Regret Analysis of Multi-task Representation Learning for Linear-Quadratic Adaptive Control

Bruce D. Lee[*,1], Leonardo F. Toso[*,2], Thomas T.C.K. Zhang[*,1], James Anderson[2], and Nikolai Matni[1]

Abstract— **Representation learning is a powerful tool that enables learning over large multitudes of agents or domains by enforcing that all agents operate on a shared set of learned features. However, many robotics or controls applications that would benefit from collaboration operate in settings with changing environments and goals, whereas most guarantees for representation learning are stated for static settings. Toward rigorously establishing the benefit of representation learning in dynamic settings, we analyze the regret of multi-task representation learning for linear-quadratic control. This setting introduces unique challenges. Firstly, we must account for and balance the *misspecification* introduced by an approximate representation. Secondly, we cannot rely on the parameter update schemes of single-task online LQR, for which least-squares often suffices, and must devise a novel scheme to ensure sufficient improvement. We demonstrate that for settings where exploration is "benign", the regret of any agent after T timesteps scales as $\tilde{\mathcal{O}}(\sqrt{T/H})$, where H is the number of agents. In settings with "difficult" exploration, the regret scales as $\tilde{\mathcal{O}}(\sqrt{d_{\mathsf{U}} d_\theta}\sqrt{T} + T^{2/3}/\sqrt{H})$, where d_{X} is the state-space dimension, d_{U} is the input dimension, and d_θ is the task-specific parameter count. In both cases, by comparing to the minimax single-task regret $\mathcal{O}(\sqrt{d_{\mathsf{X}} d_{\mathsf{U}}^2}\sqrt{T})$, we see a benefit of a large number of agents. Notably, in the difficult exploration case, by sharing a representation across tasks, the effective task-specific parameter count can often be small $d_\theta < d_{\mathsf{X}} d_{\mathsf{U}}$. Lastly, we provide numerical validation of the trends we predict.**

I. Introduction

Many modern applications of robotics and controls involve simultaneous control over a large number of agents. For example, robot fleet learning, in which fleets of robots performing diverse tasks share information to learn more effectively, has demonstrated impressive success in recent years [1, 2]. One of the technologies that enables this success is *transfer learning*, in which dynamics models or control policies built upon learned compressed features (also known as *representation learning*) that are broadly useful for ensuing tasks of interest. Existing work which characterizes the generalization capabilities of transfer learning largely considers static environments, where data from an agent's completed task is aggregated with data from other agents to learn the shared features offline, rather than during task execution. However, it is often relevant to have a fleet of agents adapt quickly to a changing environment, e.g. a team of drones flying in close proximity adapting to weather conditions, or a team of legged robots adapting to changing terrain conditions. In such settings, the agents must communicate to adjust their shared features online.

In this work, we rigorously study such approaches for online fleet learning with dynamical systems in the analytically tractable setting of adaptive linear-quadratic (state-feedback) control. Adaptive linear-quadratic control has emerged as a benchmark for learning to control dynamical systems using online data. This consists of a learner interacting with an unknown linear system

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad t \geq 1, \tag{1}$$

with state x_t, input u_t, and noise w_t assuming values in $\mathbb{R}^{d_{\mathsf{X}}}$, $\mathbb{R}^{d_{\mathsf{U}}}$, and $\mathbb{R}^{d_{\mathsf{X}}}$, respectively. The learner is evaluated by its incurred *regret*, which compares the cost incurred by playing the learner for T time steps against the cost attained by the optimal LQR controller. Prior work typically studies regret of a single dynamical system of the form (1). In this work, we study a setting where there are $H \gg 1$ distinct systems which share an unknown d_θ-dimensional dynamics basis. Each agent aims to minimize their individual linear-quadratic control objective; however, by communicating they may more efficiently learn the shared dynamics basis matrices. The broad questions we address are the following:

- What are the requisite algorithmic elements that enable simultaneous online control of *multiple* systems?
- What are the concrete benefits of sharing a representation across agents compared with learning individual models for each agent?

Proofs can be found in the extended manuscript [3].

A. Related Work

Fleet Learning: Fleet Policy Learning considers a setting where a dataset is obtained from a diverse collection of robot interactions. It has been studied from the perspective of offline reinforcement learning [4] and for multi-task behavior cloning [1, 5, 6]. The centralized setup considered in this line of work is challenging to scale to many platforms. In particular, data communication and storage can become prohibitive, as can the training of the model. Frameworks have also proposed and analyzed a weight merging approach where each platform learns a policy, and then communicates the weights to a central server that merges the weights [2]. This work focuses on aggregating more skills by communicating, however the communication can also be used by multiple agents to adapt to a changing environment. This is the framework we analyze in this paper, where agents communicate their estimates for a set of shared parameters. This bears resemblance to certain federated or distributed

[*] alphabetical equal contribution

[1] Electrical and Systems Engineering, University of Pennsylvania. Emails: {brucele, ttz2, nmatni}@seas.upenn.edu

[2] Electrical Engineering, Columbia University. Emails: {lt2879, james.anderson}@columbia.edu

learning settings with heterogeneous data, where due to privacy or compute constraints agents do not centralize raw data [7–9].

Multi-Task Learning (in Dynamical Systems): Multi-task learning has long been studied in machine learning [10]. More recently, multiple works have studied the benefit of a shared representation in iid learning with regard to generalization [11, 12] and efficient algorithms [7, 13–15]. However, data generated from dynamical systems break key assumptions in these works. With respect to dynamical systems, multiple works consider a parallel setting where all agents share a parameter space and task-specialization comes from perturbations therein, see Model-agnostic meta-learning (MAML) [16].[1] Both model-free federated learning of the linear-quadratic regulator with data from heterogeneous systems [17] and MAML for linear-quadratic control have been considered [18]. However, both of these settings only recover optimality up to a heterogeneity bias. By instead imposing all dynamics matrices share a common basis [19], one can ensure the error decreases to zero as data increases. Analogous multi-task learning over dynamical systems settings have also been considered in imitation learning [20, 21]. Most relevant to our work is Zhang et al. [22], where the shortcomings of algorithms for iid representation learning are addressed for a related linear system-identification set-up. A component of our algorithm is adapted from their work.

Regret Analysis of Adaptive Control: Our setting and analysis builds off recent work that attempts to provide finite sample guarantees for adaptive control by controlling the *regret* of the learning algorithm. While adaptive control has a rich history beginning with autopilot development for high-performance aircraft in the 1950s [23], finite sample regret analysis of adaptive control arose much later [24]. Subsequent work [25–27] has introduced algorithms that yield $\sqrt{T}$ regret, and are computationally feasible. Simchowitz and Foster [28] establish corresponding lower bounds, indicating that a rate of $\sqrt{d_{\mathsf{U}}^2 d_{\mathsf{X}} T}$ is optimal for completely unknown systems. Improved regret bounds of $\texttt{poly}(\log T)$ are achievable when either $A^\star$ or $B^\star$ is known [29, 30]. The aforementioned work studies adaptive control in a setting where the noise is zero-mean and stochastic. Alternative formulations of the adaptive LQR problem consider bounded adversarial disturbances [31, 32] and settings where there is misspecification between the underlying data generating process and the model class [33, 34]. Our work extends analogous regret analysis to the multi-agent setting.

