\ &\le M \sum_{k=0}^{n_{\mathrm{blk},t}-1} L_{s_k}(\alpha)^2 + \frac{t\zeta}{2}. \end{aligned} \tag{57}$$

Fix $m, n \in [d_x]$ and write $\mathcal{J}_t := \bigcup_{k=0}^{n_{\mathrm{blk},t}-1} \mathcal{I}_k$. We enumerate its elements in chronological order as $\mathcal{J}_t = \{i_1 < \cdots < i_N\}$ and define $X_\ell := w_{i_\ell}^{(m)} \widetilde{d}_{i_\ell}^{(n)}$, $q_\ell := \sigma_\infty^2 \big|\widetilde{d}_{i_\ell}^{(n)}\big|^2$ for any $1 \le \ell \le N$. By (56), we have

$$\mathbb{E}\Big[X_\ell \mid \widetilde{\mathscr{F}}_{i_\ell}\Big] = \widetilde{d}_{i_\ell}^{(n)} \mathbb{E}\Big[w_{i_\ell}^{(m)} \mid \widetilde{\mathscr{F}}_{i_\ell}\Big] = 0.$$

Thus, $(X_\ell)_{\ell=1}^N$ is a martingale-difference sequence. Moreover, since $\widetilde{d}_{i_\ell}^{(n)}$ is $\widetilde{\mathscr{F}}_{i_\ell}$-measurable, (56) implies that, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\lambda X_\ell) \mid \widetilde{\mathscr{F}}_{i_\ell}\right] = \mathbb{E}\left[\exp\left(\lambda \widetilde{d}_{i_\ell}^{(n)} w_{i_\ell}^{(m)}\right) \middle| \widetilde{\mathscr{F}}_{i_\ell}\right] \le \exp\left(\frac{\lambda^2 \sigma_\infty^2}{2} \big|\widetilde{d}_{i_\ell}^{(n)}\big|^2\right) = \exp\left(\frac{\lambda^2 q_\ell}{2}\right).$$

For fixed $\lambda \in \mathbb{R}$, define the random process

$$Z_0(\lambda) := 1, \quad Z_\ell(\lambda) := \exp\left\{\lambda \sum_{j=1}^{\ell} X_j - \frac{\lambda^2}{2} \sum_{j=1}^{\ell} q_j\right\}, \qquad 1 \le \ell \le N.$$

The preceding conditional MGF estimation gives

$$\mathbb{E}\Big[Z_\ell(\lambda) \mid \widetilde{\mathscr{F}}_{i_\ell}\Big] = Z_{\ell-1}(\lambda) \mathbb{E}\left[\exp\left(\lambda X_\ell - \frac{\lambda^2 q_\ell}{2}\right) \middle| \widetilde{\mathscr{F}}_{i_\ell}\right] \le Z_{\ell-1}(\lambda).$$

Hence, $(Z_\ell(\lambda))_{\ell=0}^N$ is a nonnegative supermartingale and, in particular, $\mathbb{E}_\tau[Z_N(\lambda)] \le 1$. Moreover, it follows that

$$\begin{aligned} \sum_{\ell=1}^{N} q_\ell &= \sigma_\infty^2 \sum_{k=0}^{n_{\mathrm{blk},t}-1} \sum_{i \in \mathcal{I}_k} \big|\widetilde{d}_i^{(n)}\big|^2 \\ &\le \sigma_\infty^2 \sum_{k=0}^{n_{\mathrm{blk},t}-1} \sum_{i \in \mathcal{I}_k} \|\widetilde{d}_i\|_2^2 \\ &\le \sigma_\infty^2 \left\{M \sum_{k=0}^{n_{\mathrm{blk},t}-1} L_{s_k}(\alpha)^2 + \frac{t\zeta}{2}\right\} =: V_{\mathrm{cr},t}(\alpha, \zeta), \end{aligned}$$

where the last inequality follows from (57). Therefore, for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}_\tau\left[\exp\left(\lambda \sum_{\ell=1}^{N} X_\ell\right)\right] = \mathbb{E}_\tau\left[Z_N(\lambda)\exp\left(\frac{\lambda^2}{2}\sum_{\ell=1}^{N} q_\ell\right)\right] \le \exp\left(\frac{\lambda^2}{2} V_{\mathrm{cr},t}(\alpha,\zeta)\right).$$

Suppose first that $V_{\mathrm{cr},t}(\alpha,\zeta) > 0$. By the Chernoff inequality, for every $\lambda > 0$ and $a > 0$,

$$\mathbb{P}_\tau\left(\sum_{\ell=1}^{N} X_\ell > a\right) \le \exp\left\{-\lambda a + \frac{\lambda^2}{2} V_{\mathrm{cr},t}(\alpha,\zeta)\right\}.$$

Optimizing the right-hand side at $\lambda = a/V_{\mathrm{cr},t}(\alpha,\zeta)$ yields

$$\mathbb{P}_\tau\left(\sum_{\ell=1}^{N} X_\ell > a\right) \le \exp\left(-\frac{a^2}{2V_{\mathrm{cr},t}(\alpha,\zeta)}\right).$$

Applying the same argument to $-X_\ell$ and combining the two one-sided bounds gives

$$\mathbb{P}_\tau\left(\left|\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k} w_i^{(m)}\widetilde{d}_i^{(n)}\right| > a\right) \le 2\exp\left(-\frac{a^2}{2V_{\mathrm{cr},t}(\alpha,\zeta)}\right).$$

We therefore have

$$\mathbb{P}_\tau\left(\left\|\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k} w_i\widetilde{d}_i^\top\right\|_{\mathrm{F}} > \frac{t\epsilon_t^{\mathrm{cov}}}{4}\right) \le 2d_x^2\exp\left(-\frac{t^2(\epsilon_t^{\mathrm{cov}})^2}{32d_x^2 V_{\mathrm{cr},t}(\alpha,\zeta)}\right). \tag{58}$$

On $\mathcal{E}_t(\alpha)\cap\{B_t^{\mathrm{bad}}(\alpha)\le\zeta\}$, one has $J_i = \chi_k = 1$ for every held-out index. Hence $\widetilde{d}_i = d_i$, and (58) controls the actual cross term:

$$\left\|\frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k}(w_i d_i^\top + d_i w_i^\top)\right\|_{\mathrm{op}} \le \frac{4}{t}\left\|\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k} w_i d_i^\top\right\|_{\mathrm{F}} \le \epsilon_t^{\mathrm{cov}}. \tag{59}$$

On the same event, we have

$$\left\|\frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k} d_i d_i^\top\right\|_{\mathrm{op}} \le \frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k}\|d_i\|_2^2 \le \underbrace{\frac{2M}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1} L_{s_k}(\alpha)^2}_{:=\beta_t(\alpha)} + B_t^{\mathrm{bad}}(\alpha) \le \beta_t(\alpha)+\zeta. \tag{60}$$

It remains to control the pure-noise term

$$N_t := \frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k}(w_i w_i^\top - \Sigma_w).$$

There are $n = t/2$ independent Gaussian summands. Let $\mathcal{N}(1/4)$ be a $1/4$-net of $\mathbb{S}^{d_x-1}$ with $|\mathcal{N}(1/4)| \le 9^{d_x}$. Then $\|N_t\|_{\mathrm{op}} \le 2\max_{v\in\mathcal{N}(1/4)}|v^\top N_t v|$. For fixed v, write $w_i = \Sigma_w^{1/2} g_i$ and $H_v := \frac{1}{n} I_n \otimes \left(\Sigma_w^{1/2} v v^\top \Sigma_w^{1/2}\right)$. Moreover, we may conclude that $\|H_v\|_{\mathrm{F}} \le \sqrt{\frac{2}{t}}\|\Sigma_w\|_{\mathrm{op}}$ and $\|H_v\|_{\mathrm{op}} \le \frac{2}{t}\|\Sigma_w\|_{\mathrm{op}}$. The Hanson–Wright inequality (see Theorem 1.1 of Rudelson & Vershynin (2013)) yields

$$\mathbb{P}_\tau(\|N_t\|_{\mathrm{op}} > \epsilon_t^{\mathrm{cov}}) \le 2\,9^{d_x}\exp\left[-c_{\mathrm{HW}} t\min\left\{\frac{(\epsilon_t^{\mathrm{cov}})^2}{8\|\Sigma_w\|_{\mathrm{op}}^2}, \frac{\epsilon_t^{\mathrm{cov}}}{4\|\Sigma_w\|_{\mathrm{op}}}\right\}\right]$$

with a universal constant c_{HW}. Define

$$\begin{aligned}
\mathcal{E}_{\mathrm{env}} &:= \mathcal{E}_t(\alpha),\\
\mathcal{E}_{\mathrm{bad}} &:= \{B_t^{\mathrm{bad}}(\alpha) \le \zeta\},\\
\mathcal{E}_{\mathrm{cross}} &:= \left\{ \left\| \sum_{k=0}^{n_{\mathrm{blk},t}-1} \sum_{i\in\mathcal{I}_k} w_i \widetilde{d}_i^{\top} \right\|_{\mathrm{F}} \le \frac{t\epsilon_t^{\mathrm{cov}}}{4} \right\},\\
\mathcal{E}_{\mathrm{noise}} &:= \{\|N_t\|_{\mathrm{op}} \le \epsilon_t^{\mathrm{cov}}\}.
\end{aligned}$$

Expanding (55) gives the identity

$$\widehat{\Sigma}_{w,t} - \Sigma_w = N_t + \frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k}(w_i d_i^{\top} + d_i w_i^{\top}) + \frac{2}{t}\sum_{k=0}^{n_{\mathrm{blk},t}-1}\sum_{i\in\mathcal{I}_k} d_i d_i^{\top}.$$

On $\mathcal{E}_{\mathrm{env}} \cap \mathcal{E}_{\mathrm{bad}} \cap \mathcal{E}_{\mathrm{cross}} \cap \mathcal{E}_{\mathrm{noise}}$, one has $\widetilde{d}_i = d_i$. Therefore (59), (60), and the triangle inequality imply $\|\widehat{\Sigma}_{w,t} - \Sigma_w\|_{\mathrm{op}} \le \beta_t(\alpha) + \zeta + 2\epsilon_t^{\mathrm{cov}}$. Hence, we can derive that

$$\begin{aligned}
&\mathbb{P}_\tau\left(\|\widehat{\Sigma}_{w,t} - \Sigma_w\|_{\mathrm{op}} > \beta_t(\alpha) + \zeta + 2\epsilon_t^{\mathrm{cov}}\right)\\
\le&\mathbb{P}_\tau(\mathcal{E}_{\mathrm{env}}^c) + \mathbb{P}_\tau(\mathcal{E}_{\mathrm{bad}}^c) + \mathbb{P}_\tau(\mathcal{E}_{\mathrm{cross}}^c) + \mathbb{P}_\tau(\mathcal{E}_{\mathrm{noise}}^c)\\
\le&\alpha + \min\{1, \Lambda_t(\alpha,\zeta)\} + 2d_x^2 \exp\left(-\frac{t^2(\epsilon_t^{\mathrm{cov}})^2}{32 d_x^2 V_{\mathrm{cr},t}(\alpha,\zeta)}\right)\\
&+ 2\,9^{d_x}\exp\left[-c_{\mathrm{HW}} t \min\left\{\frac{(\epsilon_t^{\mathrm{cov}})^2}{8\|\Sigma_w\|_{\mathrm{op}}^2}, \frac{\epsilon_t^{\mathrm{cov}}}{4\|\Sigma_w\|_{\mathrm{op}}}\right\}\right].
\end{aligned}$$

Setting $\alpha_t = \alpha$ and $\zeta_t = \zeta$, we get (10). □

C.4 Proof of Theorem 3

Proof. Based on the definition of $\mathcal{D}_t^H$, we have

$$\mathcal{J}^H(\pi^{\mathrm{A}}) - \mathcal{J}^{H,*} \le \mathbb{E}_\tau \sum_{t=0}^{H-1} \mathcal{D}_t^H(x_t, u_t). \tag{61}$$

For each $s \in \mathcal{T}_H$ and $s \le t < s + \ell_{s,H}$, split the expectation $\mathbb{E}_\tau[\mathcal{D}_t^H(x_t,u_t)]$ at $\mathcal{G}_s$. we get

$$\mathbb{E}_\tau[\mathcal{D}_t^H(x_t,u_t)] = \mathbb{E}_\tau[\mathbf{1}_{\mathcal{G}_s}\mathcal{D}_t^H(x_t,u_t)] + \mathbb{E}_\tau[\mathbf{1}_{\mathcal{G}_s^c}\mathcal{D}_t^H(x_t,u_t)].$$

By using Assumption 3 to $E_\tau[\mathbf{1}_{\mathcal{G}_s}\mathcal{D}_t^H(x_t,u_t)]$ and Hölder's inequality to $\mathbb{E}_\tau[\mathbf{1}_{\mathcal{G}_s^c}\mathcal{D}_t^H(x_t,u_t)]$, it holds that

$$\mathbb{E}_\tau \mathcal{D}_t^H(x_t,u_t) \le \beta_s^{\mathrm{Bell}}(\alpha) + C_s^{\mathrm{id}}\mathbb{E}_\tau\|e_t^{\mathrm{id}}\|_2^2 + C_s^{\mathrm{bad}}\delta_s^{1-1/p}. \tag{62}$$

The noise $\{e_t^{\mathrm{id}}\}_{t=0}^{H-1}$ are independent of $\mathscr{G}_\tau$, so $\mathbb{E}_\tau\|e_t^{\mathrm{id}}\|_2^2 = d_u\sigma_{\mathrm{id},t}^2$. Summing the B_0^{init} and $\mathbb{E}_\tau[\mathcal{D}_t^H(x_t,u_t)]$ from $t = 2M+1$ to $H-1$ and using (62), we get (15). □

D System specializations and regret certificates

D.1 Linear–quadratic specialization

In this subsection, we analyze the regret bound of Algorithm 1 in linear quadratic regulator (LQR) problem. We now introduce the model settings of LQR problem. For a stabilizable pair (A^*, B^*) (i.e., $\exists K$ s.t. $\rho(A^* + B^*K) < 1$) with $A^* \in \mathbb{R}^{d_x \times d_x}$ and $B^* \in \mathbb{R}^{d_x \times d_u}$, consider a linear dynamic model:

$$\begin{aligned}
&x_{t+1} = A^* x_t + B^* u_t + w_t, \quad c(x_t,u_t) = x_t^{\top} Q x_t + u_t^{\top} R u_t, \quad h(x_H) \equiv 0,\\
&Q \succ 0, \quad R \succ 0, \quad t = 0,\dots,H-1,
\end{aligned} \tag{63}$$

In this example, we set $f^{\mathrm{pub}}(x,u) \equiv 0$, $\mu_{\theta_{-1}}(x_0,u_0,\dots,x_i) \equiv 0, 0 \le i < 2M \wedge H$, $\theta_t^* = \theta^* := (A^*, B^*)$, and $\phi^{(t)}(x,u) = (x^\top, u^\top)^\top, t = 0,\dots,H-1$. Meanwhile, we set $\Theta_t = \Theta$ for any $t = 0,\dots,H-1$, where Θ is a compact and convex subset of

$$\left\{(A,B) \in \mathbb{R}^{d_x \times d_x} \times \mathbb{R}^{d_x \times d_u} \;\middle|\; \exists K \in \mathbb{R}^{d_u \times d_x} \text{ s.t. } \rho(A+BK) < 1\right\}$$

satisfying Assumption 5. Furthermore, we set $M = 3, \eta_t = 1, \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \sigma_{\mathrm{id},t}^2 = \sigma_0^2 (t+1)^{-1/2}, t = 0,\dots,H-1$ with $\sigma_0 > 0$.

For the linear setting, we introduce a specific implementation of the `FiniteHorizonSolve` step in Algorithm 1, which differs from the method described in Appendix B.1.3. In `FiniteHorizonSolve`, we solve the following optimal control problem

$$\begin{aligned}
\min_{\mu=\{\mu^{(i)}\}_{i=0}^{H-1} \in \prod_{i=0}^{H-1} \mathscr{C}_i} \quad & \mathcal{J}_t^{\mathrm{mod}}(\mu) := \frac{1}{H}\mathbb{E}_\tau\left[\sum_{i=0}^{H-1} x_i^\top Q x_i + u_i^\top R u_i\right] \\
\text{s.t.} \quad & x_0 \sim \nu_0, \\
& \widehat{w}_i \sim \mathcal{N}(0, \widehat{\Sigma}_{w,t}), && i = 0,\dots,H-1, \\
& v_i = \mathrm{Proj}_{\mathbb{B}^{d_u}(0,B_{\mathrm{ctl}})}\big(\mu^{(i)}(x_i)\big), && i = 0,\dots,H-1, \\
& x_{i+1} = A_s x_i + B_s(\widehat{\mathbf{K}} x_i + v_i) + \widehat{w}_i, && i = 0,\dots,H-1
\end{aligned}$$

with $\theta_s = (A_s, B_s)$ denoting the sample from `GibbsSample` in Algorithm 1 at update time $s \in \mathcal{T}_H$. Here, $\mathscr{C}_i$ denotes the candidate set of $\mu^{(i)}$. Set the candidate set of $\mu^{(i)}$ as

$$\mathscr{C}_i = \left\{x \mapsto Kx \;\middle|\; K \in \mathbb{R}^{d_u \times d_x}, x \in \mathbb{R}^{d_x}\right\}.$$

According to the previous study of the LQR problem Bertsekas et al. (1995), we can solve the above optimization problem on the feasible set $\mathscr{C}_i, i = 1,\dots,H-1$ via the following Riccati recursion:

$$\begin{aligned}
&\mathcal{P}_0(\theta_s) := \mathbf{0}^{d_x \times d_x}, \\
&\mathsf{S}_r(\theta_s) := R + B_s^\top \mathcal{P}_r(\theta_s) B_s, \\
&\mathsf{K}_r(\theta_s) := -\mathsf{S}_r(\theta_s)^{-1} B_s^\top \mathcal{P}_r(\theta_s) A_s, \\
&\mathcal{P}_{r+1}(\theta_s) := Q + A_s^\top \mathcal{P}_r(\theta_s) A_s - A_s^\top \mathcal{P}_r(\theta_s) B_s \mathsf{S}_r(\theta_s)^{-1} B_s^\top \mathcal{P}_r(\theta_s) A_s,\ r = 0,\dots,H-1. \quad (64)
\end{aligned}$$

With the above recursion (64), we return $\mu_{\theta_s}^{(i)}(x) = (\mathsf{K}_{H-i-1}(\theta_s) - \widehat{\mathbf{K}})x$ for any $x \in \mathbb{R}^{d_x}$ and any $i = 0,\dots,H-1$. The subsequent part of Algorithm 1 follows the discussion in Appendix B.1.1 and B.1.2.

Denoting the policy from Algorithm 1 by $\pi^{\mathrm{A_{LQR}}}$, we propose the following assumption and proposition.

Assumption 5. For model 63, it holds that $\theta^* \in \Theta$. For every $i = 0,\dots,H-1$, the initial state and process noise satisfy

$$x_0 \sim \nu_0 = \mathcal{N}(0, I_{d_x}), \quad w_i \sim \mathcal{N}(0, \Sigma_w), \quad \lambda_w I \preceq \Sigma_w \preceq \bar{\lambda}_w I, \quad t = 0,\dots,H-1$$

with $0 < \lambda_w \le \bar{\lambda}_w$. Meanwhile, let $\bar{K}$ be the optimal infinite-horizon LQR gain for (A^*, B^*, Q, R). There exist $\bar{P} \in \mathbb{R}^{d_x \times d_x}$ and $0 < \bar{\rho} < 1$, independent of H, such that

$$\bar{P} = \bar{P}^\top \succ 0, \qquad (A^* + B^* K)^\top \bar{P} (A^* + B^* K) \preceq \bar{\rho}\bar{P}, \quad K \in \{\widehat{\mathbf{K}}, \bar{K}\}.$$

Proposition 3 (LQ continuation bound)**.** *Under Assumption 5 and the above model setting, Assumptions 1– 3 hold, and it holds that*

$$\mathcal{J}^H(\pi^{\mathrm{A_{LQR}}}) - \mathcal{J}^{H,*} \le \widetilde{O}(H^{-1/2}). \quad (65)$$

Proof. **Step 1.** We first verify Assumption 1. For $f(x,u) = A^* x + B^* u$, set $\mathscr{L} := \max\left\{1, \max_{[A\ B] \in \Theta} \|B\|_{\mathrm{op}}\right\}$. Then we have

$$f(0,0) = 0, \qquad \|f(x,u) - f(x,u')\|_2 \le \|B^*\|_{\mathrm{op}} \|u - u'\|_2 \le \mathscr{L}\|u - u'\|_2,$$

which verifies Assumption 1(a). Meanwhile, since (A^*, B^*) is stabilizable, we can choose $K_\star$ and $\bar{q}$ such that $\rho(A^* + B^* K_\star) < \bar{q} < 1$. The discrete-time Lyapunov theorem Boyd et al. (1994) gives

$$(A^* + B^* K_\star)^\top P_\star (A^* + B^* K_\star) = \bar{q}^2 (P_\star - I_{d_x}) \preceq \bar{q}^2 P_\star$$

with $P_\star := \sum_{j=0}^{\infty} \bar{q}^{-2j} \left[(A^* + B^* K_\star)^j\right]^\top (A^* + B^* K_\star)^j \succeq I_{d_x}$. Choose

$$\mathscr{K} \geq \|K_\star\|_{\mathrm{op}}, \qquad \underline{v} = 1, \qquad \overline{v} \geq \lambda_{\max}(P_\star), \qquad q_1 = \frac{1+\bar{q}}{2}, \quad q_2 = \frac{1-\bar{q}}{2}.$$

With the above parameters, we get $(K_\star, P_\star) \in \mathfrak{S}$, $0 < q_2 < q_1 < 1$, and

$$\|f(x, K_\star x)\|_{P_\star} \leq \bar{q} \|x\|_{P_\star} = (q_1 - q_2) \|x\|_{P_\star}, \qquad x \in \mathbb{R}^{d_x}.$$

This verifies Assumption 1(b).