B. Contribution

We propose and analyze fleet linear-quadratic adaptive control in a setting where multiple linear systems driven by dynamics in the span of d_θ common basis matrices can communicate to drastically improve their individual control objectives. We propose such an algorithm and analyze the regret incurred, uncovering an interesting transition distinguishing the difficulty of the problem:

- When the system specific parameters are "benign" to identify, our proposed scheme incurs a regret of
$$R_T = \tilde{\mathcal{O}}\left(C_{\mathsf{sys}}\sqrt{\frac{T}{H}}\right),$$
where H is the number of communicating agents. When there are many agents, this is drastically lower than the regret $\mathcal{O}(\sqrt{d_{\mathsf{X}} d_{\mathsf{U}}^2 T})$ incurred if each agent had to learn to control its respective system without communication.
- When the system-specific parameters are challenging to identify, our proposed algorithm incurs a regret of at most
$$R_T = \tilde{\mathcal{O}}\left(\sqrt{d_{\mathsf{U}} d_\theta}\sqrt{T} + C_{\mathsf{sys}}\frac{T^{2/3}}{\sqrt{H}}\right).$$
When T is moderate, or if the number of agents H is large, this can demonstrate a marked gain over the single-agent setting. However, when T is large, the $T^{2/3}$ term dominates, which arises due to the mismatch between the difficulty of parameter identification and the misspecification of the learned basis directions.

In order to establish such guarantees, we propose and analyze a new algorithm that synthesizes tools from regret analysis of misspecified linear system identification and algorithmic analysis of multi-task linear regression. In particular, the multi-agent setting introduces unique challenges:

- Due to the approximate representation at any given timestep, the problem is misspecified. Therefore, in addition to the standard explore-commit tradeoff, we must account for improving the representation.
- Whereas for prior work in the stochastic single-agent setting least-squares–whose optimization and generalization is well-understood–suffices algorithmically, such an analog is not well-posed for the multiple agent setting.

We validate our theory with numerical simulations, and demonstrate the value of communicating with similar agents to learn to control more efficiently.

Notation: The Euclidean norm of a vector x is denoted $\|x\|$. For a matrix A, the spectral norm is denoted $\|A\|$, and the Frobenius norm is denoted $\|A\|_F$. The spectral radius of a square matrix is denoted $\rho(A)$. A symmetric, positive semi-definite (psd) matrix $A = A^\top$ is denoted $A \succeq 0$. The $\{\min, \max\}$ eigenvalue of a psd matrix A is denoted $\{\lambda_{\min}(A), \lambda_{\max}(A)\}$. For a positive definite matrix A, we denote the condition number as $\kappa(A) \triangleq \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$. We denote the normal distribution with mean μ and covariance Σ by $\mathcal{N}(\mu, \Sigma)$. For $f, g : D \to \mathbb{R}$, we write $f \lesssim g$ if for some $c > 0$, $f(x) \leq c g(x) \, \forall x \in D$. We denote the solutions to the discrete Lyapunov equation by $\texttt{dlyap}(A, Q)$ and the discrete algebraic Riccati equation by $\texttt{DARE}(A, B, Q, R)$. For an integer $n \in \mathbb{N}$, we define the shorthand $[n] \triangleq \{1, \ldots, n\}$. Generally, we use $\{\wedge, \vee\}$ to denote a $\{\min, \max\}$ over an indicated quantity.

[1] This is distinct from our setting, where agents share a *representation* function and task-specialization comes from linear functions of the representation.

II. Problem Formulation

A. System and Data assumptions

Consider H systems with dynamics defined by

$$x_{t+1}^{(h)} = A_\star^{(h)} x_t^{(h)} + B_\star^{(h)} u_t^{(h)} + w_t^{(h)}, \quad t \geq 1, \tag{2}$$

for $h \in [H]$. We suppose that each rollout starts from initial state $x_1^{(h)} = 0$ for $h \in [H]$, and that that the noise $w_t^{(h)}$ has iid elements that are mean zero and σ^2-sub-Gaussian for some $\sigma^2 \in \mathbb{R}$ with $\sigma^2 \geq 1$ [35]. We additionally assume that the noise has identity covariance: $\mathbf{E}\left[w_t^{(h)} w_t^{(h),\top}\right] = I$.[2] We suppose the dynamics matrices admit the decomposition

$$\begin{bmatrix} A_\star^{(k)} & B_\star^{(k)} \end{bmatrix} = \mathsf{vec}^{-1}\left(\Phi_\star \theta_\star^{(k)}\right), \tag{3}$$

where $\Phi_\star \in \mathbb{R}^{d_\mathsf{X}(d_\mathsf{X}+d_\mathsf{U}) \times d_\theta}$ is a column-orthonormal matrix that contains an optimal set of d_θ (vectorized) basis matrices in $\mathbb{R}^{d_\mathsf{X}(d_\mathsf{X}+d_\mathsf{U})}$, and $\theta_\star^{(k)} \in \mathbb{R}^{d_\theta}$ are agent-specific parameters. The operator vec^{-1} maps a vector in $\mathbb{R}^{d_\mathsf{X}(d_\mathsf{X}+d_\mathsf{U})}$ into a matrix in $\mathbb{R}^{d_\mathsf{X} \times (d_\mathsf{X}+d_\mathsf{U})}$ by stacking contiguous length-d_X blocks of a vector (top-to-bottom) into columns of a matrix (left-to-right). We can equivalently write this as a linear combination of basis matrices:

$$\begin{bmatrix} A_\star^{(k)} & B_\star^{(k)} \end{bmatrix} = \sum_{i=1}^{d_\theta} \theta_{\star,i}^{(k)} \begin{bmatrix} \Phi_{\star,i}^A & \Phi_{\star,i}^B \end{bmatrix},$$

where $\begin{bmatrix} \Phi_{\star,i}^A & \Phi_{\star,i}^B \end{bmatrix} = \mathsf{vec}^{-1} \Phi_{\star,i}$ and $\Phi_{\star,i}$ is the i^{th} column of $\Phi_\star$. This decomposition of the data generating process is a natural extension of the low-rank linear representations considered in [11, 20, 22] to the setting of multiple related dynamical systems with shared structure determined by $\Phi_\star$. A version of this model for autonomous systems was considered by [19] for multi-task system identification.

B. Control Objective

The goal of the learners is to interact with system (2) while keeping the total cumulative cost small, where the system specific cumulative cost for system h is defined for matrices $Q \succeq I$ and $R = I$ as[3]

$$C_T^{(h)} \triangleq \sum_{t=1}^{T} c_t^{(h)}, \text{ and } c_t^{(h)} \triangleq x_t^{(h),\top} Q x_t^{(h)} + u_t^{(h),\top} R u_t^{(h)}.$$

To define an algorithm that keeps the cost small, we first introduce the infinite horizon LQR cost:

$$\mathcal{J}^{(h)}(K) \triangleq \limsup_{T\to\infty} \frac{1}{T} \mathbf{E}^K C_T^{(h)}, \tag{4}$$

where the superscript K denotes evaluation under the state-feedback controller $u_t^{(h)} = K x_t^{(h)}$. To ensure that there exists a controller such that (4) is finite, we assume $(A_\star^{(h)}, B_\star^{(h)})$ is stabilizable for all $h \in [H]$. Under this assumption, (4) is minimized by the LQR controller $K_\infty(A_\star^{(h)}, B_\star^{(h)})$, where

$$K_\infty(A, B) \triangleq -(B^\top P_\infty(A,B) B + R)^{-1} B^\top P_\infty(A,B) A,$$

$$P_\infty(A, B) \triangleq \mathtt{DARE}(A, B, Q, R).$$

We define the shorthands $P_\star^{(h)} \triangleq P_\infty(A_\star^{(h)}, B_\star^{(h)})$ and $K_\star^{(h)} \triangleq K_\infty(A_\star^{(h)}, B_\star^{(h)})$ for all $h \in [H]$. To characterize the infinite-horizon LQR cost of an arbitrary stabilizing controller K, we additionally define the solution $P_K^{(h)}$ to the Lyapunov equation for the closed loop system under an arbitrary K where $\rho(A_\star^{(h)} + B_\star^{(h)} K) < 1$:

$$P_K^{(h)} \triangleq \mathtt{dlyap}(A_\star^{(h)} + B_\star^{(h)} K, Q + K^\top R K).$$

For a controller K satisfying $\rho(A_\star^{(h)} + B_\star^{(h)} K) < 1$, $\mathcal{J}^{(h)}(K) = \operatorname{tr}(P_K^{(h)})$. We have that $P_{K_\star^{(h)}}^{(h)} = P_\star^{(h)}$.

The infinite horizon LQR controller provides a baseline level of performance that our learner cannot surpass in the limit as $T \to \infty$. We quantify the performance of our learning algorithm by comparing the cumulative cost $C_T^{(h)}$ to the scaled infinite horizon cost attained by the LQR controller if the system matrices $\begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix}$ were known:

$$\mathcal{R}_T^{(h)} \triangleq C_T^{(h)} - T \mathcal{J}^{(h)}(K_\star^{(h)}). \tag{5}$$

This metric has previously been considered for adaptive control of a single system [24]. The above formulation casts the goal of the learner as interacting with each system (2) to maximize the information required for control while simultaneously regulating each system to minimize $\mathcal{R}_T^{(h)}$. The learner uses its history of interaction with each system to do so by constructing dynamics models, e.g. by determining estimates $\hat{A}^{(h)}$ and $\hat{B}^{(h)}$. It may then use these estimates as part of a *certainty equivalent* (CE) design by synthesizing controllers $\hat{K}^{(h)} = K_\infty(\hat{A}^{(h)}, \hat{B}^{(h)})$. It is known from prior work that if the model estimate is sufficiently close to the true dynamics, then the excess cost of playing the controller $\hat{K}^{(h)}$ is bounded by its parameter estimation error [27, 28].

Lemma II.1 (Theorem 3 of [28])**.** *Define* $\varepsilon^{(h)} \triangleq \frac{\|P_\star^{(h)}\|^{-5}}{3000}$. *If* $\left\| \begin{bmatrix} \hat{A}^{(h)} & \hat{B}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq \varepsilon^{(h)}$, *then*

$$\mathcal{J}^{(h)}(\hat{K}^{(h)}) - \mathcal{J}^{(h)}(K_\star^{(h)}) \leq$$
$$142 \left\| P_\star^{(h)} \right\|^8 \left\| \begin{bmatrix} \hat{A}^{(h)} & \hat{B}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2.$$

C. Algorithm Description

Our proposed algorithm, Algorithm 1, is a CE algorithm similar to those proposed by Cassel et al. [29], Lee et al. [34], which we extend to the multi-task representation learning setting. The algorithm takes a stabilizing controller $K_0^{(h)}$ for each system h as an input, in addition to an initial epoch length τ_1, an exploration sequence σ_k^2 for $k \in [k_{\mathsf{fin}}]$, state and controller bounds x_b and K_b, an initial representation estimate Φ_0, and a number of gradient steps N to run on the representation per epoch. Starting from the initial controllers, Algorithm 1 follows a doubling epoch approach. During each epoch, each agent plays their current controller with exploratory noise added with scale determined by the exploration sequence. Each agent then uses the collected data to estimate its dynamics $\begin{bmatrix} \hat{A}^{(h)} & \hat{B}^{(h)} \end{bmatrix}$ by running

[2] Noise that enters the process through a non-singular matrix H can be addressed by rescaling the dynamics by H^{-1}.

[3] Generalizing to arbitrary $Q \succ 0$ and $R \succ 0$ can be performed by scaling the cost and changing the input basis.