Step 2. We now analyze the Lipschitz property of controller for the subsequent proof. For $\theta = (A, B) \in \Theta$, let $\mathcal{P}_\infty(\theta) := \lim_{r \to \infty} \mathcal{P}_r(\theta)$. Note that Θ is compact. Based on the continuity of $\mathcal{P}_\infty(\theta)$ with respect to θ, we have $\| \left(1 + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x} \|_{\mathrm{op}} < \infty$. Consequently, it holds that

$$Q \preceq \mathcal{P}_r(\theta) \preceq \left(Q + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x}, \qquad r \geq 1, \quad \theta \in \Theta. \tag{66}$$

For a horizon $r \geq 1$, defining

$$\Psi_0 := I_{d_x}, \qquad \Psi_{j+1} := (A + B \mathsf{K}_{r-j-1}(\theta)) \Psi_j, \qquad 0 \leq j < r$$

and using (66), we have

$$\sum_{j=0}^{r-1} \Psi_j^\top \Psi_j \preceq \frac{\| \left(Q + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x} \|_{\mathrm{op}}}{\lambda_{\min}(Q)} I_{d_x},$$

$$\sup_{r \geq 0,\, \theta \in \Theta} \|D_\theta \mathcal{P}_r(\theta)\|_{\mathrm{op}} \leq \frac{2 \| \left(Q + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x} \|_{\mathrm{op}}^2 \sqrt{1 + K_{\max}^2}}{\lambda_{\min}(Q)} \cdot \sup_{[A\ B] \in \Theta} (\|A\|_{\mathrm{op}} + \|B\|_{\mathrm{op}} K_{\max}) < \infty$$

with $K_{\max} = \max\left\{\mathscr{K}, \sqrt{\frac{\| \left(Q + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x} \|_{\mathrm{op}}}{\lambda_{\min}(R)}}\right\}$. Here $D_\theta \mathcal{P}_r(\theta)$ denotes the Jacobian matrix with respect to θ. Differentiating K_r by using $\mathsf{K}_r = -(\mathsf{S}_r)^{-1} B^\top \mathcal{P}_r A$, $\mathsf{S}_r \succeq R$, we can find finite L_K and $\bar{s} := \|R\|_{\mathrm{op}} + \mathscr{L}^2 \| \left(Q + \sup_{\theta \in \Theta} \|\mathcal{P}_\infty(\theta)\|_{\mathrm{op}}\right) I_{d_x} \|_{\mathrm{op}}$, such that, for any $\theta, \widetilde{\theta} \in \Theta$, it holds that

$$\begin{aligned} &\|\mathsf{K}_r(\theta) - \mathsf{K}_r(\widetilde{\theta})\|_{\mathrm{op}} \leq L_K \|\theta - \widetilde{\theta}\|_{\mathrm{F}}, \\ &\|\mathsf{K}_r(\theta)\|_{\mathrm{op}} \leq K_{\max}, \qquad \|\mathsf{S}_r(\theta)\|_{\mathrm{op}} \leq \bar{s}, \qquad r \geq 0. \end{aligned} \tag{67}$$

These constants are independent of H.

Step 3. We now verify approximation error (5). For the basis functions $\phi^{(t)} = [x^\top, u^\top]^\top$, $t = 0, \ldots, H-1$, we have $d_t = d_x + d_u$. Since $\theta_t^* = [A^*\ B^*]$, we have $\Delta_t = \alpha_{\mathrm{app},t} = 0$.

Step 4. We next verify Assumption 2. Firstly, we verify the BMSB condition (6) of Assumption 2 instead of the two moment conditions. Nonexpansiveness of projection gives that

$$\begin{aligned} &\|\widehat{\mathbf{K}} x + \mathrm{Proj}_{\mathbb{B}(0, B_{\mathrm{ctl}})}((\mathsf{K}_r(\theta_s) - \widehat{\mathbf{K}}) x) - \widehat{\mathbf{K}} x' + \mathrm{Proj}_{\mathbb{B}(0, B_{\mathrm{ctl}})}((\mathsf{K}_r(\theta_s) - \widehat{\mathbf{K}}) x')\|_2 \\ \leq &(2\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max}) \|x - x'\| \end{aligned} \tag{68}$$

for any $x \in \mathbb{R}^{d_x}$, $r = 0, \ldots, H-1$, and $s \in \mathcal{T}_H$. For $s \in \mathcal{T}_H$, define

$$\gamma_s := \frac{\min\{\lambda_w, \sigma_0^2 (s+1)^{-1/2}\}}{64\left[1 + (2\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})^2\right]}.$$

To verify the persistence condition at update time s, we consider the historical feature rows

$$\phi_j^{(s)} := \phi^{(s)}(x_j, u_j), \qquad 0 \leq j < n_s.$$

Using the Lipschitz bound in (68), for every $v = (a^\top, b^\top)^\top \in \mathbb{S}^{d_s-1}$, we have

$$\mathbb{P}_\tau\Big(|v^\top \phi_j^{(s)}| < \sqrt{\gamma_s}\,\Big|\,\mathscr{H}_{j-1}\Big) \le \begin{cases} \dfrac{2\sqrt{\gamma_s}}{\sigma_{\mathrm{id},j}\|b\|_2\sqrt{2\pi}}, & \|b\|_2 \ge \dfrac{1}{2\sqrt{1+(2\|\widehat{\mathbf{K}}\|_{\mathrm{op}}+K_{\max})^2}}, \\ \dfrac{6\sqrt{\gamma_s}}{\sqrt{2\pi\lambda_w}}, & \text{otherwise,} \end{cases} < \frac{1}{2}.$$

Hence, for every $0 \le j < n_s$,

$$\mathbb{P}_\tau\Big(|v^\top \phi_j^{(s)}| \ge \sqrt{\gamma_s}\,\Big|\,\mathscr{H}_{j-1}\Big) \ge \frac{1}{2}.$$

Therefore, the direct persistence condition (6) holds with

$$k_s = 2, \qquad p_s^{\mathrm{sb}} = \frac{1}{4}, \qquad \Gamma_{\mathrm{PE},s} = 2\gamma_s I_{d_t}, \qquad \kappa_{\mathrm{PE},s} = 1.$$

Secondly, we analyze the upper bound of the Gram matrix $E\left[\sum_{i=0}^{n_s-1}\phi_i^{(s)}(\phi_i^{(s)})^\top\,\middle|\,\mathscr{H}_{-1}\right]$ for $s \in \mathcal{T}_H$. Define

$$\begin{aligned} m_{x,4} &:= \max_{i<H}(\mathbb{E}_\tau\|x_i\|_2^4)^{1/2} \\ &\le \frac{\overline{v}}{\underline{v}}\left[[d_x(d_x+2)]^{1/4} + \frac{\mathscr{L}\{B_{\mathrm{ctl}}(H) + \sigma_0[d_u(d_u+2)]^{1/4}\} + \sqrt{\overline{\lambda}_w}[d_x(d_x+2)]^{1/4}}{1-q_1}\right]^2, \\ 2C_{\phi,H} &:= (1+3\|\widehat{\mathbf{K}}\|_{\mathrm{op}}^2)m_{x,4} + 3B_{\mathrm{ctl}}^2(H) + 3d_u\sigma_0^2. \end{aligned}$$

For $s \in \mathcal{T}_H$, substituting $\phi^{(s)} = [x^\top, u^\top]^\top$ into $\mathbb{E}\left[\sum_{i=0}^{n_s-1}\phi_i^{(s)}(\phi_i^{(s)})^\top\,\middle|\,\mathscr{H}_{-1}\right]$ and noting that

$$\mathbb{E}_\tau[\|x_i\|_2^2 + \|u_i\|_2^2] \le 2C_{\phi,H}, \qquad 0 \le i < H,$$

we get

$$E\left[\sum_{i=0}^{n_s-1}\phi_i^{(s)}(\phi_i^{(s)})^\top\,\middle|\,\mathscr{H}_{-1}\right] \preceq n_s\Gamma_{\mathrm{up},s} = 2n_sC_{\phi,H}I_{d_t}.$$

That completes the verification of Assumption 2.

Step 5. We next derive the error bound of θ_s via Theorem 1. We choose t_0 sufficiently large and set

$$\begin{aligned} &\lambda_s = 1, \qquad \alpha_{\mathrm{up},s} = \delta_{\mathrm{sb},s} = \delta_{\mathrm{mg},s} = \frac{\delta}{64H^2}, \\ &g_{-,s} = \frac{(p_s^{\mathrm{sb}})^2}{32}\,2\lfloor n_s/2\rfloor\,2\gamma_s, \qquad \epsilon_s^2 = \begin{cases} (\mathrm{diam}_{\mathrm{F}}(\Theta))^2, & s < t_0, \\ \dfrac{4z_s^2 + 2\log(64H^2\mathcal{M}_s/\delta)}{g_{-,s}}, & s \ge t_0 \end{cases} \end{aligned}$$

with $\delta = (H+1)^{-2}$. Using $B_{\mathrm{ctl}}(H) = O(\sqrt{\log(H/\delta)})$ (see (74)), we have

$$\begin{aligned} &C_{\phi,H} = O(\log(H/\delta)), \quad g_{-,s}^{-1} = O(s^{-1/2}), \quad z_s^2 + \log\mathcal{M}_s = O(\log(H/\delta)), \\ &\text{RHS of (7)} = O(\log(H/\delta)), \qquad \epsilon_s^2 = O(\log(H/\delta)/\sqrt{s}) \quad (s \ge t_0). \end{aligned} \tag{69}$$

Set $t_0 = O(\log^2(H/\delta))$ sufficiently large such that the inequality (7) holds, the sample-size condition, $g_{-,s} \ge 1$, and

$$\sqrt{\frac{\lambda_{\max}(\bar{P})}{\lambda_{\min}(\bar{P})}}\mathscr{L}\left[L_K\epsilon_s + \|\mathsf{K}_r(\theta^*) - \bar{K}\|_{\mathrm{op}}\right] \le \frac{1-\sqrt{\bar{\rho}}}{2}, \qquad s, r \ge t_0 \tag{70}$$

holds for $s \ge t_0$. The choice of t_0 exists by (69) and $\mathsf{K}_r(\theta^*) \to \bar{K}$. Then, using (46), we have, for $s \ge t_0$,

$$\begin{aligned} &\mathfrak{G}_s = \tfrac{1}{2}g_{-,s}\epsilon_s^2 - 2z_s^2 = \log(64H^2\mathcal{M}_s/\delta) > 0, \\ &\mathcal{M}_s\exp(-\eta_s\mathfrak{G}_s) = \delta/(64H^2). \end{aligned}$$

Hence, using Theorem 1 when $s \geq t_0$ and the deterministic diameter bound $(\mathrm{diam}_{\mathrm{F}}(\Theta))^2$ when $s < t_0$, we have

$$\mathbb{P}_\tau(\|\theta_s - \theta^*\|_{\mathrm{F}} > \epsilon_s) \leq \frac{\delta}{16H^2}\mathbf{1}\{s \geq t_0\}. \tag{71}$$

Step 6. We now derive the error bound of $\widehat{\Sigma}_{w,s}$ via Theorem 2. Set

$$\alpha_s = \delta/(4H), \qquad \zeta_s = s^{-1/2} + \frac{2M}{sH^2}\sum_{\substack{j\in\mathcal{T}_H\\ t_0\leq j\leq s}} U_j(\alpha_s)^2,$$

$$\epsilon_s^{\mathrm{cov}} = \max\left\{4d_x\sqrt{\frac{\bar{\lambda}_w[\beta_s(\alpha_s)+\zeta_s]\log(32d_x^2/\delta)}{s}},\ \bar{\lambda}_w\sqrt{\frac{8\log(32\,9^{d_x}/\delta)}{c_{\mathrm{HW}}s}},\ \frac{4\bar{\lambda}_w\log(32\,9^{d_x}/\delta)}{c_{\mathrm{HW}}s}\right\}$$

with $s \in \mathcal{T}_H$. Then we get $\Lambda_s(\alpha_s,\zeta_s) \leq \delta/16,\ V_{\mathrm{cr},s}(\alpha_s,\zeta_s) \leq \frac{1}{2}\bar{\lambda}_w s[\beta_s(\alpha_s)+\zeta_s]$. Using Theorem 2, we get

$$\mathbb{P}_\tau\Big(\|\widehat{\Sigma}_{w,s} - \Sigma_w\|_{\mathrm{op}} > \beta_s(\alpha_s) + \zeta_s + 2\epsilon_s^{\mathrm{cov}}\Big) \leq \delta/4. \tag{72}$$

Step 7. We now verify Assumption 3. The exact LQR Bellman residual is

$$\mathcal{D}_t^H(x,u) = \frac{1}{H}\|u - \mathsf{K}_{H-t-1}(\theta^*)x\|^2_{\mathsf{S}_{H-t-1}(\theta^*)} \geq 0, \quad (x,u)\in\mathbb{R}^{d_x+d_u}, \quad t = 0,\dots,H-1. \tag{73}$$

Choose a fixed constant

$$C \geq \max\left\{1, \frac{4}{1-\sqrt{\bar{\rho}}}\sqrt{\frac{\lambda_{\max}(\bar{P})}{\lambda_{\min}(\bar{P})}}\left(\mathscr{L}\sigma_0\sqrt{2d_u} + \sqrt{2d_x\bar{\lambda}_w}\right)\right\},$$

and set

$$\begin{aligned}
&R_x^0(H,\alpha') := C\sqrt{\log\frac{16(d_x+d_u)H}{\alpha'}}, \qquad B_{\mathrm{ctl}}(H) = (\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})R_x^0(H,\alpha'),\\
&R_x(H,\alpha') = \sqrt{d_x}R_{x,H-1}(\alpha'/2), \quad r_e(H,\alpha') = \sigma_0\sqrt{2d_u\log\frac{16(d_x+d_u)H}{\alpha'}}
\end{aligned} \tag{74}$$

with $\alpha' = \delta/(4H)$. For $\|\theta_s - \theta^*\|_{\mathrm{F}} \leq \epsilon_s$ and $s, r \geq t_0$, Assumption 5 and (70) give

$$\left\|\bar{P}^{1/2}\left[A^* + B^*\{(1-\lambda)\widehat{\mathbf{K}} + \lambda\mathsf{K}_r(\theta_s)\}\right]\bar{P}^{-1/2}\right\|_{\mathrm{op}} \leq \frac{1+\sqrt{\bar{\rho}}}{2} < 1, \quad 0\leq\lambda\leq 1.$$

Then, by Lemma 3, (71) and the above inequality, it holds that

$$\begin{aligned}
&\max_{t<H}\|x_t\|_2 \leq R_x(H,\alpha'),\\
&\|x_{t_0+j}\|_2 \leq \sqrt{\frac{\lambda_{\max}(\bar{P})}{\lambda_{\min}(\bar{P})}}\left(\frac{1+\sqrt{\bar{\rho}}}{2}\right)^j R_x(H,\alpha') + \tfrac{1}{2}R_x^0(H,\alpha'), \quad t_0 + j \leq H - t_0
\end{aligned}$$

with probability at least $1-\alpha'$. Since $R_x/R_x^0 = O(1)$, choose t_1 by a fixed number such that

$$\|x_t\|_2 \leq R_x^0(H,\alpha'), \qquad \|\mu_{\theta_s}^{(t)}(x_t)\|_2 \leq B_{\mathrm{ctl}}(H), \qquad t_0 + t_1 \leq t \leq H - t_0.$$

Define $\mathcal{T}_H^\circ := \{s\in\mathcal{T}_H : s \geq t_0 + t_1, s + \ell_{s,H} \leq H - t_0\}$ and set

$$\begin{aligned}
&\mathcal{K}_s(\alpha')\\
&= \begin{cases}
\Big\{(x,u) : \|x\|_2 \leq R_x(H,\alpha'),\ \|u\|_2 \leq \|\widehat{\mathbf{K}}\|_{\mathrm{op}}R_x(H,\alpha') + B_{\mathrm{ctl}}(H) + r_e(H,\alpha')\Big\}\\
\qquad \cap\big\{(x,u) : \|x\|_2 \leq R_x^0(H,\alpha')\big\}, & s\in\mathcal{T}_H^\circ,\\
\Big\{(x,u) : \|x\|_2 \leq R_x(H,\alpha'),\ \|u\|_2 \leq \|\widehat{\mathbf{K}}\|_{\mathrm{op}}R_x(H,\alpha') + B_{\mathrm{ctl}}(H) + r_e(H,\alpha')\Big\}, & s\notin\mathcal{T}_H^\circ.
\end{cases}\\
&a_s = \epsilon_s, \quad c_s = \beta_s(\alpha') + \zeta_s + 2\epsilon_s^{\mathrm{cov}}, \quad \chi_s(\alpha') = 0, \quad B_{\mathrm{ctl}}(H) = O(\sqrt{\log H}).
\end{aligned} \tag{75}$$

On $\mathcal{G}_s(a_s, c_s, \alpha')$, projection $\mathrm{Proj}_{\mathbb{B}(0,B_{\mathrm{ctl}})}(\cdot)$ is inactive for $s \in \mathcal{T}_H^\circ$. For $s \notin \mathcal{T}_H^\circ$, it holds that

$$\|u_t - \mathsf{K}_{H-t-1}(\theta^*)x_t\|_2 \le (\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})R_x(H,\alpha') + B_{\mathrm{ctl}}(H) + \|e_t^{\mathrm{id}}\|_2.$$

Thus Assumption 3 holds with

$$\begin{aligned}
&\beta_s^{\mathrm{Bell}}(\alpha') = \frac{2\bar{s}}{H}\Big[R_x(H,\alpha')^2 c_s^2 + L_K^2 R_x^0(H,\alpha')^2 \epsilon_s^2 \mathbf{1}\{s \in \mathcal{T}_H^\circ\} \\
&\qquad\qquad + \{(\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})R_x(H,\alpha') + B_{\mathrm{ctl}}(H)\}^2 \mathbf{1}\{s \notin \mathcal{T}_H^\circ\}\Big], \\
&C_s^{\mathrm{bad}} = \max_{s \le t < s+\ell_{s,H}} \|\mathcal{D}_t^H(x_t,u_t)\|_{L^2} \le \frac{C_H}{H}, \qquad B_0^{\mathrm{init}} \le \frac{(2M \wedge H)C_H}{H}, \\
&C_H := 3\bar{s}\left[(\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})^2 m_{x,4} + B_{\mathrm{ctl}}^2(H) + m_{e,4}\right], \quad m_{e,4} := \sigma_0^2\sqrt{d_u(d_u+2)}, \\
&p = 2, \qquad C_s^{\mathrm{id}} = 2\bar{s}/H.
\end{aligned}$$

Here $C_H = O(\log H)$ and $\delta_s \le \alpha' + \delta/(16H^2) + \delta/4$.