Algorithm 1 Shared-Representation Certainty-Equivalent Control with Continual Exploration

```
Input: Stabilizing controllers $K_0^{(h)}$ for $h \in [H]$, initial epoch length $\tau_1$, number of epochs $k_{\mathsf{fin}}$, exploration sequence $\sigma_1^2, \sigma_2^2, \sigma_3^2, \dots \sigma_{k_{\mathsf{fin}}}^2$, state bound $x_b$, controller bound $K_b$, initial representation estimate $\Phi_0$, gradient steps per epoch $N$
Initialize: $\hat{K}_1^{(h)} \leftarrow K_0^{(h)}$, $\tau_0 \leftarrow 0$, $T \leftarrow \tau_1 2^{k_{\mathsf{fin}}-1}$, $\hat{\Phi}_1 \leftarrow \Phi_0$.
for $k = 1, 2, \dots, k_{\mathsf{fin}}$ do
    for $h = 1, \dots, H$ (in parallel) do
        for $t = \tau_{k-1}, \tau_{k-1}+1, \dots, \tau_k$ do
            // Data collection
            if $\|x_t^{(h)}\|^2 \geq x_b^2 \log T$ or $\|\hat{K}_k^{(h)}\| \geq K_b$ then
                Abort and play $K_0^{(h)}$ forever
            Play $u_t^{(h)} = \hat{K}_k^{(h)} x_t^{(h)} + \sigma_k g_t^{(h)}$,
              where $g_t^{(h)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$
        // Task-wise parameter updates
        $\hat{\theta}_k^{(h)} \leftarrow \texttt{LS}(\hat{\Phi}_k, x^{(h)}_{\tau_{k-1}:\lceil\frac{3}{2}\tau_{k-1}\rceil}, u^{(h)}_{\tau_{k-1}:\lceil\frac{3}{2}\tau_{k-1}\rceil})$
        $\begin{bmatrix}\hat{A}_k^{(h)} & \hat{B}_k^{(h)}\end{bmatrix} \leftarrow \mathsf{vec}^{-1}\left(\hat{\Phi}_k \hat{\theta}_k^{(h)}\right)$
        $\hat{K}_{k+1}^{(h)} \leftarrow K_\infty(\hat{A}_k^{(h)}, \hat{B}_k^{(h)})$
    // Representation update
    $\hat{\Phi}_{k+1} \leftarrow \texttt{DFW}(\hat{\Phi}_k, x^{(1:H)}_{\lceil\frac{3}{2}\tau_{k-1}\rceil:\tau_k}, u^{(1:H)}_{\lceil\frac{3}{2}\tau_{k-1}\rceil:\tau_k}, N)$
    $\tau_{k+1} \leftarrow 2\tau_k$
```

Algorithm 2 Least squares: $\texttt{LS}(\hat{\Phi}, x_{1:t+1}, u_{1:t})$

```
Input: Model structure estimate $\hat{\Phi}$, state data $x_{1:t+1}$, input data $u_{1:t}$
Return: $\hat{\theta}$, where
    $\hat{\theta} = \Lambda^{\dagger}\left(\sum_{s=1}^{t} \hat{\Phi}^\top \left(\begin{bmatrix} x_s \\ u_s \end{bmatrix} \otimes I_{d_{\mathsf{X}}}\right) x_{s+1}\right)$   and
    $\Lambda = \sum_{s=1}^{t} \hat{\Phi}^\top \left(\begin{bmatrix} x_s \\ u_s \end{bmatrix}\begin{bmatrix} x_s \\ u_s \end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}.$
```

least-squares (Algorithm 2), fixing the current representation estimate $\hat{\Phi}$.[4] This is used to synthesize a new CE controller $\hat{K}^{(h)} = K_\infty(\hat{A}^{(h)}, \hat{B}^{(h)})$. At the end of each epoch, the agents engage in a round of N representation updates (Algorithm 3), in which they update their estimate for the shared basis using local data and communicate to take the average of their estimates. To analyze expected regret it is necessary to prevent catastrophic failures even under unlikely failure events. For this reason, the algorithm checks the state and controller norm against the supplied bounds x_b and K_b at the start of each interaction round, and aborts the CE scheme if either is too large.

A key subtlety and contribution of our algorithm comes in how the parameters are updated (Algorithm 1 and 3).

Algorithm 3 `De-bias & Feature Whiten`: $\texttt{DFW}(\hat{\Phi}, x_{1:t}^{(1:H)}, u_{1:t}^{(1:H)}, N)$

```
Input: Representation estimate $\hat{\Phi}$, state data $x_{1:t+1}^{(1:H)}$, input data $u_{1:t}^{(1:H)}$, gradient steps $N$, step-size $\eta$
Split the data into $2N$ equal length trajectories $D_1, \dots, D_{2N}$.
for $n = 1, \dots, N$ do
    for $h = 1, \dots, H$ in parallel do
        Compute weights
            $\hat{\theta}_n^{(h)} \leftarrow \texttt{LS}(\hat{\Phi}_n, \{x_s^{(h)}, u_s^{(h)}\}_{s \in D_{2n-1}})$.
        Compute local rep. update $\overline{\Phi}_n^{(h)}$ (6) on $D_{2n}$.
    Compute global rep. update
        $\hat{\Phi}^n, \_ \leftarrow \texttt{thin\_QR}(\frac{1}{H}\sum_{h=1}^{H} \bar{\Phi}_n^{(h)})$.
Return: $\hat{\Phi}_+ \leftarrow \hat{\Phi}_N$
```

In the single-agent setting, the optimal dynamics matrix $\begin{bmatrix}\hat{A} & \hat{B}\end{bmatrix}$ with respect to the current data batch follows by least squares, such that with a doubling epoch the parameter error approximately halves [28]. However, due to the multi-agent structure of our setting, least squares is no longer implementable, let alone optimal. This motivates the need for an alternative subroutine that ensures the representation error between epochs. Subroutines satisfying this are remote in the literature, especially since existing linear representation learning (or bilinear matrix sensing) algorithms heavily rely on the assumption that the data (or sensing matrix) across all tasks is iid isotropic Gaussian $x_i^{(h)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$ [7, 14, 15], which is violated in our setting where states distribution from different systems converge to their respective stationary distributions. A recent algorithm `De-bias & Feature Whiten (DFW)` proposed by Zhang et al. [36] addresses many analogous issues for a related multi-task representation learning problem, which we adapt for our setting. Beyond its guarantees (see Section II-D), `DFW` enables distributed optimization of a shared linear representation across data sources with *non-identical distributions*, and temporally dependent covariates. Additionally, `DFW` does not require communication of raw data between the agents, and instead each agent only communicates their respective updated representation, allowing the algorithm to be implemented in a federated manner. During each `DFW` iteration $n \in [N]$, each agent uses $\frac{1}{2N}^{\text{th}}$ of its data to estimate its local parameters via least-squares given the current representation $\hat{\Phi}_{n-1}$ (see Algorithm 2). Then, each agent uses the other $\frac{1}{2N}^{\text{th}}$ of its data to compute its *local* representation descent step:

$$\begin{aligned}
\nabla_{\Phi,n}^{(h)} &\triangleq \nabla_\Phi \sum_{t \in D_n} \left\| x_{t+1}^{(h)} - \mathsf{vec}^{-1}\left(\hat{\Phi}\hat{\theta}_n^{(h)}\right)\begin{bmatrix} x_t^{(h)} \\ u_t^{(h)} \end{bmatrix}\right\|^2 \\
\hat{\Sigma}_n^{(h)} &\triangleq \sum_{t \in D_n}\left(I_{d_{\mathsf{Y}}} \times \begin{bmatrix} x_t^{(h)} \\ u_t^{(h)} \end{bmatrix}\right)\left(I_{d_{\mathsf{Y}}} \times \begin{bmatrix} x_t^{(h)} \\ u_t^{(h)} \end{bmatrix}\right)^\top \qquad (6) \\
\overline{\Phi}_n^{(h)} &\leftarrow \hat{\Phi}_{n-1} - \eta\,(\hat{\Sigma}_n^{(h)})^{-1}\nabla_{\Phi,n}^{(h)}.
\end{aligned}$$

The updated local representations from each agent are then averaged and orthonormalized, and transmitted back to each

[4] This procedure throws away data from previous epochs, and does not allow updating the model at arbitrary times. This eases the analysis, but may be undesirable. Such undesirable characteristics have been removed in single task expected regret analysis [30].

agent for the next iteration (see Algorithm 3, line 7).

D. Representation Error Guarantees

In this section, we motivate the roles of our representation update (Algorithm 3) and task-specific weight update (Algorithm 2) subroutines. Consider current representation estimate $\hat{\Phi}$ and data $(x_{1:t}^{(1:H)}, u_{1:t}^{(1:H)})$ generated from initial states $x_1^{(1)}, \dots, x_1^{(H)}$, under stabilizing controllers $K^{(1)}, \dots, K^{(H)}$ with exploratory noise $\sigma_u g_s^{(h)}$, $g_s^{(h)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_{\mathsf{U}}})$ for $s \in [t]$, $h \in [H]$, and some $\sigma_u \in [0,1]$. This can be seen as the general set-up for the data collected during an epoch of Algorithm 1. We want to establish the following:

1) Running `DFW` yields an updated representation whose error decomposes as a contraction of the previous representation's error plus a variance term that scales inversely with the amount of *total data* tH.
2) The parameter error $\left\|\hat{\Phi}\hat{\theta}^{(h)} - \Phi_\star \theta_\star^{(h)}\right\|$ accrued by fitting the least-squares task-specific weights, holding the representation fixed, decomposes into a sum of least-squares error scaling inversely with t and the *representation error*.

These two guarantees together inform how to set the epoch length and exploratory noise strength σ_u to balance the explore-commit tradeoff for the ensuing regret analysis. To quantify the representation error, we consider the *subspace distance* between the spaces spanned by the columns of $\hat{\Phi}$ and $\Phi_\star$ (which are constrained to be column-orthonormal).

Definition II.1 (Stewart and Sun [37])**.** *For a given matrix with orthonormal columns Φ, let $\Phi_\perp$ be a matrix such that $\begin{bmatrix}\Phi & \Phi_\perp\end{bmatrix}$ is an orthogonal matrix. Then, given another column-orthonormal matrix Φ', the* subspace distance *between Φ', Φ may be written $d(\Phi, \Phi') \triangleq \|\Phi_\perp^\top \Phi'\|$.*

For all dimensions of $\Phi_\star$ to be identifiable, we also make the following full-rank assumption on the optimal weights $\theta_\star^{(1)}, \dots, \theta_\star^{(H)}$.

Assumption II.1. *Consider $\Phi_\star$, $\{\theta_\star^{(h)}\}$ such that $\mathsf{vec}^{-1}(\Phi_\star \theta_\star^{(h)}) = \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix}$, $h = 1, \dots, H$. We assume $\operatorname{rank}\left(\sum_{h=1}^H \theta_\star^{(h)} \theta_\star^{(h)\top}\right) = d_\theta$.*

We now state a bound on the improvement of the subspace distance after running Algorithm 3.

Theorem II.1 (`DFW` guarantee, informal)**.** *Let Assumption II.1 hold and fix $\delta \in (0,1)$. Then, provided an appropriately chosen step-size $\eta > 0$, $t \geq \tau_{\mathsf{dfw}}$, and $d(\hat{\Phi}, \Phi_\star) \leq d_{\mathsf{dfw}}$, with probability at least $1-\delta$ running Algorithm 3 yields the following guarantee on the updated representation $\hat{\Phi} \to \hat{\Phi}_N$:*

$$d(\hat{\Phi}_N, \Phi_\star) \leq \rho^N d(\hat{\Phi}, \Phi^\star) + \frac{\overline{K}_{\mathsf{avg}}}{1 - \sqrt{2}\rho^N} \frac{\sqrt{N}}{\sigma_u \sqrt{tH}},$$

where

$$\rho = \frac{1}{4}\kappa\left(\sum_{h=1}^H \theta_\star^{(h)} \theta_\star^{(h)\top}\right)^{-1}$$

$$\begin{aligned}\overline{K}_{\mathsf{avg}} = \sqrt{\frac{1}{H}\sum_{h=1}^H \sigma^2 \|\theta_\star^{(h)}\|^2 (1 + \|K^{(h)}\|^2 + \sigma_u^2)} \\ \cdot \operatorname{polylog}(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, H, 1/\delta).\end{aligned}$$

In particular, we have demonstrated that running `DFW` contracts the subspace distance by a factor of ρ^N, up to a variance factor. Notably, $\overline{K}_{\mathsf{avg}}$ serves as a task-averaged "noise-level", and the denominator of the variance factor scales *jointly* with the number of tasks H and data per task t. For downstream analysis, it suffices to choose a number of iterations N such that $\rho^N \leq 1/2$, i.e., $N \geq \log(2)/\log(1/\rho)$, and thus is independent of the size of the data. The subspace distance manifests in the error between the learned system parameters $\hat{\Phi}\hat{\theta}$ and the optimal $\Phi_\star \theta_\star$. In particular, given the output $\hat{\theta}$ of Algorithm 2, it can be shown (e.g. Theorem 5, [34]) that the parameter least squares error decomposes into a term scaling inversely with data and a term involving the subspace distance between $\hat{\Phi}$ and $\Phi_\star$.

Theorem II.2. *(LS error, informal) Consider running Algorithm 2 on the t data samples generated from a system of the form (2) for $t \geq \tau_{\mathsf{ls}}$, where τ_{ls} is a burn-in time. Then with probability at least $1 - \delta$,*

$$\left\|\hat{\Phi}\hat{\theta} - \Phi_\star \theta_\star\right\|^2 \lesssim \frac{\sigma^2 d_\theta \log(1/\delta)}{t \times \mathsf{excitation\ lvl}} + C_{\mathsf{sys}} \frac{d(\hat{\Phi}, \Phi_\star)^2}{\mathsf{excitation\ lvl}},$$

where C_{sys} is a constant that depends on the system (2), and $\mathsf{excitation\ lvl}$ *characterizes the extent to which the the state is excited as required to identify the parameters θ.*

Formal statements of Theorem II.1 and Theorem II.2 are instantiated in the ensuing regret analysis and can be found in the full paper [3]. We have thus established the desiderata stated at the beginning of the section. It remains to show that salient choices of epoch length and exploratory noise level in Algorithm 1 yield no-regret guarantees.