Step 8. We now combine all of the analysis above to derive the regret bound of $\pi^{\mathrm{A_{LQR}}}$ Noting that

$$\sum_{s \notin \mathcal{T}_H^\circ} \ell_{s,H} \le 2t_0 + 2t_1 + 4M, \qquad R_x(H,\alpha')^2 + B_{\mathrm{ctl}}(H)^2 = \widetilde{O}(\log H),$$

we have

$$\begin{aligned}
&\frac{1}{H}\sum_{s \in \mathcal{T}_H} \ell_{s,H}\beta_s^{\mathrm{Bell}}(\alpha') \\
\le &\frac{2\bar{s}}{H}\Bigg[L_K^2 R_x^0(H,\alpha')^2 \sum_{s \in \mathcal{T}_H^\circ} \ell_{s,H}\epsilon_s^2 + R_x(H,\alpha')^2 \sum_{s \in \mathcal{T}_H} \ell_{s,H} c_s^2 \\
&+ \Big\{(\|\widehat{\mathbf{K}}\|_{\mathrm{op}} + K_{\max})R_x(H,\alpha') + B_{\mathrm{ctl}}(H)\Big\}^2 (2t_0 + 2t_1 + 4M)\Bigg] \\
= &\widetilde{O}(H^{-1/2}).
\end{aligned}$$

Moreover, using $C_s^{\mathrm{id}} = 2\bar{s}/H$ and $\sum_{t<H} \sigma_{\mathrm{id},t}^2 = O(\sqrt{H})$, it holds that

$$\sum_{s \in \mathcal{T}_H} C_s^{\mathrm{id}} \sum_{t=s}^{s+\ell_{s,H}-1} d_u \sigma_{\mathrm{id},t}^2 = \widetilde{O}(H^{-1/2}).$$

The initialization and bad-event terms satisfy

$$B_0^{\mathrm{init}} + C_s^{\mathrm{bad}} \sum_{s \in \mathcal{T}_H} \ell_{s,H}\sqrt{\delta_s} \le \frac{(2M \wedge H)C_H}{H} + \frac{C_H}{H}\sum_{s \in \mathcal{T}_H} \ell_{s,H}\sqrt{\delta_s} = \widetilde{O}(H^{-1}).$$

Substituting the above bounds into Theorem 3 yields

$$\mathcal{J}^H(\pi^{\mathrm{A_{LQR}}}) - \mathcal{J}^{H,*} = \widetilde{O}(H^{-1/2}).$$

□

D.2 Nonlinear dynamics with a Fourier sieve

In this subsection, we analyze the regret bound of Algorithm 1 in the nonlinear Fourier control problem. Let $A_{\mathrm{F}}, \nu_{\mathrm{F}} > 0$ be known constants and $\{\alpha_{\ell,\star}\}_{\ell \ge 1}$ be unknown coefficients satisfying $|\alpha_{\ell,\star}| \le A_{\mathrm{F}} e^{-\nu_{\mathrm{F}}\ell}, \ell \ge 1$. Set constants $d_x = d_u = 1, a > 1, q, \mu_{\mathrm{F}} > 0, k_{\mathrm{ref}} \in \mathbb{R}, b_-, b_+ \in \mathbb{R}, b_- \le b_+$. For an unknown function $g_\star(x) := \sum_{\ell \ge 1} \alpha_{\ell,\star}\sin(\ell x)$, an unknown $b_\star \in [b_-, b_+]$ and a known constant a, consider a nonlinear dynamic model:

$$\begin{aligned}
&x_{t+1} = ax_t + g_\star(x_t) + b_\star u_t + w_t, \\
&c(x_t,u_t) = q(1-\cos x_t) + \mu_{\mathrm{F}}\{\sqrt{1 + (u_t - k_{\mathrm{ref}}x_t)^2} - 1\}, \\
&h(x_H) = 0, \quad t = 0, \dots, H-1.
\end{aligned} \tag{76}$$

In this example, we consider a fixed truncation order $d_H \geq 2$ and set $d_t = d_H$, $f^{\mathrm{pub}}(x,u) = ax$, $\theta_t^* = \theta^* := (\alpha_{1,\star}, \ldots, \alpha_{d_H-1,\star}, b_\star)$, and $\phi^{(t)}(x,u) = \mathrm{col}(\sin x, \ldots, \sin((d_H-1)x), u)$ for any $t = 0, \ldots, H-1$. Meanwhile, we set $\Theta_t = \Theta$ for any $t = 0, \ldots, H-1$, where Θ is the compact and convex set

$$\Theta = \{(\alpha, b) \in \mathbb{R}^{d_H-1} \times \mathbb{R} \mid |\alpha_\ell| \leq A_{\mathrm{F}} e^{-\nu_{\mathrm{F}}\ell},\ \ell = 1, \ldots, d_H - 1, \quad b_- \leq b \leq b_+\}.$$

Furthermore, we set the parameters in Algorithm 1 as $M = 3, \eta_t = 1, \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \sigma_{\mathrm{id},t}^2 = \sigma_0^2(t+1)^{-1/2}, t = 0, \ldots, H-1$, with $\sigma_0 > 0$. The realization of Algorithm 1 follows the discussion of Appendix B.

We introduce notation and an assumption to analyze the regret bound for this example. Fix

$$H \geq 2M+1, \qquad \delta = (H+1)^{-2}, \qquad \alpha' = \frac{\delta}{4H}.$$

For $\alpha \in (0,1)$, set

$$\begin{aligned} &R_x(H,\alpha) = R_{x,H-1}(\alpha/2), \qquad r_e(H,\alpha) = \sigma_0\sqrt{2\log(4H/\alpha)}, \\ &\mathcal{K}_s(\alpha) = \{(x,u) \mid |x| \leq R_x(H,\alpha),\ |u| \leq |\widehat{\mathbf{K}}|R_x(H,\alpha) + B_{\mathrm{ctl}} + r_e(H,\alpha)\}, \quad s \in \mathcal{T}_H. \end{aligned} \tag{77}$$

For $f(x,u) = ax + g_\star(x) + b_\star u$, we define

$$\begin{aligned} &\mathcal{X}_s(\alpha) := \{x : \exists u,\ (x,u) \in \mathcal{K}_s(\alpha)\}, \quad u_t^*(x) \in \operatorname*{arg\,min}_{u\in\mathbb{R}} \mathsf{Q}_{t,H}(x,u), \\ &\mathsf{Q}_{t,H}(x,u) := c(x,u) + H\mathbb{E}_\tau[V_{t+1}^*(f(x,u) + w_t)], \quad \forall s \in \mathcal{T}_H, \quad t = 0, \ldots, H-1. \end{aligned} \tag{78}$$

For the sampled model $(f_{\theta_s}, \widehat{\Sigma}_{w,s})$, we define

$$\begin{aligned} &\widehat{V}_{s,H}^*(x) := h(x), \\ &\widehat{\mathsf{Q}}_{s,t,H}(x,u) := c(x,u) + H\mathbb{E}_{\tau,\widehat{w}\sim\mathcal{N}(0,\widehat{\Sigma}_{w,s})}\widehat{V}_{s,t+1}^*(f_{\theta_s}(x,u) + \widehat{w}), \\ &\widehat{V}_{s,t}^*(x) := H^{-1}\inf_{u\in\mathbb{R}} \widehat{\mathsf{Q}}_{s,t,H}(x,u), \quad \forall s \in \mathcal{T}_H, \quad t = 0, \ldots, H-1. \end{aligned} \tag{79}$$

Define the bounded action interval

$$\mathcal{U}_{\mathrm{B}}(x) := [\widehat{\mathbf{K}}x - B_{\mathrm{ctl}}(H),\ \widehat{\mathbf{K}}x + B_{\mathrm{ctl}}(H)]$$

and control input $\bar{u}_{s,t}(x) := \widehat{\mathbf{K}}x + \mathrm{Proj}_{\mathbb{B}^{d_u}(0,B_{\mathrm{ctl}}(H))}\big(\mu_{\theta_s}^{(t)}(x)\big)$ for $s \in \mathcal{T}_H$ and $t = 0, \ldots, H-1$. It is obvious to see that $\bar{u}_{s,t}(x) \in \mathcal{U}_{\mathrm{B}}(x)$.

Assumption 6. After certification, the continuation is restarted; x_0, the fresh process noises, and the exploration noises are mutually independent and jointly independent of $\mathscr{G}_\tau$. They satisfy

$$x_0 \sim \nu_0 = \mathcal{N}(0,1), \quad w_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,\sigma_w^2), \qquad 0 < \lambda_w \leq \sigma_w^2 \leq \bar{\lambda}_w, \qquad i = 0, \ldots, H-1.$$

Let

$$G_{\mathrm{F}} := \frac{A_{\mathrm{F}}e^{-\nu_{\mathrm{F}}}}{(1-e^{-\nu_{\mathrm{F}}})^2}, \quad \rho_{\mathrm{F}} := \sup_{b\in[b_-,b_+]} |a + b\widehat{\mathbf{K}}| + G_{\mathrm{F}}, \quad b_{\max} := \max_{b\in[b_-,b_+]} |b| \tag{80}$$

satisfy $\rho_{\mathrm{F}} < 1$. Define

$$\ell_{\mathrm{F}}^- := \inf_{x\neq 0} \frac{g_\star(x)}{x}, \qquad \ell_{\mathrm{F}}^+ := \sup_{x\neq 0} \frac{g_\star(x)}{x}.$$

There exists $\bar{q}_{\mathrm{F}} \in (0,1)$ such that

$$b_\star \neq 0, \qquad \frac{\ell_{\mathrm{F}}^+ - \ell_{\mathrm{F}}^-}{2} \leq \bar{q}_{\mathrm{F}}, \qquad \left|a + \frac{\ell_{\mathrm{F}}^- + \ell_{\mathrm{F}}^+}{2}\right| \leq |b_\star|\mathscr{K}.$$

Set $L_W := (q + \mu_{\mathrm{F}}|\widehat{\mathbf{K}} - k_{\mathrm{ref}}|)/(1-\rho_{\mathrm{F}})$ and require

$$B_{\mathrm{ctl}}(H) \geq |\widehat{\mathbf{K}} - k_{\mathrm{ref}}|R_x(H,\alpha').$$

Let $0 < \mu_H \leq L_H$ and $L_{\theta,H}, L_{\Sigma,H}, L_{\chi,H} \geq 0$ satisfy

$$\max\left\{L_H, \mu_H^{-1}, L_{\theta,H}, L_{\Sigma,H}, L_{\chi,H}, B_{\mathrm{ctl}}(H), |\widehat{\mathbf{K}}|, (1-\rho_{\mathrm{F}})^{-1}\right\} = \widetilde{O}(1).$$

For every $s \in \mathcal{T}_H$, let $\varepsilon_s^{\mathrm{Bell}} \geq 0$ be a deterministic error. On $\mathcal{G}_s(a_s, c_s, \alpha)$, it holds that

$$\sup_{\substack{s \leq t < (s+2M)\wedge H \\ x \in \mathcal{X}_s(\alpha)}} \left[\widehat{\mathsf{Q}}_{s,t,H}(x, \bar{u}_{s,t}(x)) - \inf_{u \in \mathbb{R}} \widehat{\mathsf{Q}}_{s,t,H}(x,u) \right] \leq \varepsilon_s^{\mathrm{Bell}}. \tag{81}$$

Moreover, for a fixed constant $C_{\mathrm{sol}} > 0$ independent of H, it holds that

$$\sum_{s \in \mathcal{T}_H} \varepsilon_s^{\mathrm{Bell}} \leq C_{\mathrm{sol}} \sqrt{H}. \tag{82}$$

For every $s \leq t < (s+2M) \wedge H$ and $x \in \mathcal{X}_s(\alpha)$, the function $\mathsf{Q}_{t,H}(x, \cdot)$ is twice continuously differentiable and satisfies

$$\begin{aligned}
&u_t^\star(x) \in \mathcal{U}_{\mathrm{B}}(x), \\
&\mathsf{Q}_{t,H}(x,u) - \mathsf{Q}_{t,H}(x, u_t^\star(x)) \geq \frac{\mu_H}{2} |u - u_t^\star(x)|^2, \qquad u \in \mathcal{U}_{\mathrm{B}}(x), \\
&\nabla_{uu}^2 \mathsf{Q}_{t,H}(x,u) \leq L_H, \qquad u \in \mathbb{R}, \\
&\left\| \nabla_u \mathsf{Q}_{t,H}(x,u) - \nabla_u \widehat{\mathsf{Q}}_{s,t,H}(x,u) \right\|_2 \leq b_{s,H}^{\mathrm{grad}}(\alpha), \qquad u \in \mathcal{U}_{\mathrm{B}}(x)
\end{aligned} \tag{83}$$

with $b_{s,H}^{\mathrm{grad}}(\alpha) := L_{\theta,H} a_s + L_{\Sigma,H}(c_s) + L_{\chi,H} \chi_s(\alpha)$. Here $u_t^\star(x)$ is a minimizer of $\mathsf{Q}_{t,H}(x, \cdot)$ over $\mathbb{R}$.

Set $\Delta_H = \frac{A_{\mathrm{F}} e^{-\nu_{\mathrm{F}} d_H}}{1 - e^{-\nu_{\mathrm{F}}}}$, $d_H = \left\lceil \frac{\log H}{2\nu_{\mathrm{F}}} \right\rceil + 1$. Denoting the above realization of Algorithm 1 by A_{F} and the corresponding policy by $\pi^{\mathrm{A}_{\mathrm{F}}}$, we propose the following theoretical result.

Proposition 4. *Under Assumption 6 and the above settings, Assumptions 1–3 hold. Meanwhile, it holds that*

$$\mathcal{J}^H(\pi^{\mathrm{A}_{\mathrm{F}}}) - \mathcal{J}^{H,*} \leq \widetilde{O}\Big(H^{-1/2}\Big). \tag{84}$$

Proof. **Step 1.** We first verify Assumption 1. Set

$$\mathscr{L} = b_{\max}, \quad \underline{v} = \overline{v} = 1, \quad q_1 = \frac{1 + \bar{q}_{\mathrm{F}}}{2}, \quad q_2 = \frac{1 - \bar{q}_{\mathrm{F}}}{2}.$$

Then

$$f(0,0) = 0, \qquad |f(x,u) - f(x,u')| = |b_\star| \, |u - u'| \leq \mathscr{L} |u - u'|,$$

which verifies Assumption 1(a). Taking

$$K_\star = -\frac{a + (\ell_{\mathrm{F}}^- + \ell_{\mathrm{F}}^+)/2}{b_\star}, \qquad P_\star = 1,$$

we have $(K_\star, P_\star) \in \mathfrak{S}$ and

$$\begin{aligned}
\|f(x, K_\star x)\|_{P_\star} &= \left| g_\star(x) - \frac{\ell_{\mathrm{F}}^- + \ell_{\mathrm{F}}^+}{2} x \right| \\
&\leq \frac{\ell_{\mathrm{F}}^+ - \ell_{\mathrm{F}}^-}{2} |x| \leq (q_1 - q_2) \|x\|_{P_\star}, \qquad x \in \mathbb{R}.
\end{aligned}$$

Thus Assumption 1(b) holds.

Step 2. We now analyze the approximation error. For the basis functions $\phi^{(t)}, t = 0, \ldots, H-1$, the Fourier coefficient bound gives

$$\begin{aligned}
&\sup_{(x,u) \in \mathbb{R}^2} |f(x,u) - f_{\theta^*}(x,u)| \leq \sum_{\ell = d_H}^{\infty} A_{\mathrm{F}} e^{-\nu_{\mathrm{F}} \ell} = \Delta_H, \\
&\Delta_t := \Delta_H, \qquad \alpha_{\mathrm{app},t} = 0, \qquad \chi_s(\alpha) \leq \Delta_H.
\end{aligned} \tag{85}$$

Step 3. We next verify Assumption 2. Firstly, we verify (6) instead of the two moment conditions in Assumption 2. Let μ_i^{dep} denote the residual policy actually deployed at time i. For $s \in \mathcal{T}_H$, define

$$\phi_i^{(s)} := \mathrm{col}\big(\sin(x_i), \ldots, \sin((d_H - 1)x_i), u_i\big).$$

For $s \in \mathcal{T}_H$ and $0 \le i < s$, let

$$m_i := \begin{cases} 0, & i = 0, \\ f(x_{i-1}, u_{i-1}), & i \ge 1, \end{cases} \qquad x_i = m_i + y_i.$$

For $i = 0$, $y_i = x_0 \sim \mathcal{N}(0,1)$ is independent of $\mathscr{H}_{-1}$; for $i \ge 1$, $y_i = w_{i-1} \sim \mathcal{N}(0, \Sigma_w)$. Consequently, conditionally on $\mathscr{H}_{i-1}$, the density of y_i is at least $\frac{\exp\left[-\pi^2/\left(2\min\{1,\lambda_w\}\right)\right]}{\sqrt{2\pi \max\{1,\bar{\lambda}_w\}}}$ on $[-\pi, \pi]$. Define

$$D_{\mathrm{F}} := |\widehat{\mathbf{K}}|_{\mathrm{op}}\pi + B_{\mathrm{ctl}}(H), \qquad \beta_0 := \min\left\{\frac{1}{2}, \frac{1}{16\sqrt{2}(1+D_{\mathrm{F}})}\right\},$$

$$c_{\mathrm{F}} := \min\left\{\frac{1}{32\sigma_0}, \frac{\beta_0\sqrt{2\pi}}{8}\right\}, \qquad \gamma_{\mathrm{F},s} := c_{\mathrm{F}}^2\sigma_0^2(s+1)^{-1/2}.$$

Let $v = (r^\top, \beta)^\top \in \mathbb{S}^{d_H-1}$, $r = (r_1, \dots, r_{d_H-1})^\top$. Suppose first that $|\beta| \le \beta_0$ and define

$$\mathcal{E}_i^e := \left\{|e_i^{\mathrm{id}}| \le (32\beta_0)^{-1}\right\}.$$

With $\Phi(z) := (2\pi)^{-1/2}\int_{-\infty}^{z} e^{-v^2/2}\,\mathrm{d}v$, we have

$$\sigma_{\mathrm{id},i} \le \sigma_0, \qquad \mathbb{P}_\tau(\mathcal{E}_i^e \mid \mathscr{H}_{i-1}) \ge 2\Phi\left(\frac{1}{32\beta_0\sigma_0}\right) - 1.$$

For every fixed $e \in [-(32\beta_0)^{-1}, (32\beta_0)^{-1}]$, define $c(e) := \beta(\widehat{\mathbf{K}}m_i + e)$. On $|y_i| \le \pi$, we have

$$\left|\beta\big(\widehat{\mathbf{K}}y_i + \mu_i^{\mathrm{dep}}(m_i + y_i)\big)\right| \le \beta_0 D_{\mathrm{F}}, \qquad \left|\sum_{\ell=1}^{d_H-1} r_\ell \sin(\ell(m_i + y_i))\right| \le \sqrt{d_H - 1}.$$

If $|c(e)| > 2\big(\sqrt{d_H-1} + \beta_0 D_{\mathrm{F}}\big)$, then, for every $|y_i| \le \pi$, it holds that $|v^\top \phi_i^{(s)}| \ge \sqrt{\gamma_{\mathrm{F},s}}$. Hence, conditionally on $\mathscr{H}_{i-1}$ and $e_i^{\mathrm{id}} = e$,

$$\mathbb{P}_\tau\left(|v^\top\phi_i^{(s)}| \ge \sqrt{\gamma_{\mathrm{F},s}} \,\middle|\, \mathscr{H}_{i-1}, e_i^{\mathrm{id}} = e\right) \ge \frac{2\pi\exp\left[-\pi^2/\left(2\min\{1,\lambda_w\}\right)\right]}{\sqrt{2\pi\max\{1,\bar{\lambda}_w\}}}.$$

If $|c(e)| \le 2\big(\sqrt{d_H-1} + \beta_0 D_{\mathrm{F}}\big)$, then Fourier orthogonality on $[m_i - \pi, m_i + \pi]$ gives

$$\begin{aligned}
&\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{\ell=1}^{d_H-1} r_\ell \sin(\ell(m_i+y)) + c(e)\right|^2 \mathrm{d}y = \frac{\|r\|_2^2}{2} + c(e)^2 \ge \frac{3}{8}, \\
&\sup_{|y|\le\pi}\left|\sum_{\ell=1}^{d_H-1} r_\ell \sin(\ell(m_i+y)) + c(e)\right| \le 3\big(1 + \sqrt{d_H-1} + \beta_0 D_{\mathrm{F}}\big), \\
&\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\sum_{\ell=1}^{d_H-1} r_\ell \sin(\ell(m_i+y)) + c(e)\right|^4 \mathrm{d}y \\
&\le 9\big(1 + \sqrt{d_H-1} + \beta_0 D_{\mathrm{F}}\big)^2\left(\frac{\|r\|_2^2}{2} + c(e)^2\right).
\end{aligned}$$

For every bounded residual selector, on $|y_i| \le \pi$,

$$\left|\sum_{\ell=1}^{d_H-1} r_\ell \sin(\ell(m_i+y_i)) + c(e)\right| \ge \sqrt{\frac{3}{16}} \quad\Longrightarrow\quad |v^\top\phi_i^{(s)}| \ge \sqrt{\frac{3}{16}} - \beta_0 D_{\mathrm{F}} > \frac{1}{32} \ge \sqrt{\gamma_{\mathrm{F},s}}.$$

Therefore, the Paley–Zygmund inequality gives, uniformly for $e \in [-(32\beta_0)^{-1}, (32\beta_0)^{-1}]$,

$$\begin{aligned}
&\mathbb{P}_\tau\left(|v^\top\phi_i^{(s)}| \ge \sqrt{\gamma_{\mathrm{F},s}} \,\middle|\, \mathscr{H}_{i-1}, e_i^{\mathrm{id}} = e\right) \\
&\ge \frac{2\pi\exp\left[-\pi^2/\left(2\min\{1,\lambda_w\}\right)\right]}{\sqrt{2\pi\max\{1,\bar{\lambda}_w\}}}\min\left\{1, \frac{7}{8192\big(1+\sqrt{d_H-1}+\beta_0 D_{\mathrm{F}}\big)^2}\right\}.
\end{aligned}$$

Combining the two cases and integrating with respect to e_i^{id} gives

$$\mathbb{P}_\tau\Big(|v^\top \phi_i^{(s)}| \geq \sqrt{\gamma_{\mathrm{F},s}} \,\Big|\, \mathscr{H}_{i-1}\Big) \geq p_{\mathrm{F},1}$$

with

$$p_{\mathrm{F},1} := \frac{2\pi \exp\big[-\pi^2/\big(2\min\{1,\lambda_w\}\big)\big]}{\sqrt{2\pi \max\{1,\bar{\lambda}_w\}}} \min\left\{1,\ \frac{7}{8192\big(1+\sqrt{d_H-1}+\beta_0 D_{\mathrm{F}}\big)^2}\right\} \cdot \left[2\Phi\left(\frac{1}{32\beta_0\sigma_0}\right)-1\right].$$

Now suppose that $|\beta| > \beta_0$. Conditionally on $\mathscr{H}_{i-1}$, x_i, and the residual action $\mu_i^{\mathrm{dep}}(x_i)$ chosen before drawing e_i^{id}, the map

$$e \longmapsto r^\top\big(\sin(x_i),\dots,\sin((d_H-1)x_i)\big) + \beta\big(\widehat{\mathbf{K}}x_i + \mu_i^{\mathrm{dep}}(x_i) + e\big)$$

is affine with β. Since $e_i^{\mathrm{id}} \sim \mathcal{N}(0,\sigma_{\mathrm{id},i}^2)$, we have

$$\mathbb{P}_\tau\Big(|v^\top \phi_i^{(s)}| < \sqrt{\gamma_{\mathrm{F},s}} \,\Big|\, \mathscr{H}_{i-1}, x_i, \mu_i^{\mathrm{dep}}(x_i)\Big) \leq \frac{2\sqrt{\gamma_{\mathrm{F},s}}}{\sqrt{2\pi}|\beta|\sigma_{\mathrm{id},i}} \leq \frac{2c_{\mathrm{F}}}{\sqrt{2\pi}|\beta|} \leq \frac{1}{4}.$$

For $i < s$, it holds that $\sigma_{\mathrm{id},i}^2 = \sigma_0^2(i+1)^{-1/2} \geq \sigma_0^2(s+1)^{-1/2}$. Then, for every $v \in \mathbb{S}^{d_H-1}$ and every $0 \leq i < s$, we have

$$\mathbb{P}_\tau\Big(|v^\top \phi_i^{(s)}| \geq \sqrt{\gamma_{\mathrm{F},s}} \,\Big|\, \mathscr{H}_{i-1}\Big) \geq p_{\mathrm{F}}, \qquad p_{\mathrm{F}} := \min\left\{p_{\mathrm{F},1}, \frac{3}{4}\right\}.$$

Thus (6) holds with

$$k_s = 1, \qquad \Gamma_{\mathrm{PE},s} = 2\gamma_{\mathrm{F},s} I_{d_H}, \qquad \kappa_{\mathrm{PE},s} = \frac{1}{4p_{\mathrm{F}}}.$$

Secondly, we verify the upper bound of $\mathbb{E}_\tau\Big[\sum_{i=0}^{n_s-1} \phi^{(s)}(x_i,u_i)\phi^{(s)}(x_i,u_i)^\top \,\Big|\, \mathscr{H}_{-1}\Big]$. Define

$$C_{x,H} := 1 + \frac{4(1+\rho_{\mathrm{F}}^2)\{b_{\max}^2[B_{\mathrm{ctl}}(H)^2+\sigma_0^2]+\bar{\lambda}_w\}}{(1-\rho_{\mathrm{F}}^2)^2}, \qquad C_u := 3\big(|\widehat{\mathbf{K}}|^2 C_{x,H} + B_{\mathrm{ctl}}(H)^2 + \sigma_0^2\big),$$

Using

$$|x_{i+1}| \leq \rho_{\mathrm{F}}|x_i| + b_{\max}\big(B_{\mathrm{ctl}}(H) + |e_i^{\mathrm{id}}|\big) + |w_i|$$

and the Gaussian second moments, we have

$$\mathbb{E}_\tau[x_i^2 \mid \mathscr{H}_{-1}] \leq C_{x,H}, \qquad \mathbb{E}_\tau[u_i^2 \mid \mathscr{H}_{-1}] \leq C_u, \qquad 0 \leq i < H.$$

Since $\|\phi^{(t)}(x_i,u_i)\|_2^2 = \sum_{\ell=1}^{d_H-1} \sin^2(\ell x_i) + u_i^2$, it holds that

$$\mathbb{E}_\tau\Big[\|\phi^{(t)}(x_i,u_i)\|_2^2 \mid \mathscr{H}_{-1}\Big] \leq d_H - 1 + C_u \leq (1+C_u)d_H.$$

Setting $\Gamma_{\mathrm{up},s} = (1+C_u)d_H I_{d_H}$, we obtain

$$\mathbb{E}_\tau\left[\sum_{i=0}^{n_s-1} \phi^{(s)}(x_i,u_i)\phi^{(s)}(x_i,u_i)^\top \,\middle|\, \mathscr{H}_{-1}\right] \preceq n_s \Gamma_{\mathrm{up},s}.$$

This verifies Assumption 2.

Step 4. We next derive the estimation error bound of θ_s via Theorem 1. Set $\lambda_s = 1$ and choose $t_0 \geq 2(M+1)$ so that (7) and $g_{-,s} \geq 1$ hold for every $s \in \mathcal{T}_H$ with $s \geq t_0$. For $s \in \mathcal{T}_H$, set

$$\alpha_{\mathrm{up},s} = \delta_{\mathrm{sb},s} = \delta_{\mathrm{mg},s} = \frac{\delta}{32H^3}, \quad g_{-,s} = \frac{p_{\mathrm{F}}^2}{16} n_s \gamma_{\mathrm{F},s}, \quad \epsilon_s^2 = \begin{cases} \big[1+\mathrm{diam}_{\mathrm{F}}(\Theta)\big]^2, & s < t_0, \\ \dfrac{4z_s^2 + 2\log(32H^3\mathcal{M}_s/\delta)}{g_{-,s}}, & s \geq t_0. \end{cases}$$

Hence, using Theorem 1 when $s \geq t_0$ and the deterministic diameter bound $(\mathrm{diam}_{\mathrm{F}}(\Theta))^2$ when $s < t_0$, we have

$$\mathbb{P}_\tau(\|\theta_s - \theta^*\|_{\mathrm{F}} > \epsilon_s) \leq q_s = \frac{\delta}{8H^3}\mathbf{1}\{s \geq t_0\}. \tag{86}$$

Indeed, it holds that

$$\begin{aligned}
&p_{\mathrm{F}}^{-1} = O(d_H), \qquad c_{\mathrm{F}}^{-2} = O(1 + |\widehat{\mathbf{K}}|^2 + B_{\mathrm{ctl}}(H)^2), \qquad C_u = \widetilde{O}(1),\\
&\log\det(2\overline{\Gamma}_s\Gamma_{\mathrm{PE},s}^{-1}) = d_H \log\left(1 + \frac{d_H^2(1+C_u)}{\alpha_{\mathrm{up},s}\gamma_{\mathrm{F},s}}\right),\\
&g_{-,s}^{-1} = \widetilde{O}(s^{-1/2}), \qquad z_s^2 + \log\mathcal{M}_s = \widetilde{O}(s\Delta_H^2 + d_H), \qquad s \geq t_0.
\end{aligned}$$

Thus $t_0 = \widetilde{O}(1)$ can be chosen, and

$$\epsilon_s^2 \leq \left[1 + \mathrm{diam}_{\mathrm{F}}(\Theta)\right]^2 \mathbf{1}\{s < t_0\} + \widetilde{O}\left(\sqrt{s}\,\Delta_H^2 + \frac{d_H}{\sqrt{s}}\right)\mathbf{1}\{s \geq t_0\}.$$

Step 5. We now derive the error bound of $\widehat{\Sigma}_w$ via Theorem 2. For $s \in \mathcal{T}_H$, choose

$$\begin{aligned}
&\alpha_s = \frac{\delta}{32H^3}, \qquad \zeta_s = s^{-1/2} + \frac{2M}{sH^2}\sum_{\substack{j\in\mathcal{T}_H\\ t_0\leq j\leq s}} U_j(\alpha_s)^2,\\
&\epsilon_s^{\mathrm{cov}} = \sqrt{\frac{64\bar{\lambda}_w[\beta_s(\alpha_s)+\zeta_s]\log(128H^3d_H/\delta)}{s}} \vee \bar{\lambda}_w\sqrt{\frac{8(1+d_x\log 9)\log(128H^3d_H/\delta)}{c_{\mathrm{HW}}s}}\\
&\qquad\vee \frac{4\bar{\lambda}_w\log(128H^3d_H/\delta)}{c_{\mathrm{HW}}s}.
\end{aligned}$$

Using $q_j = 0$ for $j < t_0$ in Theorem 2, we have

$$\Lambda_s(\alpha_s,\zeta_s) \leq \frac{2M}{s\zeta_s}\sum_{\substack{j\in\mathcal{T}_H\\ t_0\leq j\leq s}} U_j(\alpha_s)^2\frac{\delta}{8H^3} \leq \frac{\delta}{8H}.$$

We therefore have

$$\mathbb{P}_\tau\left(\|\widehat{\Sigma}_{w,s} - \Sigma_w\|_{\mathrm{op}} > \beta_s(\alpha_s) + \zeta_s + 2\epsilon_s^{\mathrm{cov}}\right) \leq \frac{\delta}{4H}.$$

Step 6. We now verify Assumption 3. Set

$$a_s = \epsilon_s, \qquad c_s = \beta_s(\alpha_s)+\zeta_s+2\epsilon_s^{\mathrm{cov}}, \qquad \delta_s := \mathbb{P}_\tau\big(\mathcal{G}_s(a_s,c_s,\alpha')^c\big) \leq \alpha' + \frac{\delta}{8H^3}\mathbf{1}\{s\geq t_0\} + \frac{\delta}{4H}.$$

For $x, x', r \in \mathbb{R}$, we have

$$\begin{aligned}
&|f(x,\widehat{\mathbf{K}}x + r) - f(x',\widehat{\mathbf{K}}x' + r)| \leq \rho_{\mathrm{F}}|x - x'|,\\
&|c(x,\widehat{\mathbf{K}}x + r) - c(x',\widehat{\mathbf{K}}x' + r)| \leq (q + \mu_{\mathrm{F}}|\widehat{\mathbf{K}} - k_{\mathrm{ref}}|)|x - x'|.
\end{aligned}$$

Backward induction in the Bellman recursion, using $h = 0$ and $c \geq 0$, gives

$$\begin{aligned}
&H|V_t^*(x) - V_t^*(x')| \leq L_W|x - x'|, \qquad V_t^*(x) \geq V_{t+1}^*(x),\\
&0 \leq H\mathcal{D}_t^H(x,u) \leq c(x,u) + L_W\mathbb{E}_\tau|f(x,u) + w_t - x|, \qquad 0 \leq t < H.
\end{aligned}$$

On $\mathcal{G}_s(a_s,c_s,\alpha')$, combining (81) and (83) gives, for $s \leq t < (s+2M)\wedge H$,

$$\frac{\mu_H}{2}|\bar{u}_{s,t}(x_t) - u_t^\star(x_t)|^2 \leq H\mathcal{D}_t^H(x_t,\bar{u}_{s,t}(x_t)) \leq 2\varepsilon_s^{\mathrm{Bell}} + \frac{2(b_{s,H}^{\mathrm{grad}}(\alpha'))^2}{\mu_H}.$$

Since $u_t = \bar{u}_{s,t}(x_t) + e_t^{\mathrm{id}}$ and $\nabla_{uu}^2\mathsf{Q}_{t,H} \leq L_H$,

$$\mathcal{D}_t^H(x_t,u_t) \leq \frac{4L_H\varepsilon_s^{\mathrm{Bell}}}{\mu_H H} + \frac{4L_H(b_{s,H}^{\mathrm{grad}}(\alpha'))^2}{\mu_H^2 H} + \frac{L_H}{H}|e_t^{\mathrm{id}}|^2.$$

Define

$$C_{\mathrm{F},H} := 2q + \mu_{\mathrm{F}}(B_{\mathrm{ctl}}(H) + \sigma_0) + \big(\mu_{\mathrm{F}}|\widehat{\mathbf{K}} - k_{\mathrm{ref}}| + 2L_W\big)\sqrt{C_{x,H}}.$$

By Jensen's and Minkowski's inequalities, it holds that

$$\begin{aligned}\|H\mathcal{D}_t^H(x_t,u_t)\|_{L^2(\mathbb{P}_\tau)} &\le \|c(x_t,u_t)\|_{L^2(\mathbb{P}_\tau)} + L_W\|x_{t+1}-x_t\|_{L^2(\mathbb{P}_\tau)}\\ &\le C_{\mathrm{F},H} = \widetilde{O}(1), \qquad 0\le t<H.\end{aligned}$$

Therefore Assumption 3 holds at $\alpha=\alpha'$ with

$$\beta_s^{\mathrm{Bell}}(\alpha') = \frac{4L_H\varepsilon_s^{\mathrm{Bell}}}{\mu_H H} + \frac{4L_H\big(b_{s,H}^{\mathrm{grad}}(\alpha')\big)^2}{\mu_H^2 H},$$

$$C_s^{\mathrm{id}} = L_H/H, \qquad p=2, \qquad C_s^{\mathrm{bad}} \le \frac{C_{\mathrm{F},H}}{H}, \qquad B_0^{\mathrm{init}} \le \frac{(2M\wedge H)C_{\mathrm{F},H}}{H}.$$

Step 7. We combine the above bounds to derive the regret bound of $\pi^{\mathrm{A_F}}$. By the definitions of L_s, U_s, β_s in Theorem 2 and Steps 2–5, it holds that

$$\begin{aligned}&L_s(\alpha_s) \le \Delta_H + \bar{d}_s(\alpha_s)\epsilon_s, \qquad \bar{d}_s(\alpha_s) + U_s(\alpha_s) = \widetilde{O}(1),\\ &\beta_s(\alpha_s) = \widetilde{O}\Big(\sqrt{s}\,\Delta_H^2 + s^{-1/2} + t_0/s\Big), \qquad \zeta_s = s^{-1/2} + \widetilde{O}(H^{-2}),\\ &c_s = \widetilde{O}\Big(\sqrt{s}\,\Delta_H^2 + s^{-1/2} + t_0/s\Big), \qquad \chi_s(\alpha') \le \Delta_H.\end{aligned}$$

Since $t_0 = \widetilde{O}(1)$ and $\Delta_H^2 = \widetilde{O}(H^{-1})$, we have

$$\begin{aligned}&\frac{1}{H}\sum_{s\in\mathcal{T}_H}\ell_{s,H}a_s^2 = \widetilde{O}\bigg(\sqrt{H}\,\Delta_H^2 + \frac{d_H}{\sqrt{H}} + \frac{t_0}{H}\bigg) = \widetilde{O}(H^{-1/2}),\\ &\frac{1}{H}\sum_{s\in\mathcal{T}_H}\ell_{s,H}c_s^2 = \widetilde{O}\bigg(H\Delta_H^4 + \frac{d_H}{H} + \frac{t_0^2}{H}\bigg) = \widetilde{O}(H^{-1}),\\ &\sum_{s\in\mathcal{T}_H}\ell_{s,H}\beta_s^{\mathrm{Bell}}(\alpha') = \widetilde{O}\bigg(H^{-1/2} + \frac{t_0}{H}\bigg),\\ &\sum_{s\in\mathcal{T}_H}\ell_{s,H}C_s^{\mathrm{bad}}\delta_s^{1/2} = \widetilde{O}\bigg(\sqrt{\alpha' + \frac{\delta}{H}}\bigg).\end{aligned}$$

Moreover,

$$\sum_{s\in\mathcal{T}_H}\ell_{s,H}\varepsilon_s^{\mathrm{Bell}} \le 2MC_{\mathrm{sol}}\sqrt{H},$$

$$B_0^{\mathrm{init}} + \sum_{s\in\mathcal{T}_H}C_s^{\mathrm{id}}\sum_{t=s}^{s+\ell_{s,H}-1}\sigma_{\mathrm{id},t}^2 \le \frac{(2M\wedge H)C_{\mathrm{F},H}}{H} + \frac{2L_H\sigma_0^2}{\sqrt{H}} = \widetilde{O}(H^{-1/2}).$$

Theorem 3 gives

$$\mathcal{J}^H(\pi^{\mathrm{A_F}}) - \mathcal{J}^{H,*} = \widetilde{O}(H^{-1/2}).$$

□

D.3 Nonlinear transient price impact

In this subsection, we analyze the regret bound of Algorithm 1 in a nonlinear transient price-impact problem (see Curato et al. (2017); Alfonsi et al. (2010); Obizhaeva & Wang (2013) for more details). We now introduce the model settings of this problem. Let y_t, z_t, u_t, w_t^y and w_t^z denote inventory, transient impact, control, inventory noise and impact noise, respectively, and set $x_t = (y_t, z_t)^\top$, $d_x = 2$, $d_u = 1$. For known parameters $0 < \nu < 1$, $\epsilon_{\mathrm{imp}}, \eta_{\mathrm{imp}} > 0$, $\phi_{\mathrm{imp}}, \phi_{\mathrm{T,imp}} \ge 0$, and

unknown parameters (ρ^*, λ^*), define $\widetilde{q}_\nu(u) := u + u(u^2 + \epsilon_{\mathrm{imp}}^2)^{(\nu-1)/2}$, $L_q := 1 + \epsilon_{\mathrm{imp}}^{\nu-1}$, and consider

$$\begin{aligned} &y_{t+1} = y_t - u_t + w_t^y, \qquad z_{t+1} = \rho^* z_t + \lambda^* \widetilde{q}_\nu(u_t) + w_t^z, \qquad w_t = (w_t^y, w_t^z)^\top \sim \mathcal{N}(0, \Sigma_w), \\ &c(x_t, u_t) = z_t u_t + \frac{\eta_{\mathrm{imp}}}{2} u_t^2 + \frac{\phi_{\mathrm{imp}}}{2} y_t^2, \qquad h(x_H) = \frac{\phi_{\mathrm{T,imp}}}{2} y_H^2, \qquad t = 0, \ldots, H-1, \end{aligned} \tag{87}$$

In this example, we set $f^{\mathrm{pub}}(x,u) = \begin{pmatrix} y - u \\ 0 \end{pmatrix}$, $\theta_t^* = \theta^* := \begin{pmatrix} 0 & 0 \\ \rho^* & \lambda^* \end{pmatrix}$, $\phi^{(t)}(x,u) = (z, \widetilde{q}_\nu(u))^\top$, $d_t = 2$ for any $t = 0, \ldots, H-1$. Meanwhile, let $0 \le \bar{\rho} < 1$ and $0 < \underline{\lambda} \le \overline{\lambda}$ be known bounds. We set $\Theta_t = \Theta$ for every $t = 0, \ldots, H-1$, where Θ is the compact and convex set $\left\{ \left(\begin{smallmatrix} 0 & 0 \\ \rho & \lambda \end{smallmatrix} \right) \in \mathbb{R}^{2\times 2} : |\rho| \le \bar{\rho},\ \underline{\lambda} \le \lambda \le \overline{\lambda} \right\}$. Furthermore, we set the parameters in Algorithm 1 as $M = 3, \eta_t = 1, \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \sigma_{\mathrm{id},t}^2 = \sigma_0^2 (t+1)^{-1/2}, t = 0, \ldots, H-1$. The realization of Algorithm 1 follows the discussion in Appendix B.