III. Regret Analysis

As previewed in the introduction, we consider two settings: one where the system-specific parameters $\theta_\star^{(h)}$ are easily identifiable given the representation, and one in which they are not. The setting where the system-specific parameters are easily identifiable corresponds to a situation in which $\mathsf{excitation\ lvl}$ from Theorem II.2 is nonzero even when the input is determined by the optimal LQR controller. In both settings, we require that the bounds for the abort procedure (Line 7, Algorithm 1) are sufficiently large to ensure that the abort procedure occurs with small probability. To state the bounds, we introduce the following notation.

$$\begin{aligned} \Psi_{B_\star^{(h)}} &\triangleq \max\left\{1, \left\|B_\star^{(h)}\right\|\right\}, & \Psi_B^\vee &\triangleq \max_{h=1,\dots,H} \Psi_{B_\star^{(h)}} \\ \theta^\vee &\triangleq \max_{h=1,\dots,H} \left\|\theta_\star^{(h)}\right\|, & P_0^\vee &\triangleq \max_{h=1,\dots,H} \left\|P_{K_0^{(h)}}^{(h)}\right\| \\ P_\star^\wedge &\triangleq \min_{h=1,\dots,H} \left\|P_{K_\star^{(h)}}^{(h)}\right\|, & \varepsilon^\wedge &\triangleq \min_{h=1,\dots,H} \varepsilon^{(h)}. \end{aligned}$$

Assumption III.1. *We assume that*

$$x_b \geq 400 (P_0^\vee)^2 \Psi_B^\vee \sigma \sqrt{d_{\mathsf{X}} + d_{\mathsf{U}}}, \quad K_b \geq \sqrt{P_0^\vee}.$$

A. Not Easily Identifiable

In this setting, we do not make additional assumptions about the structure of $\Phi_\star$. We require an assumption ensuring that it is possible to obtain a stabilizing CE controller after the first epoch with high probability. To do so, we make an assumption about the subspace distance of Φ_0 from $\Phi_\star$.

Assumption III.2. *Define*

$$\beta_1 \triangleq C_{\mathsf{bias},1} \sigma^4 (P_0^\vee)^{12} (\Psi_B^\vee)^8 (\theta^\vee)^2 (d_{\mathsf{X}} + d_{\mathsf{U}}) \sqrt{\frac{d_\theta}{d_{\mathsf{U}}}}$$

for a sufficiently large universal constant $C_{\mathsf{bias},1}$. We assume our representation error satisfies $d(\Phi_0, \Phi^\star) \leq \frac{\varepsilon^\wedge}{2\beta_1 \log T}$.

This assumption leads to the following regret bound.

Theorem III.1. *Consider applying Algorithm 1 with initial stabilizing controllers $K_0^{(1)}, \dots K_0^{(H)}$ for $T = \tau_1 2^{k_{\mathsf{fin}}-1}$ timesteps for some positive integers k_{fin}, and τ_1. Let $\tau_k = 2^k \tau_1$ for $k \in [k_{\mathsf{fin}}]$. Suppose that the exploration sequence supplied to the algorithm satisfies*

$$\sigma_k^2 = \max\left\{ \tau_k^{-1/3} H^{-1/2},\ \sqrt{\frac{d_\theta}{d_{\mathsf{U}} \tau_k}},\ \rho^{(k-1)N} d(\Phi_0, \Phi_\star) \right\} \tag{7}$$

for $k \in [k_{\mathsf{fin}}]$, where ρ is the contraction rate of Theorem II.1. Suppose the state bound x_b and the controller bound K_b satisfy Assumption III.1 and that Φ_0 satisfies Assumption III.2. Additionally suppose that the parameter N is sufficiently large that $\rho^N \leq \frac{1}{2}$ and that the weights satisfy Assumption II.1. There exists a polynomial function $\mathtt{poly}_{\mathsf{warm}}$ such that if $\tau_1 = \tau_{\mathsf{warm}} \log^2 T$ with

$$\tau_{\mathsf{warm}} \geq \mathtt{poly}_{\mathsf{warm}}(\sigma, P_0^\vee, \Psi_B^\vee, \theta^\vee, x_b, d_\theta, d_{\mathsf{X}}, d_{\mathsf{U}}, \log(H)),$$

then the expected regret satisfies for $h = 1, \dots, H$

$$\boldsymbol{E}\left[\mathcal{R}_T^{(h)}\right] \leq c_0 \log^2(T) + c_1 \sqrt{d_\theta d_{\mathsf{U}}} \sqrt{T} \log^2(T) + c_2 \frac{T^{2/3}}{\sqrt{H}} \log^2(HT),$$

where $c_0 = \mathtt{poly}\Big(\sigma, d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, x_b, K_b, \|Q\|, \theta^\vee, P_0^\vee, \Psi_B^\vee, \tau_{\mathsf{warm}}, x_b, d(\hat{\Phi}_0, \Phi_\star)\Big)$, $c_1 = \mathtt{poly}(P_0^\vee, \Psi_B^\vee, \sigma)$, and $c_2 = \mathtt{poly}\Big(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, P_0^\vee, \Psi_B^\vee, \theta^\vee, \sigma, N\Big)$.

Consider the above bound in the regime where T is small, e.g., on the order of the number of communicating agents. In this regime, the $T^{2/3}$ term becomes negligible, and the regret is dominated by the term that scales as $\sqrt{d_\theta d_{\mathsf{U}}} \sqrt{T}$. The should be contrasted with the minimax regret bound for single task adaptive control $\sqrt{d_{\mathsf{X}} d_{\mathsf{U}}^2 T}$ [28]: if the system-specific parameter count d_θ is smaller than $d_{\mathsf{X}} d_{\mathsf{U}}$, then the dominant term in the low data regime is smaller than the minimax regret of the single-task setting. In the adaptive control setting under consideration, the low data regime is often the one of interest, as we want the controller to rapidly adapt to a changing environment. However, it is remains an open question whether it is possible to achieve overall $\sqrt{T}$ regret in the multi-task learning setting. The following section examines one case where this is true.

B. Easily identifiable

In this setting, we assume that $\Phi^\star$ admits additional structure that makes the identification of $\theta_\star^{(h)}$ easy.

Assumption III.3. *Let $\alpha \geq \frac{1}{3(P_0^\vee)^{3/2}}$. Assume that $\min_{v:\|v\|=1} \left\| \sum_{i=1}^{d_\theta} v_i (\Phi_{\star,i}^A + \Phi_{\star,i}^B K) \right\|_F^2 \geq \alpha^2$ with $K \in \{K_0^{(h)}, K_\star^{(h)}\}$, $h \in [H]$, recalling $\begin{bmatrix} \Phi_{\star,i}^A & \Phi_{\star,i}^B \end{bmatrix} = \mathsf{vec}^{-1}\, \Phi_{\star,i}$, and $\Phi_{\star,i}$ is the i^{th} column of $\Phi_\star$.*

Under the above assumption, the weights θ are easily identifiable once the shared structure Φ is learned. As in the previous section, we require that the initial representation error is small enough to guarantee the closeness condition in Lemma II.1 may be satisfied with our estimated model after a single epoch.

Assumption III.4. *Define*

$$\beta_2 \triangleq C_{\mathsf{bias},2} \max_{h=1,\dots,H} \frac{\varepsilon^\wedge (P_0^\vee)^9 (\Psi_B^\vee)^8 (\theta^\vee)^2 (d_{\mathsf{X}} + d_{\mathsf{U}})}{d_\theta \min\{\alpha^2, \alpha^4\}}$$

for a sufficiently large universal constant $C_{\mathsf{bias},2}$. Our representation error satisfies $d(\Phi_0, \Phi^\star) \leq \sqrt{\frac{\varepsilon^\wedge}{2\beta_2}}$.

This allows us to state the following regret bound.

Theorem III.2. *Consider applying Algorithm 1 with initial stabilizing controller $K_0^{(1)}, \dots, K_0^{(H)}$ for $T = \tau_1 2^{k_{\mathsf{fin}}}$ timeteps for some positive integers k_{fin}, and τ_1. Let $\tau_k = 2^k \tau_1$ for $k \in [k_{\mathsf{fin}}]$ and suppose the exploration sequence is*

$$\sigma_k^2 = \max\left\{ \tau_k^{-1/2} H^{-1/2}, \rho^{(k-1)N} d(\Phi_0, \Phi_\star) \right\}, \tag{8}$$

for all $k \in [k_{\mathsf{fin}}]$, where ρ is the contraction rate of Theorem II.1. Suppose the state bound x_b and the controller bound K_b satisfy Assumption III.1, and that $\Phi_\star$ satisfies Assumption III.3 and Φ_0 satsisfies Assumption III.4. Additionally suppose that the parameter N is sufficiently large that $\rho^N \leq \frac{1}{2}$ and that the weights satisfy Assumption II.1. There exists a polynomial $\mathtt{poly}_{\mathsf{warm}}$ such that if $\tau_1 = \tau_{\mathsf{warm}} \log^2 T$ with

$$\tau_{\mathsf{warm}} \geq \mathtt{poly}_{\mathsf{warm}}\left(\sigma, P_0^\vee, \Psi_B^\vee, \theta^\vee, x_b, d_\theta, d_{\mathsf{X}}, d_{\mathsf{U}}, \log(H), \frac{1}{\alpha}\right),$$

then the expected regret satisfies for $h = 1, \dots, H$ satisfies

$$\boldsymbol{E}\left[\mathcal{R}_T^{(h)}\right] \leq c_1 \log^2(T) + c_2 \frac{\sqrt{T}}{\sqrt{H}} \log^2(TH),$$

where $c_1 = \mathtt{poly}\Big(\sigma, d_\theta, d_{\mathsf{U}}, d_{\mathsf{X}}, \frac{1}{\alpha}, \Psi_B^\vee, P_0^\vee, x_b, K_b, \theta^\vee, \|Q\|, \tau_{\mathsf{warm}}, d(\hat{\Phi}_0, \Phi_\star)\Big)$ and $c_2 = \mathtt{poly}\Big(\sigma, d_\theta, d_{\mathsf{U}}, d_{\mathsf{X}}, \frac{1}{\alpha}, \Psi_B^\vee, P_0^\vee, x_b, N\Big)$.

Consider once more the setting when the amount of data is on the order of the number of communicating agents. Here, the regret is dominated by a $\log T$ term. In particular, by sharing the "hard to learn" information, the communicating agents significantly simplify their respective adaptive control problems. Even in the regime of large T, the above regret bound improves upon what is possible in the single task setting as long as the number of agents is sufficiently large.

IV. Numerical Validation

We now present numerical results to illustrate and validate our bounds. In particular, we compare our proposed multi-task representation learning approach for the adaptive LQR design (Algorithm 1) over the setting where a single system attempts to learn its dynamics by using its local simulation data and computes a CE controller on top of the estimated model. To this end, our experimental setup considers H dynamical systems, described by (2), where the system matrices $(A_\star^{(h)}, B_\star^{(h)})$ are obtained by linearizing (around the origin) and discretizing (with Euler's approach) multiple cartpole dynamics with equations:

$$(M^{(h)} + m^{(h)})\ddot{x} + m^{(h)}\ell^{(h)}(\ddot{\theta}\cos(\theta) - \dot{\theta}^2\sin(\theta)) = u,$$
$$m^{(h)}(\ddot{x}\cos(\theta) + \ell^{(h)}\ddot{\theta} - g\sin(\theta)) = 0, \tag{9}$$

for all $h \in [H]$, where $c_{\mathrm{p}}^{(h)} = (M^{(h)}, m^{(h)}, l^{(h)})$ denote the tuple of cartpole parameters. Such parameters represent the cart mass, pole mass, and pole length, respectively. We set the gravity $g = 1$ and perform the discretization of (9) with step-size 0.25. Following [34], we generate H $(A_\star^{(h)}, B_\star^{(h)})$, by first considering a set of *nominal* cartpole parameters: $c_{\mathrm{p}}^{(1)} = (0.4, 1.0, 1.0)$, $c_{\mathrm{p}}^{(2)} = (1.6, 1.3, 0.3)$, $c_{\mathrm{p}}^{(3)} = (1.3, 0.7, 0.65)$, $c_{\mathrm{p}}^{(4)} = (0.2, 0.055, 1.36)$, and $c_{\mathrm{p}}^{(5)} = (0.2, 0.47, 1.825)$.

We then perturb such parameters with a random scalar within the interval $(0, 0.1)$ to generate different cartpole parameters $c_p^{(h)}$. With the system matrices $(A_\star^{(h)}, B_\star^{(h)})$ in hands, for all $h \in [H]$, we generate the disturbance signal as $w_t^{(h)} \sim \mathcal{N}(0, 0.01 I_{d_{\mathrm{X}}})$ and set the step-size and number of iterations of Algorithm 3 as $\eta = 0.25$, and $N = 1000$. It is worth noting that step 2 of Algorithm 3 is considered for the simplicity of the theoretical analysis only, in our experiments we exploit the entire dataset for all `DFW` iterations.

Figure 1 depicts the expected regret of Algorithm 1 as a function of the timesteps T for a varying number of tasks H. Note that such expected regret is with respect to a nominal task $h = 1$. This figure shows the results for the easily identifiable setting, i.e., where Assumption III.3 is satisfied. The labeled "fully-unknown" curve corresponds to the setting where a single system estimates its dynamics and computes its controller only using its own trajectory data. As predicted in our bounds (Theorem III.2), by learning the representation in a multi-task setting and exploiting it to learn a more accurate model can provide a significant reduction in the expected regret when compared to the fully-unknown case. In particular, the regret incurred in the single-task setting is in the order of $\mathcal{O}(\sqrt{T})$, whereas the regret of Algorithm 1 in the easily identifiable setting is dominated by $\mathcal{O}\left(\frac{\sqrt{T}}{\sqrt{H}}\right)$. Therefore, as the number of tasks H increases, the regret of Algorithm 1 decreases. This can be seen comparing the regret from $H = 25$ to $H = 100$–which both improve upon the regret in the fully-unknown setting.