We introduce notation and an assumption to analyze the regret bound for this example. Let $\alpha \in (0,1)$ and set

$$\begin{aligned} &R_x(H, \alpha) = R_{x, H-1}(\alpha/2), \quad r_e(H, \alpha) = \sigma_0 \sqrt{2 \log(4H/\alpha)}, \quad x = (y, z) \in \mathbb{R}^2, \\ &\mathcal{K}_s(\alpha) = \{(x,u) \in \mathbb{R}^3 \mid \|x\|_2 \le R_x(H,\alpha),\ |u| \le \|\widehat{\mathbf{K}}\|_{\mathrm{op}} R_x(H,\alpha) + B_{\mathrm{ctl}}(H) + r_e(H,\alpha)\}. \end{aligned} \tag{88}$$

For $f(x,u) = f_{\theta^*}(x,u)$, we define

$$\begin{aligned} &\mathcal{X}_s(\alpha) := \{x : \exists u,\ (x,u) \in \mathcal{K}_s(\alpha)\},\ u_t^\star(x) \in \arg\min_{u \in \mathbb{R}^{d_u}} \mathsf{Q}_{t,H}(x,u), \\ &\mathsf{Q}_{t,H}(x,u) := c(x,u) + H \mathbb{E}_\tau [V_{t+1}^*(f(x,u) + w_t)], \quad \forall s \in \mathcal{T}_H, \quad t = 0, \ldots, H-1. \end{aligned} \tag{89}$$

For the sampled model $(f_{\theta_s}, \widehat{\Sigma}_{w,s})$, we define

$$\begin{aligned} &\widehat{V}_{s,H}^*(x) := h(x), \\ &\widehat{\mathsf{Q}}_{s,t,H}(x,u) := c(x,u) + H \mathbb{E}_{\tau, \widehat{w} \sim \mathcal{N}(0, \widehat{\Sigma}_{w,s})} [\widehat{V}_{s,t+1}^*(f_{\theta_s}(x,u) + \widehat{w})], \\ &\widehat{V}_{s,t}^*(x) := H^{-1} \inf_{u \in \mathbb{R}} \widehat{\mathsf{Q}}_{s,t,H}(x,u), \quad \forall s \in \mathcal{T}_H, \quad t = 0, \ldots, H-1. \end{aligned} \tag{90}$$

Assumption 7. For Θ, we have $\theta^* \in \Theta$. For every $i = 0, \ldots, H-1$, the initial state and process noise satisfy

$$x_0 \sim \nu_0 = \mathcal{N}(0, I_2), \quad w_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \Sigma_w), \quad \lambda_w I_2 \preceq \Sigma_w \preceq \bar{\lambda}_w I_2, \quad t = 0, \ldots, H-1$$

with $0 < \lambda_w \le \bar{\lambda}_w$. For $\eta_{\mathrm{imp}}, L_q, \overline{\lambda}, \overline{\rho}$, and any $u \in \mathbb{R}$, we have $1 \le \frac{\partial \widetilde{q}_\nu(u)}{\partial u} \le L_q$ and $\eta_{\mathrm{imp}} > \frac{2\overline{\lambda} L_q}{1 - \overline{\rho}}$. There are constants $B_{\mathrm{ctl}}(H)$, $L_H \ge \mu_H > 0$, and $L_{\theta,H}, L_{\Sigma,H} \ge 0$ such that

$$\max\{B_{\mathrm{ctl}}(H), L_H, \mu_H^{-1}, L_{\theta,H}, L_{\Sigma,H}, L_{\mu,H}\} = \widetilde{O}(1). \tag{91}$$

There exists a constant $L_{\mu,H}$, such that

$$\|\mu_{\theta_s}^{(i)}(x) - \mu_{\theta_s}^{(i)}(x')\|_2 \le L_{\mu,H} \|x - x'\|_2, \qquad s \in \mathcal{T}_H, \quad 0 \le i < H. \tag{92}$$

Set $\widehat{\mathbf{K}} = (K_y, K_z)$. For every $\alpha' \in (0,1)$, every $s \in \mathcal{T}_H$, every $0 \le t \le H-2$ and $x \in \mathcal{X}_s(\alpha')$ with $s \in \mathcal{T}_H$, there is a measurable choice $u_t^\star(x) \in \arg\min_{u \in \mathbb{R}} \mathsf{Q}_{t,H}(x,u)$ satisfying

$$|u_t^\star(x) - \widehat{\mathbf{K}} x| \le B_{\mathrm{ctl}}(H), \qquad \nabla_{uu}^2 \mathsf{Q}_{t,H}(x,u) \preceq L_H I, \quad u \in \mathbb{R}. \tag{93}$$

The function $\mathsf{Q}_{t,H}(x, \cdot)$ is twice continuously differentiable. Define the interval

$$\mathcal{I}_{s,t}(x) := \{(1-\tau) u_t^\star(x) + \tau(\bar{u}_{s,t}(x)) : 0 \le \tau \le 1\}$$

with $\bar{u}_{s,t}(x) = \widehat{\mathbf{K}} x + \mathrm{Proj}_{\mathbb{B}^{d_u}(0, B_{\mathrm{ctl}}(H))}\big(\mu_{\theta_s}^{(t)}(x)\big)$ For every $s \in \mathcal{T}_H$ and every $\alpha \in (0,1)$, on $\mathcal{G}_s(a_s, c_s, \alpha)$, assume

$$\begin{aligned} &\mu_H I \preceq \nabla_{uu}^2 \mathsf{Q}_{t,H}(x,u), \\ &\left\| \nabla_u \mathsf{Q}_{t,H}(x,u) - \nabla_u \widehat{\mathsf{Q}}_{s,t,H}(x,u) \right\|_2 \le b_{s,H}^{\mathrm{grad}} := L_{\theta,H} a_s + L_{\Sigma,H}(c_s), \\ &s \le t < \min\{s + 2M, H-1\}, \quad x \in \mathcal{X}_s(\alpha), \quad u \in \mathcal{I}_{s,t}(x). \end{aligned} \tag{94}$$

Meanwhile, the prespecified deterministic errors $\varepsilon_s^{\mathrm{Bell}} \geq 0$ satisfy

$$\sum_{s\in\mathcal{T}_H} \ell_{s,H}\varepsilon_s^{\mathrm{Bell}} \leq O(\sqrt{H}), \qquad \sup_{\substack{s\leq t<(s+2M)\wedge H \\ x\in\mathcal{X}_s(\alpha)}} \left[\widehat{\mathsf{Q}}_{s,t,H}(x,\bar{u}_{s,t}(x)) - \inf_{u\in\mathbb{R}} \widehat{\mathsf{Q}}_{s,t,H}(x,u)\right] \leq \varepsilon_s^{\mathrm{Bell}}. \tag{95}$$

Denoting the policy from Algorithm 1 by $\pi^{\mathrm{A_I}}$, we propose the following assumption and theoretical result.

Proposition 5 (Nonlinear transient price impact)**.** *Under Assumption 7 and the above model settings, Assumptions 1–3 hold, and it holds that*

$$\mathcal{J}^H(\pi^{\mathrm{A_I}}) - \mathcal{J}^{H,*} \leq \widetilde{O}(H^{-1/2}). \tag{96}$$

Proof. **Step 1.** We first verify Assumption 1. Taking $\mathscr{L} = \sqrt{1+\overline{\lambda}^2 L_q^2}$, for Assumption 1(a), we have $f(0,0)=0$ and

$$\|f(x,u)-f(x,u')\|_2^2 \leq 1+\overline{\lambda}^2 L_q^2|u-u'|^2 = \mathscr{L}^2|u-u'|^2.$$

Thus Assumption 1(a) holds. To verify Assumption 1(b), choose $q_* \in (\bar{\rho}^2, 1)$ and $\eta_* > 0$ such that $(1+\eta_*)\bar{\rho}^2 \leq q_*$, and set

$$c_* := \min\left\{1, \frac{q_*}{(1+\eta_*^{-1})\overline{\lambda}^2 L_q^2}\right\}, \qquad K_\star = (1,0).$$

Setting $\mathscr{K} \geq 1$, $\overline{v}c_* \geq \underline{v}$, and

$$K_\star = [1 \quad 0], \qquad P_\star = \mathrm{diag}(\overline{v}, \overline{v}c_*),$$

we have

$$\begin{aligned}\|f(x,K_\star x)\|_{P_\star}^2 &= \overline{v}c_*\big(\rho z + \lambda\widetilde{q}_\nu(y)\big)^2 \\ &\leq q_*\overline{v}c_* z^2 \leq q_*\overline{v}(y^2+c_* z^2) = q_*\|x\|_{P_\star}^2\end{aligned}$$

with $x=(y,z)^\top \in \mathbb{R}^2$ and $(K_\star, P_\star) \in \mathfrak{S}$. Setting $q_1 = (1+\sqrt{q_*})/2$ and $q_2 = (1-\sqrt{q_*})/2$, it holds that

$$\|f(x,K_\star x)\|_{P_\star} \leq (q_1-q_2)\|x\|_{P_\star}.$$

That completes the verification of Assumption 1.

Step 2. We analyze the approximation error. Since $\theta_t^* = \begin{pmatrix} 0 & 0 \\ \rho^* & \lambda^* \end{pmatrix}$, it holds that $\Delta_t = \alpha_{\mathrm{app},t} = 0$.

Step 3. We next verify Assumption 2, starting with (6). Let μ_j^{dep} denote the residual policy actually deployed at time j. Based on Assumption 7, we have

$$|\mu_j^{\mathrm{dep}}(x)| \leq B_{\mathrm{ctl}}(H), \qquad |\mu_j^{\mathrm{dep}}(x) - \mu_j^{\mathrm{dep}}(x')| \leq L_{\mu,H}\|x-x'\|_2$$

for all $0 \leq j < H$. Define

$$c_{\mathrm{sb},H} := \min\left\{\frac{\sqrt{2\pi}}{8\sqrt{1+(L_q(|K_z|+L_{\mu,H}))^2}}, \frac{\sqrt{2\pi}\left(\sqrt{1-\beta_H^2} - \beta_H L_q(|K_z|+L_{\mu,H})\right)}{8\sqrt{1+(L_q(|K_z|+L_{\mu,H}))^2}}\right\},$$

$$\lambda_{\mathrm{I}} := \min\{1,\lambda_w\}, \gamma_{\mathrm{I},s} := 2c_{\mathrm{sb},H}^2 \min\left\{\lambda_{\mathrm{I}}, \sigma_0^2(s+1)^{-1/2}\right\},\ \beta_H := \frac{1}{2\sqrt{1+(L_q(|K_z|+L_{\mu,H}))^2}}.$$

Note that

$$1 \leq \widetilde{q}_\nu'(e) \leq L_q, \qquad |\widetilde{q}_\nu(e)-\widetilde{q}_\nu(e')| \leq L_q|e-e'|, \quad \forall e,e' \in \mathbb{R}. \tag{97}$$

For $v = (a, b) \in \mathbb{S}^1$, if $|b| \geq \beta_H$, using (97) and the fact that e_j^{id} is Gaussian, for $s \in \mathcal{T}_H$ and $0 \leq j \leq n_s - 1$, we have

$$\mathbb{P}_\tau(|az_j + b\widetilde{q}_\nu(u_j)| < c_{\mathrm{sb},H}\sigma_{\mathrm{id},j} \,|\, \mathscr{H}_{j-1}, x_j) \leq \frac{2c_{\mathrm{sb},H}}{\sqrt{2\pi}\,\beta_H} \leq \frac{1}{2}.$$

If $|b| < \beta_H$, then $|a| \geq \sqrt{1-\beta_H^2}$, we have

$$\begin{aligned} &\Big|az_1 + b\widetilde{q}_\nu\Big(K_y y_j + K_z z_1 + \mu_j^{\mathrm{dep}}(y_j, z_1) + e_j^{\mathrm{id}}\Big) \\ &\quad - \Big(az_2 + b\widetilde{q}_\nu\Big(K_y y_j + K_z z_2 + \mu_j^{\mathrm{dep}}(y_j, z_2) + e_j^{\mathrm{id}}\Big)\Big)\Big| \\ &\geq \left(\sqrt{1-\beta_H^2} - \beta_H L_q\big(|K_z| + L_{\mu,H}\big)\right)|z_2 - z_1|, \qquad \forall z_1, z_2 \in \mathbb{R}. \end{aligned} \tag{98}$$

given (y_j, e_j^{id}). Meanwhile, it is obvious to show that

$$\mathrm{Var}_\tau(z_j \mid \mathscr{H}_{j-1}, y_j) \geq \lambda_{\mathrm{I}}.$$

We therefore have

$$\mathbb{P}_\tau\Big(|az_j + b\widetilde{q}_\nu(u_j)| < c_{\mathrm{sb},H}\sqrt{\lambda_{\mathrm{I}}}\,\Big|\,\mathscr{H}_{j-1}\Big) \leq \frac{2c_{\mathrm{sb},H}}{\sqrt{2\pi}\,\left(\sqrt{1-\beta_H^2} - \beta_H L_q\big(|K_z| + L_{\mu,H}\big)\right)} \leq \frac{1}{2}.$$

Since $\sigma_{\mathrm{id},j}^2 \geq \sigma_0^2(s+1)^{-1/2}$ for $j \leq s-1$, both cases give

$$\mathbb{P}_\tau\left(|v^\top \phi^{(s)}(x_j, u_j)| \geq \sqrt{\gamma_{\mathrm{I},s}/2}\,\middle|\,\mathscr{H}_{j-1}\right) \geq \frac{1}{2}.$$

Thus (6) holds with

$$k_s = 1, \qquad p_s^{\mathrm{sb}} = \frac{1}{2}, \qquad \Gamma_{\mathrm{PE},s} = \gamma_{\mathrm{I},s} I_2.$$

We next analyze the upper bound of $\mathbb{E}_\tau\Big[\sum_{i=0}^{n_s-1}\phi^{(s)}(x_i,u_i)\phi^{(s)}(x_i,u_i)^\top\,\Big|\,\mathscr{H}_{-1}\Big]$. Define

$$C_{x,H} := \sup_{0\leq i<H}\mathbb{E}_\tau\|x_i\|_2^2 < \infty, \qquad C_{\phi,H} := C_{x,H} + 3L_q^2\left(\|\widehat{\mathbf{K}}\|_{\mathrm{op}}^2 C_{x,H} + B_{\mathrm{ctl}}(H)^2 + \sigma_0^2\right).$$

Taking $\phi^{(t)}(x,u) = (z, \widetilde{q}_\nu(u))^\top$ into $\mathbb{E}_\tau\Big[\sum_{i=0}^{n_s-1}\phi^{(s)}(x_i,u_i)\phi^{(s)}(x_i,u_i)^\top\,\Big|\,\mathscr{H}_{-1}\Big]$, we have

$$\mathbb{E}_\tau\left[\sum_{i=0}^{n_s-1}\phi^{(s)}(x_i,u_i)\phi^{(s)}(x_i,u_i)^\top\,\middle|\,\mathscr{H}_{-1}\right] \preceq n_s C_{\phi,H} I_2.$$

This verifies Assumption 2. Moreover, using Gaussian noise setting, we have $C_{x,H} = \widetilde{O}(1)$, and hence $C_{\phi,H} = \widetilde{O}(1)$ when $B_{\mathrm{ctl}}(H) = \widetilde{O}(1)$.

Step 4. We now derive the error bound of θ_s via Theorem 1. Note that the application of Theorem 1 need to satisfies (7). So we choose t_0 sufficiently large and for $0 < \delta < 1$, $s \in \mathcal{T}_H$, set

$$\alpha_{\mathrm{up},s} = \delta_{\mathrm{sb},s} = \delta_{\mathrm{mg},s} = \frac{\delta}{64H^3}, g_{-,s} = \frac{(p_s^{\mathrm{sb}})^2}{32}\,k_s\left\lfloor\frac{n_s}{k_s}\right\rfloor\gamma_{\mathrm{I},s},$$

$$\epsilon_s^2 = \begin{cases} (\mathrm{diam}_{\mathrm{F}}(\Theta))^2, & s < t_0, \\ \dfrac{4z_s^2 + 2\log(64H^3\mathcal{M}_s/\delta)}{g_{-,s}}, & s \geq t_0. \end{cases}$$

Here, t_0 is sufficient large number such that for any $s \geq t_0$, inequality (7) holds. Since $\gamma_{\mathrm{I},s} = O((s+1)^{-1/2})$, and all remaining factors in the RHS of 7 are logarithmic with H, we have $t_0 = \widetilde{O}(1)$. It is easy to verify $\mathfrak{G}_s = \frac{1}{2}g_{-,s}\epsilon_s^2 - 2z_s^2 = \log\frac{64H^3\mathcal{M}_s}{\delta}$, $\mathcal{M}_s e^{-\eta_s\mathfrak{G}_s} = \frac{\delta}{64H^3}$. Hence, using Theorem 1 when $s \geq t_0$ and the deterministic diameter bound $(\mathrm{diam}_{\mathrm{F}}(\Theta))^2$ when $s < t_0$, we have

$$\mathbb{P}_\tau(\|\theta_s - \theta^*\|_{\mathrm{F}} > \epsilon_s) \leq \frac{\delta}{16H^3}\mathbf{1}\{s \geq t_0\}. \tag{99}$$

Step 5. We now analyze the error bound of $\widehat{\Sigma}_{w,s}$ via Theorem 2. Setting

$$\alpha_s = \delta/(64H^3),$$
$$\zeta_s = s^{-1/2} + \frac{2M}{sH^2} \sum_{\substack{j\in\mathcal{T}_H \\ t_0\le j\le s}} U_j(\alpha_s)^2 + \frac{(4\bar{\rho}^2 + (\bar{\lambda}-\overline{\lambda})^2)^2 \left[R_{x,H-1}(\alpha_s)^2 + (1+\epsilon_{\mathrm{imp}}^{\nu-1})^2 R_{u,H-1}(\alpha_s)^2\right]}{H^2},$$
$$\epsilon_s^{\mathrm{cov}} = \max\left\{ \sqrt{\frac{64\bar{\lambda}_w[\beta_s(\alpha_s)+\zeta_s]L_H}{s}},\ \bar{\lambda}_w\sqrt{\frac{8(1+d_x\log 9)\log(128H^3/\delta)}{c_{\mathrm{HW}}s}},\ \frac{4\bar{\lambda}_w\log(128H^3/\delta)}{c_{\mathrm{HW}}s} \right\}$$

for $s\in\mathcal{T}_H$, we get

$$\mathbb{P}_\tau(\|\widehat{\Sigma}_{w,s}-\Sigma_w\|_{\mathrm{op}} > \beta_s(\alpha_s)+\zeta_s+2\epsilon_s^{\mathrm{cov}}) \le \frac{\delta}{4H}. \tag{100}$$

Step 6. We now verify Assumption 3. For any $\alpha'\in(0,1)$, set

$$a_s = \mathbf{1}\{s<t_0\}\big(1+\mathrm{diam}_{\mathrm{F}}(\Theta)\big) + \mathbf{1}\{s\ge t_0\}\epsilon_s,$$
$$c_s = \beta_s(\alpha_s)+\zeta_s+2\epsilon_s^{\mathrm{cov}},$$
$$\delta_s = \alpha' + \frac{\delta}{8H^3} + \frac{\delta}{4H}.$$

For $s\in\mathcal{T}_H$ and $s\le t<(s+2M)\wedge(H-1)$, on $\mathcal{G}_s(a_s,c_s,\alpha')$, combining (94) and (95) gives

$$\frac{\mu_H|\bar{u}_{s,t}(x_t)-u_t^\star(x_t)|^2}{2} \le H\mathcal{D}_t^H(x_t,\bar{u}_{s,t}(x_t)) \le \varepsilon_s^{\mathrm{Bell}} + b_{s,H}^{\mathrm{grad}}|\bar{u}_{s,t}(x_t)-u_t^\star(x_t)|,$$
$$\mathcal{D}_t^H(x_t,u_t) \le \frac{4L_H\varepsilon_s^{\mathrm{Bell}}}{\mu_H H} + \frac{4L_H(b_{s,H}^{\mathrm{grad}})^2}{\mu_H^2 H} + \frac{L_H}{H}|e_t^{\mathrm{id}}|^2, \quad s\le t<(s+2M)\wedge(H-1).$$

At $t=H-1$, the model settings give

$$u_{H-1}^\star(x) = \frac{H\phi_{\mathrm{T,imp}}y - z}{\eta_{\mathrm{imp}}+H\phi_{\mathrm{T,imp}}},$$
$$\widehat{\mathsf{Q}}_{s,H-1,H}(x,u) - \inf_{v\in\mathbb{R}}\widehat{\mathsf{Q}}_{s,H-1,H}(x,v) = H\mathcal{D}_{H-1}^H(x,u) = \frac{\eta_{\mathrm{imp}}+H\phi_{\mathrm{T,imp}}}{2}|u-u_{H-1}^\star(x)|^2 \tag{101}$$

for any $(x,u)\in\mathbb{R}^3$. We therefore have

$$0\le\mathcal{D}_{H-1}^H(x_{H-1},u_{H-1}) \le \frac{\varepsilon_s^{\mathrm{Bell}}}{H}$$

whenever $H-1\in[s,(s+2M)\wedge H)$. Using the stabilizing control $u=\widehat{\mathbf{K}}x$ and Young's inequality, we have

$$|V_t^*(x)| \le C\big(1+\|x\|_2^2\big), \qquad 0\le t\le H.$$

By applying the nonlinear impact law $\widetilde{q}_\nu(u)$, the condition $\eta_{\mathrm{imp}} > \frac{2\bar{\lambda}L_q}{1-\bar{\rho}}$ in Assumption 7, and the coercivity of the control cost then imply, it holds that, for every $(x,u)\in\mathbb{R}^2\times\mathbb{R}$,

$$0\le\mathcal{D}_t^H(x,u)\le C\big(1+\|x\|_2^2+u^2\big), \qquad 0\le t<H,$$

where C depends only on the fixed model (87), and is independent of t, H, and $B_{\mathrm{ctl}}(H)$. Moreover, (38) and the Gaussian fourth moments imply

$$\sup_{t<H}\mathbb{E}_\tau\left[\|x_t\|_2^4+|u_t|^4\right] \le C\big(1+B_{\mathrm{ctl}}(H)^2\big)^2.$$

Consequently,

$$\max_{s\le t<(s+2M)\wedge H}\|\mathcal{D}_t^H(x_t,u_t)\|_{L^2(\mathbb{P}_\tau)} \le C\left[1+\sup_{t<H}\left(\mathbb{E}_\tau[\|x_t\|_2^4+|u_t|^4]\right)^{1/2}\right] \le C\big(1+B_{\mathrm{ctl}}(H)^2\big).$$

Thus, (91) and the preceding display verify Assumption 3 with

$$\begin{aligned}
&\beta_s^{\mathrm{Bell}}(\alpha') = \left(\frac{4L_H}{\mu_H} + 2\right)\frac{\varepsilon_s^{\mathrm{Bell}}}{H} + \frac{4L_H\big(b_{s,H}^{\mathrm{grad}}(\alpha')\big)^2}{\mu_H^2 H},\\
&C_s^{\mathrm{id}} = \frac{L_H}{H} + \frac{\eta_{\mathrm{imp}} + H\phi_{\mathrm{T,imp}}}{H}\mathbf{1}\{s + \ell_{s,H} = H\},\\
&p = 2, \qquad C_s^{\mathrm{bad}} \le \frac{C_H}{H},\\
&C_H := H\max_{t<H}\|\mathcal{D}_t^H(x_t, u_t)\|_{L^2} \le O(H(1 + (B_{\mathrm{ctl}}(H))^2)),\\
&B_0^{\mathrm{init}} = \frac{2ML_H}{H}[4(B_{\mathrm{ctl}}(H))^2 + \sigma_0^2] + \frac{2MC_H}{H}\sqrt{\alpha'}.
\end{aligned} \tag{102}$$

Step 7. Finally, we combine all of the analysis above to get the regret bound of $\pi^{\mathrm{A_I}}$. Since $t_0 = \widetilde{O}(1)$, the preceding choices give

$$\begin{aligned}
&\frac{1}{H}\sum_{s\in\mathcal{T}_H}\ell_{s,H}a_s^2 = \widetilde{O}\left(H^{-1/2} + \frac{t_0}{H}\right),\\
&\frac{1}{H}\sum_{s\in\mathcal{T}_H}\ell_{s,H}(c_s)^2 = \widetilde{O}\left(H^{-1/2} + \frac{t_0^2}{H}\right),\\
&\sum_{s\in\mathcal{T}_H}\ell_{s,H}\beta_s^{\mathrm{Bell}}(\alpha') = \widetilde{O}\left(H^{-1/2} + \frac{t_0}{H}\right),\\
&\sum_{s\in\mathcal{T}_H}\ell_{s,H}C_s^{\mathrm{bad}}\delta_s^{1/2} = \widetilde{O}\left(\frac{t_0}{H} + \sqrt{\frac{\delta}{H}}\right).
\end{aligned}$$

Taking $\alpha' = \frac{\delta}{4H}$, $\delta = (H+1)^{-4}$, and using $\sum_{s\in\mathcal{T}_H}\ell_{s,H}\varepsilon_s^{\mathrm{Bell}} = O(\sqrt{H})$, Theorem 3 yields

$$\mathcal{J}^H(\pi^{\mathrm{A_I}}) - \mathcal{J}^{H,*} = \widetilde{O}(H^{-1/2}).$$

□

E GRADIENT FLOW FOR PROBLEM (20)

We study the model-based control problem (20) at a fixed update epoch s. Write

$$\begin{aligned}
&F(x,a) := f_{\theta_s}(x, \widehat{\mathbf{K}}x + a), \qquad \ell(x,a) := H^{-1}c(x, \widehat{\mathbf{K}}x + a),\\
&\Sigma := \widehat{\Sigma}_{w,s} \succ 0, \qquad \mathcal{A} := \mathbb{B}^{d_u}(0, B_{\mathrm{ctl}}).
\end{aligned}$$

Here a is a residual action, so the total control is $\widehat{\mathbf{K}}x + a$. We condition on $x_0 = \bar{x}_0$ and write the residual feedback as $v(t,x) = \mu^{(t)}(x) \in \mathcal{A}$. Since v takes values in $\mathcal{A}$, it agrees with its projection in (20). The conditional problem is

$$\begin{aligned}
&\min_v\ J[v] := \mathbb{E}\left[\sum_{t=0}^{H-1}\ell(x_t, v(t,x_t)) + h(x_H)\right],\\
&\text{s.t.}\quad x_{t+1} = F(x_t, v(t,x_t)) + w_t, \qquad w_t \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,\Sigma), \qquad x_0 = \bar{x}_0.
\end{aligned}$$

For the initial law in (20), the corresponding objective is $\mathcal{J}_s^{\mathrm{mod}}(v) = \int J[v;x]\,\nu_0(\mathrm{d}x)$, where $J[v;x]$ denotes the same conditional cost with initial state x. The convergence argument follows the local optimal control construction and two-case strategy of Zhou & Lu (2025, Theorem 4.3 and Appendix D), using the discrete-time Bellman recursion below.

For a deterministic Markov feedback v, define

$$V_v(t,x) := \mathbb{E}\left[\sum_{r=t}^{H-1}\ell(x_r, v(r,x_r)) + h(x_H)\,\middle|\, x_t = x\right].$$

The Bellman recursion gives

$$V_v(t_1, x) = \mathbb{E}\left[\sum_{r=t_1}^{t_2-1} \ell(x_r, v(r, x_r)) + V_v(t_2, x_{t_2}) \middle| x_{t_1} = x\right],$$

for $0 \le t_1 \le t_2 \le H$. Let v^* denote an optimal feedback and $V^* := V_{v^*}$ its value function. The dynamic programming principle yields

$$V^*(t_1, x) = \inf_v \mathbb{E}\left[\sum_{r=t_1}^{t_2-1} \ell(x_r, v(r, x_r)) + V^*(t_2, x_{t_2}) \middle| x_{t_1} = x\right], \tag{103}$$

where the infimum is over adapted residual controls on $\{t_1, \dots, t_2 - 1\}$ with values in $\mathcal{A}$. Taking $t_2 = t_1 + 1$ in (103) gives a deterministic Markov minimizer under the assumptions below.

To include the deterministic initial state, set $\lambda_0 := \delta_{\bar{x}_0}$ and $\lambda_t(\mathrm{d}x) := \mathrm{d}x$ for $t \ge 1$. Let ρ_t^v be the density of the state law relative to λ_t; in particular, $\rho_0^v(\bar{x}_0) = 1$. Then

$$\rho_{t+1}^v(y) = \int_{\mathbb{R}^{d_x}} \rho_t^v(x) p_\Sigma\big(y - F(x, v(t, x))\big)\, \lambda_t(\mathrm{d}x), \qquad 0 \le t \le H - 2,$$

where p_Σ is the density of $\mathcal{N}(0, \Sigma)$. Every integral at $t = 0$ below is therefore a point value at $\bar{x}_0$. The value functions satisfy

$$\begin{aligned} &V_v(t, x) = \ell(x, v(t, x)) + \mathbb{E}_w[V_v(t + 1, F(x, v(t, x)) + w)], \quad t = 0, \dots, H - 1, \\ &V_v(H, x) = h(x), \end{aligned}$$

and

$$\begin{aligned} &V^*(t, x) = \min_{a \in \mathcal{A}} \{\ell(x, a) + \mathbb{E}_w[V^*(t + 1, F(x, a) + w)]\}, \quad t = 0, \dots, H - 1, \\ &V^*(H, x) = h(x). \end{aligned} \tag{104}$$

Here and below, $w \sim \mathcal{N}(0, \Sigma)$. Define

$$\begin{aligned} Q_v(t, x, a) &:= \ell(x, a) + \mathbb{E}_w[V_v(t + 1, F(x, a) + w)], \\ Q^*(t, x, a) &:= \ell(x, a) + \mathbb{E}_w[V^*(t + 1, F(x, a) + w)]. \end{aligned}$$

Lemma 4. *The functional derivative of J in $L^2(\lambda)$ is*

$$\frac{\delta J[v]}{\delta v}(t, x) = \rho_t^v(x) \nabla_a Q_v(t, x, v(t, x)),$$

where ∇_a denotes the gradient in the third argument.

Proof. By the performance difference lemma (Kakade & Langford, 2002),

$$J[v'] - J[v] = \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \rho_t^{v'}(x) \big[Q_v(t, x, v'(t, x)) - Q_v(t, x, v(t, x))\big]\, \lambda_t(\mathrm{d}x). \tag{105}$$

For an admissible perturbation ϕ, the definition of the functional derivative gives

$$\lim_{\varepsilon \to 0} \frac{J[v + \varepsilon\phi] - J[v]}{\varepsilon} = \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \left\langle \frac{\delta J[v]}{\delta v}(t, x), \phi(t, x) \right\rangle \lambda_t(\mathrm{d}x).$$

By (105),

$$\frac{J[v + \varepsilon\phi] - J[v]}{\varepsilon} = \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \rho_t^{v+\varepsilon\phi}(x) \frac{Q_v(t, x, v(t, x) + \varepsilon\phi(t, x)) - Q_v(t, x, v(t, x))}{\varepsilon}\, \lambda_t(\mathrm{d}x).$$

Dominated convergence therefore yields

$$\lim_{\varepsilon \to 0} \frac{J[v + \varepsilon\phi] - J[v]}{\varepsilon} = \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \langle \rho_t^v(x) \nabla_a Q_v(t, x, v(t, x)), \phi(t, x) \rangle\, \lambda_t(\mathrm{d}x),$$

where the inner products are in $\mathbb{R}^{d_u}$. This proves the claim. □

Motivated by Lemma 4, consider the gradient flow

$$\frac{\mathrm{d}}{\mathrm{d}\tau} v^\tau(t,x) = -\rho_t^{v^\tau}(x) \nabla_a Q_{v^\tau}(t,x,v^\tau(t,x)), \qquad \tau \geq 0, \tag{106}$$

with $v^0 \in \mathcal{V}$. At $t = 0$, (106) is imposed at $x = \bar{x}_0$. The admissible feedback class is

$$\mathcal{V} := \big\{ v : \ v(t,\cdot) \in C^2(\mathbb{R}^{d_x}; \mathbb{R}^{d_u}), \quad \|v(t,\cdot)\|_\infty \leq B_{\mathrm{ctl}}, \\ \max\{\|\nabla_x v(t,\cdot)\|_\infty, \|\nabla_x^2 v(t,\cdot)\|_\infty\} \leq K_{\mathrm{reg}}, \quad t = 0, \dots, H-1 \big\}.$$

Here K_{reg} is a regularity bound, and the sup norms use Euclidean norms for vectors and operator norms for derivatives.

Assumption 8. It holds that:

(a) Fix $\Delta > B_{\mathrm{ctl}}$. There exists $K_0 > 0$ such that $\ell \in C^3(\mathbb{R}^{d_x} \times \mathbb{R}^{d_u})$, $F \in C^3(\mathbb{R}^{d_x} \times \mathbb{R}^{d_u})$, $h \in C^3(\mathbb{R}^{d_x})$, and

$$\max\left\{ \|\ell(x,a)\|_{C^3(\mathbb{R}^{d_x} \times B_2(\Delta))}, \|F(x,a)\|_{C^3(\mathbb{R}^{d_x} \times B_2(\Delta))}, \|h(x)\|_{C^3} \right\} \leq K_0$$

with $B_2(\Delta) := \{a \in \mathbb{R}^{d_u} : \|a\|_2 \leq \Delta\}$.

(b) There exist constants $0 < \mu_Q \leq L_Q$ such that for any bounded function $V(t+1,\cdot)$ satisfying $\|V(t+1,\cdot)\|_\infty \leq M$ and $\|\nabla_x^2 V(t+1,\cdot)\|_\infty \leq M_2$ (where M, M_2 are given in Lemma 6), any (t,x), and any a with $\|a\|_2 \leq B_{\mathrm{ctl}}$, the function $a \mapsto Q(t,x,a) = \ell(x,a) + \mathbb{E}_w[V(t+1, F(x,a)+w)]$ is twice continuously differentiable with respect to a, and its Hessian satisfies

$$\mu_Q I \preceq \nabla_a^2 Q(t,x,a) \preceq L_Q I.$$

The pointwise minimizers of Q_v and Q^* used below lie in the interior of $\mathcal{A}$.

(c) Letting $(v^\tau)_{\tau \geq 0} \subseteq \mathcal{V}$ be the gradient flow (106) and $v^* \in \mathcal{V}$ the optimal control of Lemma 6 below, it holds that

$$D := \sup_{\tau \geq 0} \|v^\tau - v^*\|_{L^2(\lambda)} < \infty,$$
$$\|\phi\|_{L^2(\lambda)}^2 := \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \|\phi(t,x)\|_2^2 \lambda_t(\mathrm{d}x).$$

Lemma 5. *Let $\psi : \mathbb{R}^{d_x} \to \mathbb{R}$ be bounded measurable with $\|\psi\|_\infty \leq M$. Then $\Psi(z) := \mathbb{E}_w[\psi(z+w)]$ belongs to $C^\infty(\mathbb{R}^{d_x})$, and for every multi-index α there is a constant $c_{\alpha,\Sigma}$ such that $\|\partial^\alpha \Psi\|_\infty \leq c_{\alpha,\Sigma} M$. In particular, $c_{\mathbf{0},\Sigma} = 1$, which is independent of Σ.*

Proof. We omit the proof as it is a standard result of Gaussian mollification. □

Lemma 6. *Under Assumption 8:*

(a) For every $v \in \mathcal{V}$, there exist constants $M = M(K_0, H)$, $M_2 = M_2(K_0, K_{\mathrm{reg}}, H, \Sigma)$, independent of $v \in \mathcal{V}$, such that

$$\max_{0 \leq t \leq H} \|V_v(t,\cdot)\|_\infty \leq M, \qquad \max_{0 \leq t \leq H} \|\nabla_x^2 V_v(t,\cdot)\|_\infty \leq M_2.$$

(b) For K_{reg} sufficiently large, there is a unique pair (V^*, v^*) with $v^* \in \mathcal{V}$ satisfying the Bellman optimal equation (104)*.*

Proof. We proceed by backward induction on $t = H, H-1, \dots, 0$.

Base case. For any $v \in \mathcal{V}$, it holds that $V_v(H,x) = h(x)$; by Assumption 8, $\|h\|_\infty \leq K_0 =: M_H$ and $\|\nabla^2 h(x)\|_\infty \leq K_0 =: M_{2,H}$.

Inductive step. Assume $\|V_v(t+1,\cdot)\|_\infty \leq M_{t+1}$ (independent of v), is C^2, and satisfies $\|\nabla_x^2 V_v(t+1,x)\|_\infty \leq M_{2,t+1}$ (independent of v). By Lemma 5, $\Psi_t^v(z) := \mathbb{E}_w[V_v(t+1, z+w)]$

belongs to $C^\infty(\mathbb{R}^{d_x})$ with $\|\Psi_t^v\|_\infty \le M_{t+1}$ and $\|\nabla^2\Psi_t^v\|_\infty \le c_2 M_{t+1}$ (c_2 depending only on Σ). Hence

$$V_v(t,x) = \ell(x, v(t,x)) + \Psi_t^v(F(x, v(t,x)))$$

is a composition of C^2 functions ($\ell, F \in C^3$, $v(t,\cdot) \in C^2$, $\Psi_t^v \in C^\infty$), and $V_v(t,\cdot) \in C^2$. Moreover, it holds that $\|V_v(t,\cdot)\|_\infty \le K_0 + M_{t+1} =: M_t$, and chain rule implies $\|\nabla_x^2 V_v(t,\cdot)\|_\infty \le C(K_0, K_{\mathrm{reg}}, c_2, M_{t+1}) =: M_{2,t}$ (the constant C depending only on $K_0, K_{\mathrm{reg}}, H, \Sigma$). Hence, setting $M := \max_t M_t$, $M_2 := \max_t M_{2,t}$ completes the induction and (a) holds.

We now show (b). First, we claim that if $\|V^*(t+1,\cdot)\|_\infty \le M$, then there exists $K_1 = K_1(K_0, M, \Sigma)$ (independent of K_{reg}), such that every joint partial derivative of $Q^*(t,\cdot,\cdot)$ in (x,a) of order at most 3 has norm at most K_1. Indeed, recall $Q^*(t,x,a) = \ell(x,a) + \Psi_t(F(x,a))$ with $\Psi_t(z) := \mathbb{E}_w[V^*(t+1, z+w)]$. By Lemma 5, $\Psi_t \in C^\infty$ with $\|\partial^\alpha \Psi_t\|_\infty \le c_{\alpha,\Sigma} M$ for every multi-index α, $c_{\alpha,\Sigma}$ depending only on Σ, α. Hence, every derivative of $\Psi_t \circ F$ up to order 3 is a finite (dimension-dependent) sum of products of $\mathbf{D}^j F$ ($j \le 3$, bounded by K_0 via Assumption 8) and $\mathbf{D}^i \Psi_t$ evaluated at $F(x,a)$ ($i \le 3$, bounded by $c_{i,\Sigma} M$). Adding the direct contribution $\mathbf{D}^k \ell \le K_0$ gives the claim, with K_1 the resulting finite sum of bounds. For $t = H-1, \dots, 0$, by Assumption 8(b), $v^*(t,x) := \operatorname*{arg\,min}_{\|a\| \le B_{\mathrm{ctl}}} Q^*(t,x,a)$ is well defined, unique, and satisfies $\nabla_a Q^*(t,x,v^*(t,x)) = 0$. The first-order condition follows from the interior-minimizer condition in Assumption 8(b). Let $A(x) := \nabla_a^2 Q^*(t,x,v^*(t,x))$, $B(x) := D_x \nabla_a Q^*(t,x,v^*(t,x))$. By Assumption 8(b), $\mu_Q I \preceq A(x) \preceq L_Q I$, so $\|A(x)^{-1}\| \le 1/\mu_Q$; moreover, it holds that $\|B(x)\| \le K_1$. The implicit function theorem gives $v^*(t,\cdot) \in C^2$ with

$$\|\nabla_x v^*(t,x)\|_\infty = \|A(x)^{-1} B(x)\|_\infty \le \frac{K_1}{\mu_Q} =: \Lambda_1.$$

Differentiating $A(x), B(x)$ once more along the graph $x \mapsto (x, v^*(t,x))$ via the chain rule brings in third-order joint partials of $Q^*(t,\cdot,\cdot)$, each again bounded by K_1:

$$\max\left\{ \left\| \frac{\mathrm{d}}{\mathrm{d}x} A(x) \right\|_\infty, \left\| \frac{\mathrm{d}}{\mathrm{d}x} B(x) \right\|_\infty \right\} \le K_1(1+\Lambda_1) =: K_2.$$

By $\frac{\mathrm{d}}{\mathrm{d}x}(A^{-1}) = -A^{-1}(\frac{\mathrm{d}}{\mathrm{d}x}A)A^{-1}$, $\left\| \frac{\mathrm{d}}{\mathrm{d}x}(A(x)^{-1}) \right\|_\infty \le K_2/\mu_Q^2$. The chain rule gives

$$\|\nabla^2 v^*(t,\cdot)\|_\infty \;\le\; \frac{K_2}{\mu_Q^2} K_1 + \frac{K_2}{\mu_Q} =: \Lambda_2(K_0, \mu_Q, M, \Sigma).$$

Since M is independent of K_{reg}, choose

$$K_{\mathrm{reg}} \ge \Lambda(K_0, \mu_Q, M, \Sigma) := \Lambda_1 + \Lambda_2.$$

This gives the derivative bounds defining $\mathcal{V}$; the action bound follows from $v^*(t,x) \in \mathcal{A}$. Thus $v^* \in \mathcal{V}$. □

Lemma 7. *Under Assumption 8, for all $v \in \mathcal{V}$,*

$$\frac{\mu_Q}{2} A[v] \le J[v] - J[v^*] \le \frac{L_Q}{2} A[v], \tag{107}$$
$$A[v] := \sum_{t=0}^{H-1} \int_{\mathbb{R}^{d_x}} \rho_t^v(x) \|v(t,x) - v^*(t,x)\|_2^2 \lambda_t(\mathrm{d}x).$$

Proof. By (105) and Assumption 8(b), the proof of this lemma is straightforward and thus omitted. □

Theorem 4. *Under Assumption 8 and along the gradient flow* (106)*, it holds that*

$$\lim_{\tau \to \infty} J[v^\tau] = J[v^*].$$

Proof. Noticing that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}J[v^\tau] = \left\langle \frac{\delta J}{\delta v}[v^\tau], \frac{\mathrm{d}}{\mathrm{d}\tau}v^\tau \right\rangle_{L^2(\lambda)} = -\sum_{t=0}^{H-1}\int_{\mathbb{R}^{d_x}} \rho_t^{v^\tau}(x)^2 \|\nabla_a Q_{v^\tau}(t,x,v^\tau(t,x))\|_2^2 \lambda_t(\mathrm{d}x) \le 0,$$

$J[v^\tau]$ is nonincreasing in τ; by Lemma 7, $J[v^\tau] \ge J[v^*]$. Hence, $L := \lim_{\tau\to\infty} J[v^\tau]$ exists and $L \ge J[v^*]$. Write $y(\tau) := J[v^\tau] - J[v^*] \ge 0$ and $\eta := L - J[v^*] \ge 0$. Suppose, for contradiction, $\eta > 0$. For $\tau \ge 0$ and $k \in \{1,2\}$, define the weighted norm

$$\|\phi\|_{\tau,k}^2 := \sum_{t=0}^{H-1}\int_{\mathbb{R}^{d_x}} \rho_t^{v^\tau}(x)^k \|\phi(t,x)\|_2^2 \lambda_t(\mathrm{d}x)$$

with the $t=0$ term always $\|\phi(0,\bar{x}_0)\|_2^2$. By Assumption 8(b) and Cauchy–Schwarz inequality,

$$\dot{y}(\tau) = -\|\nabla_a Q_{v^\tau}(\cdot,\cdot,v^\tau)\|_{\tau,2}^2 \;\le\; -\mu_Q^2 \,\|v^\tau - v^{\tau\diamond}\|_{\tau,2}^2 \tag{108}$$

with $v^{\tau\diamond}(t,x) := \underset{\|a\|_2 \le B_{\mathrm{ctl}}}{\arg\min}\; Q_{v^\tau}(t,x,a)$.

Case (a): there exist $\mu_0, \tau_0 > 0$ such that

$$\|v^\tau - v^{\tau\diamond}\|_{\tau,2} \;\ge\; \mu_0 \,\|v^\tau - v^*\|_{\tau,2} \qquad (\tau \ge \tau_0). \tag{109}$$

In this case, by (108) and (109), it holds that $\dot{y}(\tau) \le -\mu_Q^2\mu_0^2\, B(\tau)$ with $B(\tau) := \|v^\tau - v^*\|_{\tau,2}^2$. For each fixed $t \in \{0,\dots,H-1\}$, Cauchy–Schwarz inequality gives

$$\int_{\mathbb{R}^{d_x}} \rho_t^{v^\tau}(x)\|\phi_t^\tau(x)\|_2^2\lambda_t(\mathrm{d}x) \le \left(\int_{\mathbb{R}^{d_x}} \rho_t^{v^\tau}(x)^2\|\phi_t^\tau(x)\|_2^2\lambda_t(\mathrm{d}x)\right)^{1/2}\left(\int_{\mathbb{R}^{d_x}} \|\phi_t^\tau(x)\|_2^2\lambda_t(\mathrm{d}x)\right)^{1/2}.$$

with $\phi_t^\tau(x) := v^\tau(t,x) - v^*(t,x)$. Summing over t and applying Cauchy–Schwarz once more, we may conclude that

$$\begin{aligned} A[v^\tau] \le& \sum_{t=0}^{H-1}\left(\int_{\mathbb{R}^{d_x}} \rho_t^{v^\tau}(x)^2\|\phi_t^\tau(x)\|_2^2\lambda_t(\mathrm{d}x)\right)^{1/2}\left(\int_{\mathbb{R}^{d_x}} \|\phi_t^\tau(x)\|_2^2\lambda_t(\mathrm{d}x)\right)^{1/2} \\ \le& B(\tau)^{1/2}\|v^\tau - v^*\|_{L^2(\lambda)} \le DB(\tau)^{1/2}, \end{aligned}$$

where the last inequality holds due to Assumption 8(c). By the RHS of (107), $A[v^\tau] \ge \frac{2}{L_Q} y(\tau)$; thus, $B(\tau) \ge \left(\frac{A[v^\tau]}{D}\right)^2 \ge \frac{4}{L_Q^2 D^2}\, y(\tau)^2$. Substituting back, we have

$$\dot{y}(\tau) \le -\kappa\, y(\tau)^2 \;\text{ with }\; \kappa := \frac{4\mu_Q^2\mu_0^2}{L_Q^2 D^2} > 0 \;\text{ for }\; \tau \ge \tau_0. \tag{110}$$

If $y(\tau_0) = 0$, then $y \equiv 0$ for $\tau \ge \tau_0$, so $\eta = 0$, contradicting $\eta > 0$. If $y(\tau_0) > 0$, as long as $y(\tau) > 0$, (110) implies $\frac{\mathrm{d}}{\mathrm{d}\tau}(y^{-1}) = -y^{-2}\dot{y} \ge \kappa$, so $y(\tau)^{-1} \ge y(\tau_0)^{-1} + \kappa(\tau - \tau_0) \to \infty$, whence $y(\tau) \to 0$, i.e. $\eta = 0$. Again, we arrive at a contradiction. Thus Case (a) forces $\eta = 0$.

Case (b): (109) **fails for every $\mu_0, \tau_0 > 0$.**

In this case, taking $\mu_0 = 1/k$ and $\tau_0 = k$ for $k = 1, 2, \dots$, there exist $\tau_k \to \infty$ such that

$$\|v_k - v_k^\diamond\|_{k,2} \;\le\; \frac{1}{k}\,\|v_k - v^*\|_{k,2} =: \frac{1}{k}\,\delta_k. \tag{111}$$

Here, we write $v_k := v^{\tau_k}$, $v_k^\diamond := v^{\tau_k\diamond}$ and $\|\cdot\|_{k,2} := \|\cdot\|_{\tau_k,2}$ for the sake of simplicity. We may observe that

$$\delta_k^2 = \sum_{t=0}^{H-1}\int_{\mathbb{R}^{d_x}} \rho_t^{v_k}(x)^2\|v_k(t,x) - v^*(t,x)\|_2^2\lambda_t(\mathrm{d}x) \le 4B_{\mathrm{ctl}}^2[1 + (H-1)p_{\max}] < \infty,$$

uniformly in k with $p_{\max} := \max_{x\in\mathbb{R}^{d_x}} p_\Sigma(x)$. Here, the last inequality holds due to

$$\begin{aligned}\rho_t^{v_k}(x) &= \int_{\mathbb{R}^{d_x}} \rho_{t-1}^{v_k}(x') p_\Sigma(x - F(x', v_k(t-1,x')))\lambda_{t-1}(\mathrm{d}x') \\ &\le p_{\max} \int_{\mathbb{R}^{d_x}} \rho_{t-1}^{v_k}(x')\lambda_{t-1}(\mathrm{d}x') = p_{\max}, \qquad t = 1,\dots,H-1.\end{aligned}$$

Hence, by (111), it holds that

$$\epsilon_k := \|v_k - v_k^\diamond\|_{k,2} \le \frac{2B_{\mathrm{ctl}}\sqrt{1+(H-1)p_{\max}}}{k} \to 0. \tag{112}$$

Step 1. We claim that there exist a subsequence (not relabeled) and v_∞ with $v_\infty(t,\cdot) \in C^1(\mathbb{R}^{d_x};\mathbb{R}^{d_u})$ and $\nabla_x v_\infty(t,\cdot)$ globally K_{reg}-Lipschitz for each t, such that for every $t = 0,\dots,H-1$:

(i) $v_k(t,\cdot) \to v_\infty(t,\cdot)$ locally uniformly on $\mathbb{R}^{d_x}$;

(ii) $\nabla_x v_k(t,\cdot) \to \nabla_x v_\infty(t,\cdot)$ locally uniformly on $\mathbb{R}^{d_x}$.

Moreover $\|v_\infty(t,x)\|_2 \le B_{\mathrm{ctl}}$, $\|\nabla_x v_\infty(t,x)\|_\infty \le K_{\mathrm{reg}}$ for all x, and $v_\infty(t,\cdot)$ is twice differentiable a.e. with $\|\nabla_x^2 v_\infty(t,\cdot)\| \le K_{\mathrm{reg}}$ a.e. Indeed, fixing t and by $v_k \in \mathcal{V}$: $\|v_k(t,x)\|_2 \le B_{\mathrm{ctl}}$, $\|\nabla_x v_k(t,x)\|_\infty \le K_{\mathrm{reg}}$, $\|\nabla_x^2 v_k(t,x)\|_\infty \le K_{\mathrm{reg}}$ for all x,k, it holds that $v_k(t,\cdot)$ is globally K_{reg}-Lipschitz and $\nabla_x v_k(t,\cdot)$ is globally K_{reg}-Lipschitz; both families are therefore uniformly bounded and equi-continuous on every compact set. Hence, the aforementioned claim can be directly shown by Arzelà–Ascoli Theorem.

Step 2. We claim that for each $t = H, H-1,\dots,0$, $V_{v_k}(t,\cdot) \to V_{v_\infty}(t,\cdot)$ locally uniformly on $\mathbb{R}^{d_x}$. We proceed by backward induction on t. For $t = H$, the claim holds trivially since both functions coincide with h. Now, assume the claim holds for $t+1$. Fix a closed ball $\overline{B}_m := \{x : \|x\|_2 \le m\}$, $m \ge 1$. For $x \in \overline{B}_m$, it holds that

$$\begin{aligned}V_{v_k}(t,x) - V_{v_\infty}(t,x) = &\underbrace{\big[\ell(x, v_k(t,x)) - \ell(x, v_\infty(t,x))\big]}_{\text{(I)}} \\ &+ \underbrace{\mathbb{E}_w\big[V_{v_k}(t{+}1, F(x,v_k(t,x)) + w) - V_{v_\infty}(t{+}1, F(x,v_\infty(t,x)) + w)\big]}_{\text{(II)}}.\end{aligned}$$

Term (I) tends to 0 uniformly on $\overline{B}_m$ since $\ell \in C^1$ and $v_k(t,\cdot) \to v_\infty(t,\cdot)$ uniformly on $\overline{B}_m$. For term (II), by the same token, we may conclude that $F(x,v_k(t,x)) \to F(x,v_\infty(t,x))$ uniformly on $\overline{B}_m$. Moreover, it follows that

$$\begin{aligned}|\text{(II)}| \le &\underbrace{\mathbb{E}_w\big[\big|V_{v_k}(t{+}1, F(x,v_\infty(t,x)) + w) - V_{v_\infty}(t{+}1, F(x,v_\infty(t,x)) + w)\big|\big]}_{\text{(III)}} \\ &+ \underbrace{\mathbb{E}_w\big[\big|V_{v_k}(t{+}1, F(x,v_k(t,x)) + w) - V_{v_k}(t{+}1, F(x,v_\infty(t,x)) + w)\big|\big]}_{\text{(IV)}}\end{aligned}$$

For notation simplicity, we write $y_t := F(x,v_\infty(t,x))$ and $y_{k,t} := F(x,v_k(t,x))$. We first estimate the upper bound of term (III). Since $x \in \overline{B}_m$ and $F, v_\infty(t,\cdot) \in C^1$, there exists a compact set $K_{y_t} \subseteq \mathbb{R}^{d_x}$ such that $y_t \in K_{y_t}$. For any $\epsilon > 0$, we can pick sufficiently large R such that $\mathbb{P}_w(\|w\|_2 > R) \le \frac{\epsilon}{8M}$. When $\|w\| \le R$, it holds that $V_{v_k}(t+1,\cdot) \to V_{v_\infty}(t+1,\cdot)$ uniformly on $K_{y_t} + B(\mathbf{0}_{d_x}, R)$ by the induction hypothesis. Hence, there exists N_1 such that for any $k > N_1$, it follows that

$$\sup_{y_t\in K_{y_t}, \|w\|_2\le R} |V_{v_k}(t+1, y_t + w) - V_{v_\infty}(t+1, y_t + w)| \le \epsilon/4.$$

When $\|w\| > R$, by Lemma 6, we have $|V_{v_k}|, |V_{v_\infty}| \le M$. Note that although v_∞ is only C^1-smooth, an analogous argument as in the proof of Lemma 6 yields that $|V_{v_\infty}| \le M$ as well. Thus,

$$\mathbb{E}_w\big[\big|V_{v_k}(t{+}1, y_t + w) - V_{v_\infty}(t{+}1, y_t + w)\big|\mathbf{1}\{|w| > R\}\big] \le 2M\mathbb{P}_w(|w| > R) \le \epsilon/4.$$

Hence, (III) $\le \epsilon/2$ for $k > N_1$. For term (IV), split the expectation into $\{\|w\|_2 \le R\}$ and its complement. On the first set, the locally uniform convergence in the induction hypothesis, the uniform continuity of $V_{v_\infty}(t+1,\cdot)$ on compact sets, and $y_{k,t} \to y_t$ uniformly on $\overline{B}_m$ imply that the integrand converges uniformly to zero. On the complement, it is bounded by $2M$. Thus, enlarging R if needed and then choosing $k > N_2$ gives (IV) $\le \epsilon/2$. For $k > \max\{N_1, N_2\}$, we obtain $|(\mathrm{II})| \le \epsilon$. This proves the claim by induction.

Step 3. Letting $\underline{p}(x) := \min_{\|z\|_2 \le K_0} p_\Sigma(x-z) > 0$, then, for all $v \in \mathcal{V}$, $t \ge 1$, we have $\rho_t^v(x) \ge \underline{p}(x)$. Fix a closed ball $\overline{B}_m := \{x : \|x\|_2 \le m\}$, $m \ge 1$, and let $\underline{p}_m := \min_{x \in \overline{B}_m} \underline{p}(x) > 0$. For $t \ge 1$, by (112),

$$\underline{p}_m^2 \int_{\overline{B}_m} \|v_k(t,x) - v_k^\diamond(t,x)\|_2^2 \mathrm{d}x \le \int_{\mathbb{R}^{d_x}} \rho_t^{v_k}(x)^2 \|v_k(t,x) - v_k^\diamond(t,x)\|_2^2 \mathrm{d}x \le \epsilon_k^2 \to 0,$$

so $v_k(t,\cdot) - v_k^\diamond(t,\cdot) \to 0$ in $L^2(\overline{B}_m)$. For $t = 0$, the initial state law is $\delta_{\bar{x}_0}$ and (112) directly gives pointwise convergence $v_k(0,\bar{x}_0) - v_k^\diamond(0,\bar{x}_0) \to 0$. By the proof of Lemma 6, sequences $\{v_k(t,\cdot)\}_{k\ge1}$ and $\{v_k^\diamond(t,\cdot)\}_{k\ge1}$ are uniformly Lipschitz on $\overline{B}_m$ with constants K_{reg} and Λ_1 respectively, independent of k; hence $\{g_k := v_k(t,\cdot) - v_k^\diamond(t,\cdot)\}_{k\ge1}$ is uniformly $(K_{\mathrm{reg}} + \Lambda_1)$-Lipschitz. By classical real analysis, if a family of uniformly Lipschitz functions converges to 0 in $L^2(\overline{B}_m)$, it follows that it converges uniformly to 0 on $\overline{B}_m$. Therefore, we have

$$\sup_{x \in \overline{B}_m} \|v_k(t,x) - v_k^\diamond(t,x)\|_2 \to 0, \quad t = 1, \dots, H-1. \tag{113}$$

Define $v_\infty^\diamond(t,x) := \arg\min_{a \in \mathcal{A}} Q_{v_\infty}(t,x,a)$. Step 2 and Gaussian convolution imply that Q_{v_k} converges locally uniformly in (x,a) to Q_{v_∞}, so the latter inherits μ_Q-strong convexity. By strong convexity and the first-order optimality conditions on $\mathcal{A}$, it follows that

$$\mu_Q \|v_k^\diamond(t,x) - v_\infty^\diamond(t,x)\|_2 \le \|\nabla_a Q_{v_\infty}(t,x,v_k^\diamond(t,x)) - \nabla_a Q_{v_k}(t,x,v_k^\diamond(t,x))\|_2. \tag{114}$$

Note that

$$Q_{v_k}(t,x,a) = \ell(x,a) + \int_{\mathbb{R}^{d_x}} V_{v_k}(t+1,z) p_\Sigma(z - F(x,a)) \mathrm{d}z.$$

Then, we have

$$\nabla_a Q_{v_k}(t,x,a) = \nabla_a \ell(x,a) + \mathbb{E}_w\Big[V_{v_k}(t+1, F(x,a)+w) D_a F(x,a)^\top \Sigma^{-1} w\Big].$$

Hence,

$$\begin{aligned}
&\text{RHS of (114)} \le K_0 \|\Sigma^{-1}\|_{\mathrm{op}} \, \mathbb{E}_w[\clubsuit], \\
&\clubsuit := \left|V_{v_k}(t+1, F(x,v_k^\diamond)+w) - V_{v_\infty}(t+1, F(x,v_k^\diamond)+w)\right| \|w\|_2,
\end{aligned}$$

where $v_k^\diamond$ is a shorthand for $v_k^\diamond(t,x)$. The inequality uses $\|D_a F(x,v_k^\diamond)\|_{\mathrm{op}} \le K_0$ from Assumption 8(a). Again, we may select a sufficiently large R and further do the following decomposition:

$$\mathbb{E}_w[\clubsuit] \le \underbrace{\mathbb{E}_w[\clubsuit \mathbf{1}\{\|w\| \le R\}]}_{:=\clubsuit_1} + \underbrace{\mathbb{E}_w[\clubsuit \mathbf{1}\{\|w\| > R\}]}_{:=\clubsuit_2}.$$

Regarding $\clubsuit_1$, invoking Step 2, the local uniform convergence $V_{v_k}(t+1,\cdot) \to V_{v_\infty}(t+1,\cdot)$ implies that $\clubsuit_1 \to 0$ uniformly on $\overline{B}_m$. For $\clubsuit_2$, Lemma 6 gives $\clubsuit \le 2M\|w\|_2$, so $\clubsuit_2 \le 2M\mathbb{E}_w[\|w\|_2 \mathbf{1}\{\|w\|_2 > R\}] \to 0$ as $R \to \infty$. Hence, we have shown that

$$\sup_{x \in \overline{B}_m} \|v_k^\diamond(t,x) - v_\infty^\diamond(t,x)\|_2 \to 0 \quad t = 1, \dots, H-1. \tag{115}$$

The same argument at $x = \bar{x}_0$ gives $v_k^\diamond(0,\bar{x}_0) \to v_\infty^\diamond(0,\bar{x}_0)$. Hence, for $t = 1, \dots, H-1$ and $x \in \overline{B}_m$, we have

$$\begin{aligned}
\|v_\infty(t,x) - v_\infty^\diamond(t,x)\|_2 \le &\|v_\infty(t,x) - v_k(t,x)\|_2 + \|v_k(t,x) - v_k^\diamond(t,x)\|_2 \\
&+ \|v_k^\diamond(t,x) - v_\infty^\diamond(t,x)\|_2.
\end{aligned}$$

The first term tends to 0 uniformly on $\overline{B}_m$ by Step 1; the second term by (113); the third term by (115). Letting $k \to \infty$ gives $v_\infty(t,x) = v_\infty^\diamond(t,x)$ for all $x \in \overline{B}_m$. since m was arbitrary and $\bigcup_{m\geq 1} \overline{B}_m = \mathbb{R}^{d_x}$, we have

$$v_\infty(t,x) = v_\infty^\diamond(t,x), \qquad v_\infty(0,\bar{x}_0) = v_\infty^\diamond(0,\bar{x}_0), \qquad t = 1,\dots,H-1, \quad x \in \mathbb{R}^{d_x}. \tag{116}$$

Step 4. By (116),

$$V_{v_\infty}(t,x) = Q_{v_\infty}(t,x,v_\infty(t,x)) = Q_{v_\infty}(t,x,v_\infty^\diamond(t,x)) = \min_{a\in\mathcal{A}} Q_{v_\infty}(t,x,a),$$

for $t = 1,\dots,H-1$ and all x, and at $(t,x) = (0,\bar{x}_0)$. Backward induction yields $V_{v_\infty}(t,\cdot) = V^*(t,\cdot)$ for $t = 1,\dots,H$ and $V_{v_\infty}(0,\bar{x}_0) = V^*(0,\bar{x}_0) = J[v^*]$. By Step 2 at $t = 0$, we have $J[v_k] = V_{v_k}(0,\bar{x}_0) \to V_{v_\infty}(0,\bar{x}_0) = J[v^*]$ as $k \to \infty$. But $\{J[v_k]\}_{k=0}^\infty$ is a subsequence of the convergent sequence $J[v^\tau] \to L$, so $J[v_k] \to L$ as well. Hence $L = J[v^*]$, i.e. $\eta = 0$, which contradicts $\eta > 0$.

Both cases force $\eta = 0$, contradicting the assumption $\eta > 0$. Hence $\eta = 0$, i.e.,

$$\lim_{\tau\to\infty} J[v^\tau] = J[v^*].$$

□

F Experimental Details

F.1 LQR problem

In this subsection, we introduce how A-KNOWN and A-SIEVE learn the stabilizing gain $\widehat{\mathbf{K}}$ in the LQR problem. For each LQR task, we realize the `LearnCertify` step of A-KNOWN via the regularized least square method discussed in Appendix B.1.1. Specifically, we set

$$\begin{aligned} &\sigma_{\exp} = 0.01, \qquad \lambda = 0.02, \qquad S = 2, \qquad \bar{\sigma} = 0.04, \qquad \delta_{\mathrm{conf}} = 0.05, \qquad \widetilde{N} = 400, \\ &\mathscr{K} = 8, \qquad \bar{v} = 10, \qquad \underline{v} = 0.1, \qquad \widetilde{\epsilon}_n = 0.001, \qquad n = 0,\dots,\widetilde{N}. \end{aligned} \tag{117}$$

Under the above parameters, we compute (g_n, r_n) via (22) at each time $n = 0,\dots,\widetilde{N}$. For the `CertifyGain` step of A-KNOWN, we realize it via the method for linear method described in Appendix B.1.2. Concretely, we first calculate the Lipschitz constant $L_{\Phi,n}$ of Φ_n by using (29) and the parameters in (117). Then, we set

$$\eta_n = \frac{\widetilde{\epsilon}_n}{2}, \qquad h_n = \frac{\widetilde{\epsilon}_n}{2(L_{\Phi,n} \vee 1)}, \qquad n = 0,\dots,\widetilde{N}.$$

For each time $n = 0,\dots,\widetilde{N}$, (K_n, P_n, U_n) can be calculated via (33) and (29) with the above parameters. This implements the `CertifyGain` step of A-KNOWN.

For A-SIEVE, we implement the `LearnCertify` step by the weighted ridge regression in Appendix B.1.1. Specifically, at each time $n = 0,\dots,\widetilde{N}$, we compute $(g_n, r_n, \widetilde{\epsilon}_n)$ via (24) and (27) with the following parameters:

$$\begin{aligned} &\lambda = 0.02, \quad \delta_{\mathrm{conf}} = 0.05, \quad \widetilde{N} = 200, \quad S_m = 2, \quad \bar{\sigma} = 0.04, \quad a_m = 0, \quad C_m = 5, \\ &\psi_m(z) = \mathrm{col}\left(z, \{\sin z_i\}_{i=1}^4, \chi(z_1), \chi(z_2)\right), \qquad m = 10. \end{aligned}$$

Here, $z = (x^\top, u^\top) \in \mathbb{R}^{d_x+d_u}$ and $\chi(s) := T_2(\sin s) - T_2(0) = 2\sin^2 s$, $s \in \mathbb{R}$. For the `CertifyGain` step of A-SIEVE, we implement it via the method for nonlinear system described in Appendix B.1.2. Concretely, we set

$$\begin{aligned} &\mathscr{K} = 8, \qquad \bar{v} = 10, \qquad \underline{v} = 0.1, \qquad \widetilde{\epsilon}_n = 0.001, \\ &B_n = \sqrt{12}\,\|\widehat{\Theta}_{m,n}\|_{\mathrm{F}}, \qquad H_n = 6\|\widehat{\Theta}_{m,n}\|_{\mathrm{op}}, \\ &a = \frac{\widetilde{\epsilon}_n}{4\sqrt{\kappa}(H_n \vee 1)(1+\mathscr{K}^2)}, \qquad R_n = \frac{4\sqrt{\kappa}(B_n \vee 1)}{\widetilde{\epsilon}_n}, \\ &\eta_n = \frac{\widetilde{\epsilon}_n}{4}, \qquad n = 0,\dots,\widetilde{N}. \end{aligned} \tag{118}$$

Here, the definition of $\widehat{\Theta}_{m,n}$ is in (24). Then, we calculate the Lipschitz constant $L_{\Phi,n}$ of Φ_n by (31) and (32) with the above parameters. Based on $L_{\Phi,n}$, we choose

$$h_n = \frac{\widetilde{\epsilon}_n}{8(L_{\Phi,n} \vee 1)}, \qquad n = 0, \dots, \widetilde{N}.$$

We calculate (K_n, P_n, U_n) via (33), (30), and interval branch-and-bound method (Neumaier, 2004) with h_n and the parameters in (118). This implements the `CertifyGain` step of A-SIEVE. Finally, both A-KNOWN and A-SIEVE use the true task dynamics to assess the stability of candidate feedback gains and determine stopping.

F.2 TORQUE-CONTROLLED INVERTED PENDULUM

We consider the following pendulum model:

$$\alpha_{t+1} = \alpha_t + 0.02\omega_t + \eta_{\alpha,t}, \tag{119}$$

$$\omega_{t+1} = \omega_t + 0.02\left(\frac{g}{\ell}\sin\alpha_t - \frac{b}{m\ell^2}\omega_t + \frac{1}{m\ell^2}u_t\right) + \eta_{\omega,t}, \tag{120}$$

where $x_t = (\alpha_t, \omega_t)^\top$ is the state, α_t is the pendulum angle measured from the upright position, ω_t is the angular velocity, and u_t is the torque. The gravity is $g = 9.81$. The mass m, length ℓ, damping coefficient b, and noise $w_t = (\eta_{\alpha,t}, \eta_{\omega,t})^\top$ satisfy

$$m \sim \mathrm{Unif}[0.8, 1.2], \quad \ell \sim \mathrm{Unif}[0.45, 0.55], \quad b \sim \mathrm{Unif}[0.04, 0.06],$$
$$\eta_{\alpha,t} \sim \mathcal{N}(0, 10^{-6}), \quad \eta_{\omega,t} \sim \mathcal{N}(0, \sigma_\omega^2), \quad \sigma_\omega \sim \mathrm{Unif}[0.005, 0.015], \quad t = 0, \dots, H-1$$

with $H = 100$. For the initial state $x_0 = (\alpha_0, \omega_0)$, we set $\alpha_0 \sim \mathrm{Unif}[-0.25, 0.25]$ and $\omega_0 \sim \mathrm{Unif}[-0.1, 0.1]$. Furthermore, we set the stage cost as $c(x_t, u_t) = 10\alpha_t^2 + \omega_t^2 + 0.05u_t^2, t = 0, \dots, H-1$ and the terminal cost as $h(x_H) \equiv 0$.

A-KNOWN sets $\phi^{(t)}(x, u) = [\sin\alpha, -\omega, u]^\top$. A-SIEVE sets $\phi^{(t)}$ as

$$\phi_0(x, u) = \mathrm{col}\left(1, \alpha, \omega, \sin\alpha, u, T_2(\omega), T_2(u), \omega u, \sin(2\alpha), \cos\alpha\right),$$
$$\phi^{(t)}(x, u) = \mathrm{col}\left(\phi_0(x, u), \left\{\frac{2^{-j}}{(j+2)^2}\sin((j+2)\alpha)\right\}_{j=1}^{m_t}\right),$$
$$m_t = \left\lfloor \log_2\left(1 + \frac{\lfloor t/(2M)\rfloor}{32}\right)\right\rfloor, \qquad d_t = 10 + m_t.$$

Here, $T_2(s) = 2s^2 - 1$ with $s \in \mathbb{R}$. For both algorithms, we use

$$f^{\mathrm{pub}}(z) = \begin{pmatrix}\alpha + 0.02\omega \\ \omega\end{pmatrix}, \qquad f(z) = f^{\mathrm{pub}}(z) + \Theta_m^\star \psi_m(z),$$

Both A-KNOWN and A-SIEVE implement `FiniteHorizonSolver` by using the cross-entropy method (CEM) (see Chua et al. (2018)). SPECIALIST is the adaptive nonlinear MPC method in Curato et al. (2017). Furthermore, we compute the oracle controller by using the iLQR method (see Tassa et al. (2012)).

We test the three algorithms and the oracle controller on 10 tasks with 20 trajectories per task. In each task, we first let A-KNOWN and A-SIEVE learn the stabilizing gain $\widehat{\mathbf{K}}$ by using the method shown in Appendix B.1.1 and B.1.2. For A-KNOWN and A-SIEVE, we implement the `LearnCertify` step and the `CertifyGain` step by the same methods as A-SIEVE for the LQR problem with some different settings of parameters and ψ_m. Specifically, for A-KNOWN, we set

$$M = 3, \quad e_t^{\mathrm{id}} \sim \mathcal{N}(0, 0.05^2), \quad \eta_t = 1, \quad B_{\mathrm{ctl}} = 5, \quad \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \quad t = 0, \dots, H-1,$$
$$\sigma_{\mathrm{exp}} = 0.01, \quad \lambda = 0.5, \quad \delta_{\mathrm{conf}} = 0.05, \quad \widetilde{N} = 200, \quad S_m = 0.5, \quad \bar{\sigma} = 0.015, \quad a_m = 0, \quad C_m = 1,$$
$$\psi_m(z) = [\sin\alpha, -\omega, u]^\top, \quad m = 3, \quad \mathscr{K} = 20, \quad \bar{v} = 10, \quad \underline{v} = 0.1, \quad \widetilde{\epsilon}_n = 0.001,$$
$$B_n = H_n = \|\widehat{\Theta}_{m,n}\|_{\mathrm{F}}, \qquad h_n = \frac{\widetilde{\epsilon}_n}{8(L_{\Phi,n} \vee 1)},$$

$$a = \frac{\widetilde{\epsilon}_n}{8\sqrt{\kappa}(H_n \vee 1)(1+\mathscr{K}^2)}, \qquad R_n = \frac{8\sqrt{\kappa}(B_n \vee 1)}{\widetilde{\epsilon}_n}, \qquad \eta_n = \frac{\widetilde{\epsilon}_n}{8}, \quad n = 0, \dots, \widetilde{N}.$$

with $z = (\alpha, \omega, u)^\top \in \mathbb{R}^3$. For A-SIEVE, we set

$$M = 3, \quad e_t^{\mathrm{id}} \sim \mathcal{N}(0, 0.05^2), \quad \eta_t = 1, \quad B_{\mathrm{ctl}} = 5, \quad \varepsilon_{\mathrm{cov},t} = (t+1)^{-1}, \quad t = 0, \dots, H-1,$$
$$\sigma_{\mathrm{exp}} = 0.01, \quad \lambda = 0.1, \quad \delta_{\mathrm{conf}} = 0.05, \quad \widetilde{N} = 400, \quad S_m = 1, \quad \bar{\sigma} = 0.015, \quad a_m = 0, \quad C_m = 5,$$
$$\chi(s) := T_2(\sin s) - T_2(0) = 2\sin^2 s, \qquad s \in \mathbb{R},$$
$$\psi_m(z) = \mathrm{col}\left(\alpha, \omega, u, \sin\alpha, \sin(2\alpha), \cos\alpha - 1, \chi(\omega), \chi(u), \sin\omega \sin u, \sin\alpha \sin\omega\right), \quad m = 10,$$
$$\mathscr{K} = 20, \quad \bar{v} = 10, \quad \underline{v} = 0.1, \quad \widetilde{\epsilon}_n = 0.001,$$

$$B_n = 5\|\widehat{\Theta}_{m,n}\|_{\mathrm{F}}, \qquad H_n = 8\|\widehat{\Theta}_{m,n}\|_{\mathrm{op}}, \qquad h_n = \frac{\widetilde{\epsilon}_n}{8(L_{\Phi,n} \vee 1)},$$
$$a = \frac{\widetilde{\epsilon}_n}{8\sqrt{\kappa}(H_n \vee 1)(1+\mathscr{K}^2)}, \qquad R_n = \frac{8\sqrt{\kappa}(B_n \vee 1)}{\widetilde{\epsilon}_n}, \qquad \eta_n = \frac{\widetilde{\epsilon}_n}{8}, \quad n = 0, \dots, \widetilde{N}.$$

Then, we restart A-KNOWN and A-SIEVE to the initial state and test their performances. The experimental results are shown below.

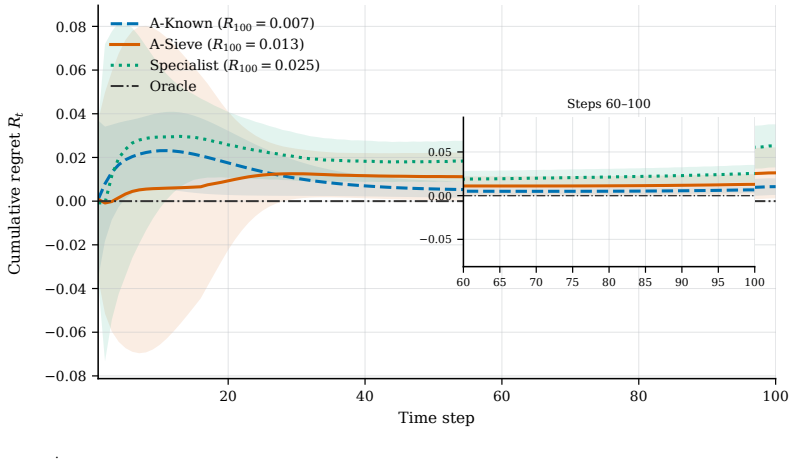


Figure 2: Pendulum cumulative regret of A-KNOWN, A-SIEVE, and SPECIALIST on 10 tasks with 10 trajectories per task.

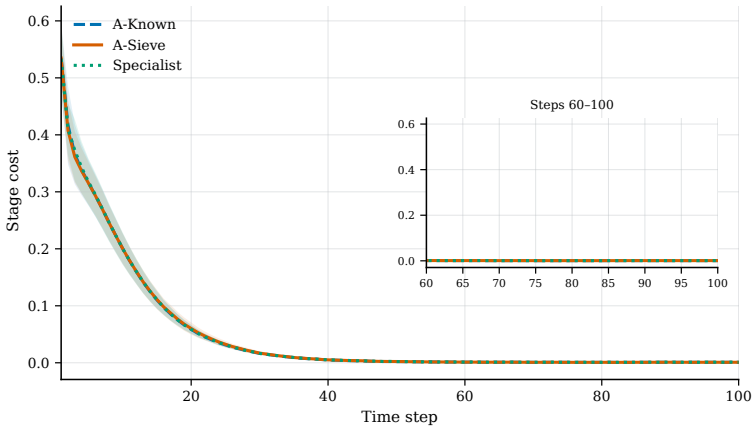


Figure 3: Pendulum stage costs of A-KNOWN, A-SIEVE, and SPECIALIST on 10 tasks with 10 trajectories per task.

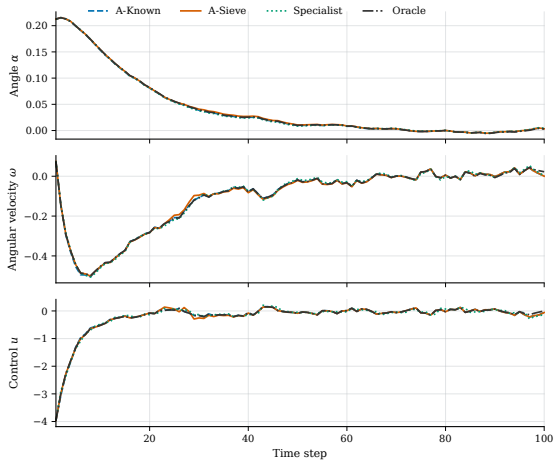


Figure 4: Representative pendulum state and control input trajectories of A-KNOWN, A-SIEVE, and SPECIALIST.

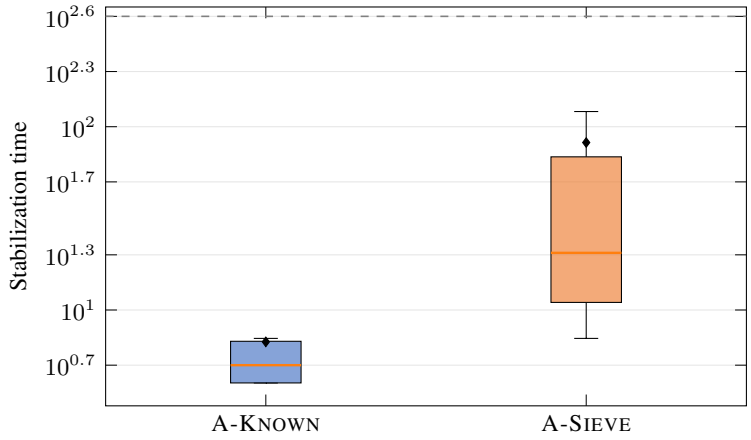


Figure 5: Box plot of the times that A-KNOWN and A-SIEVE use for learning a stabilizing controller.

Figures 2–4 compare the performances of the three algorithms in the cumulative regret, stage cost, and the state and control input trajectories. The final cumulative regrets of A-KNOWN, A-SIEVE, and SPECIALIST are approximately 0.007, 0.013, and 0.025, respectively. Meanwhile, note that the stage costs of the three algorithms decrease rapidly, and their state and control trajectories almost overlap with those of the oracle controller. These observations show that both A-SIEVE and A-KNOWN achieve the same level of performance as SPECIALIST and the oracle controller. Furthermore, Figure 5 shows that for each task, both A-SIEVE and A-Known can learn the stabilizing gain $\widehat{\mathbf{K}}$ in 150 steps.

F.3 DYNAMIC PRICING AND INVENTORY CONTROL

We consider the following pricing and inventory model:

$$D_t = \mu(p_t) + \epsilon_t, \qquad x_{t+1} = \mathrm{clip}(y_t - D_t, -20, 40). \tag{121}$$

Here, the state $x_t \in \mathbb{R}$ is the inventory level and the action $u_t = (p_t, y_t)^\top \in \mathbb{R}^2$ consists of the pricing decision p_t and the inventory ordered-up-to decision y_t. The demand noise satisfies $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We set the demand function μ as $\mu(p) = 22\alpha - 10\alpha p - 3\alpha p^2$ with $\alpha \sim \mathrm{Unif}[0.96, 1.04]$. We set $x_0 = 0$, $H = 100$, and impose the action constraint

$$u_t \in \mathcal{U}(x_t) := \{(p,y)^\top : 0 \le p \le 1,\ x_t \le y \le 40\}, \quad t = 0, \dots, H-1.$$

Define the stage and terminal costs by

$$c(x_t, u_t) = -p_t \mu(p_t) + 0.25 y_t + 10 \mathbf{1}\{y_t > 0\} + 0.05 (x_{t+1})_+ + (-x_{t+1})_+,\ h(x_H) = 0$$

with $(x_{t+1})_+ = x_{t+1} \vee 0$.

In this experiment, model (121) is a self-stable system since the state x_t is constrained to $[-20, 40]$. Hence, we skip the steps in Algorithm 1 for learning stabilizing gain $\widehat{\mathbf{K}}$ and directly set $\widehat{\mathbf{K}} = (0,0)^\top$ for A-KNOWN and A-SIEVE. We set the parameters of A-KNOWN and A-SIEVE as

$$M = 10, \quad e_t^{\text{id}} \sim \mathcal{N}(0, 0.05^2), \quad \eta_t = 1, \quad B_{\text{ctl}} = 5, \quad \varepsilon_{\text{cov},t} = (t+1)^{-1}, \quad t = 0, \dots, H-1.$$

A-KNOWN method uses $\phi^{(t)}(x,u) = (1, p, p^2)^\top$, whereas A-SIEVE uses the following $\phi^{(t)}$:

$$\phi^{(t)}(x,u) = \Big[1,\ \cos(2\pi p),\ \sin(2\pi p),\ \dots,\ \cos(2\pi K_t p),\ \sin(2\pi K_t p)\Big]^\top,$$
$$K_t = 1 + \left\lfloor \log_2 \left(1 + \left\lfloor \frac{t}{2M} \right\rfloor \right) \right\rfloor,$$

where p is the price component of u. Both A-Known and A-Sieve realize `FiniteHorizonSolve` via the method in Appendix B.1.3. SPECIALIST is the adaptive pricing method in Chen et al. (2022).

Furthermore, we compute the oracle controller via the finite-horizon dynamic programming with the true model coefficients and horizon $H = 100$ (see Gallego & Van Ryzin (1994)). We test the three algorithms and the oracle controller on 10 tasks with 20 trajectories per task. The experimental results are shown below.

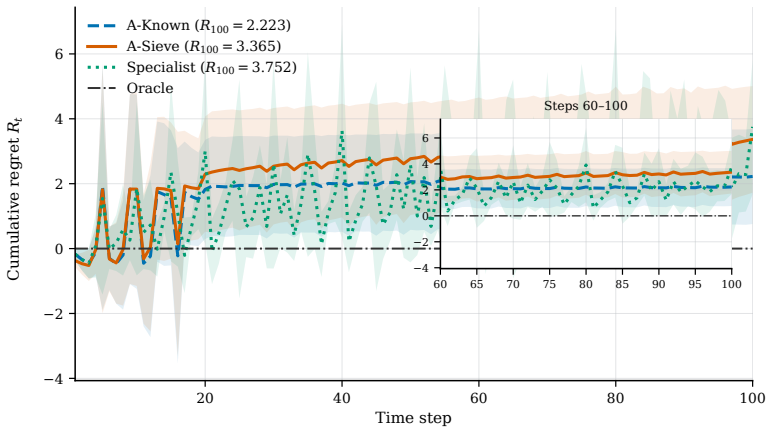


Figure 6: Pricing cumulative regret of A-KNOWN, A-SIEVE, and SPECIALIST on 10 tasks with 20 trajectories per task.

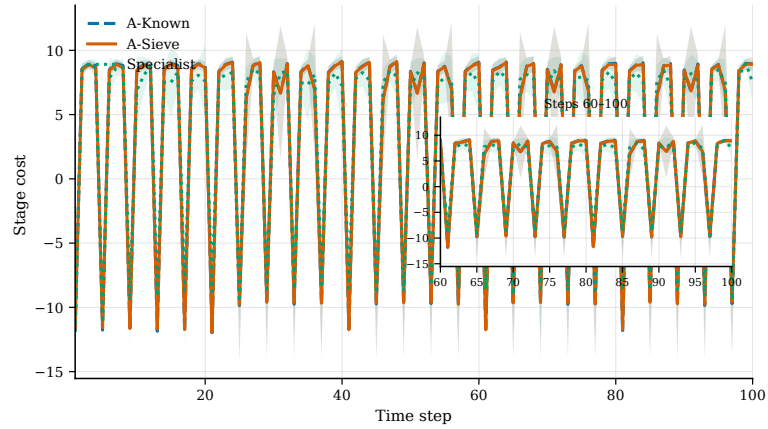


Figure 7: Pricing stage costs of A-KNOWN, A-SIEVE, and SPECIALIST on 10 tasks with 20 trajectories per task.

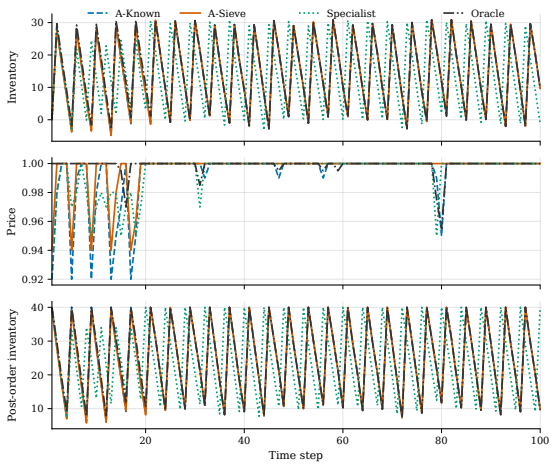


Figure 8: Representative pricing inventory, order-up-to level, and price trajectories of A-KNOWN, A-SIEVE, and SPECIALIST.

Figures 6–8 compare the performances of the three algorithms in the cumulative regret, stage cost, and representative state and control input trajectories. The final cumulative regrets of A-KNOWN, A-SIEVE, and SPECIALIST are approximately 2.223, 3.365, and 3.752, respectively. Thus both A-KNOWN and A-SIEVE have lower regrets than the one of Specialist in these tests. Meanwhile, the stage costs and (inventory, order-up-to level, and price) trajectories of A-SIEVE and A-KNOWN almost overlap with those of the oracle controller, whereas SPECIALIST has a remarkable difference from the oracle controller in the stage costs and trajectories. These observations show that A-KNOWN and A-SIEVE have better performance than SPECIALIST.