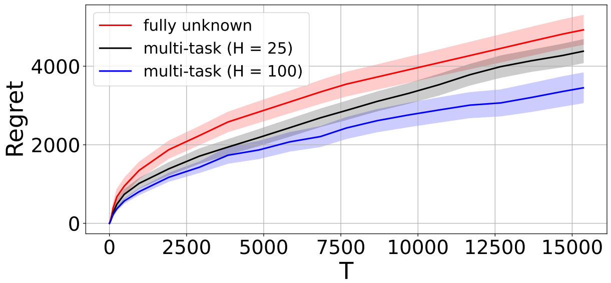

Fig. 1. Regret of Algorithm 1 with varying number of tasks H. We consider $k_{\mathrm{fin}} = 10$ epochs with initial epoch length $\tau_1 = 30$, an exploratory sequence scaling as $\sigma_k^2 \propto \frac{1}{\sqrt{2^k}}$, state and controller bounds $x_b = 25$, and $K_b = 15$, and random Φ_0 with $d(\Phi_0, \Phi_\star) \approx 0.99$.

V. Conclusion

We proposed an algorithm for the simultaneous adaptive control of multiple linear dynamical systems sharing a representation. We leveraged recent results for representation learning with non-iid data in order to provide non-asymptotic regret bounds incurred by the algorithm in two settings: one where the system specific parameters are easily identified from the shared representation, and one where they are not. In the setting where the system specific parameters are easily identifiable, the regret scales as $\sqrt{T}/\sqrt{H}$, while in the difficult-to-identify setting, the regret scales as $T^{2/3}/\sqrt{H}$. An interesting direction for future work is to determine whether the $T^{2/3}/\sqrt{H}$ regret bound can be improved to $\sqrt{T}/\sqrt{H}$ even in the difficult-to-identify setting. It would also be interesting to extend the analysis of online adaptive control with shared representations to characterize the regret of learning to control certain classes of nonlinear systems, as has been done in the single task setting [38].

References

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[2] L. Wang, K. Zhang, A. Zhou, M. Simchowitz, and R. Tedrake, "Fleet policy learning via weight merging and an application to robotic tool-use," *arXiv preprint arXiv:2310.01362*, 2023.

[3] B. Lee, L. F. Toso, T. Zhang, J. Anderson, and N. Matni, "Regret analysis of multi-task representation learning for linear-quadratic adaptive control," *preprint*, 2024. [Online]. Available: https://thomaszh3.github.io/research/multitask_sysid_regret.pdf

[4] A. Kumar, A. Singh, F. Ebert, M. Nakamoto, Y. Yang, C. Finn, and S. Levine, "Pre-training for robots: Offline rl enables learning new tasks from a handful of trials," *arXiv preprint arXiv:2210.05178*, 2022.

[5] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[6] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," *arXiv preprint arXiv:2306.14896*, 2023.

[7] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[8] X. Ma, J. Zhu, Z. Lin, S. Chen, and Y. Qin, "A state-of-the-art survey on solving non-iid data in federated learning," *Future Generation Computer Systems*, vol. 135, pp. 244–258, 2022.

[9] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8432–8440.

[10] J. Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.

[11] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," *arXiv preprint arXiv:2002.09434*, 2020.

[12] N. Tripuraneni, M. Jordan, and C. Jin, "On the theory of transfer learning: The importance of task diversity," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7852–7862, 2020.

[13] N. Vaswani, "Efficient federated low rank matrix recovery via alternating gd and minimization: A simple proof," *IEEE Transactions on Information Theory*, 2024.

[14] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, "Sample efficient linear meta-learning by alternating minimization," 2021.

[15] N. Tripuraneni, C. Jin, and M. Jordan, "Provable meta-learning of linear representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 434–10 443.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[17] H. Wang, L. F. Toso, A. Mitra, and J. Anderson, "Model-free learning with heterogeneous dynamical systems: A federated lqr approach," *arXiv preprint arXiv:2308.11743*, 2023.

[18] L. F. Toso, D. Zhan, J. Anderson, and H. Wang, "Meta-learning linear quadratic regulators: A policy gradient maml approach for the model-free lqr," *arXiv preprint arXiv:2401.14534*, 2024.

[19] A. Modi, M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Joint learning of linear time-invariant dynamical systems," *arXiv preprint arXiv:2112.10955*, 2021.

[20] T. T. Zhang, K. Kang, B. D. Lee, C. Tomlin, S. Levine, S. Tu, and N. Matni, "Multi-task imitation learning for linear dynamical systems," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 586–599.

[21] T. Guo, A. A. Al Makdah, V. Krishnan, and F. Pasqualetti, "Imitation and transfer learning for lqg control," *IEEE Control Systems Letters*, 2023.

[22] T. T. Zhang, L. F. Toso, J. Anderson, and N. Matni, "Sample-efficient linear representation learning from non-IID non-isotropic data," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Tr3fZocrI6

[23] P. Gregory, *Proceedings of the Self Adaptive Flight Control Systems Symposium*. Wright Air Development Center, Air Research and Development Command, United . . . , 1959, vol. 59, no. 49.

[24] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 1–26.

[25] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[26] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1300–1309.

[27] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[28] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[29] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1328–1337.

[30] Y. Jedra and A. Proutiere, "Minimal expected regret in linear quadratic control," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 234–10 321.

[31] E. Hazan, S. Kakade, and K. Singh, "The nonstochastic control problem," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 408–421.

[32] M. Simchowitz, K. Singh, and E. Hazan, "Improper learning for non-stochastic control," in *Conference on Learning Theory*. PMLR, 2020, pp. 3320–3436.

[33] U. Ghai, X. Chen, E. Hazan, and A. Megretski, "Robust online control with model misspecification," in *Learning for Dynamics and Control Conference*. PMLR, 2022, pp. 1163–1175.

[34] B. D. Lee, A. Rantzer, and N. Matni, "Nonasymptotic regret analysis of adaptive linear quadratic

control with model misspecification," *arXiv preprint arXiv:2401.00073*, 2023.

[35] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[36] T. T. Zhang, L. F. Toso, J. Anderson, and N. Matni, "Meta-learning operators to optimality from multi-task non-iid data," *arXiv preprint arXiv:2308.04428*, 2023.

[37] G. W. Stewart and J.-g. Sun, *Matrix perturbation theory*. Academic press, 1990.

[38] N. M. Boffi, S. Tu, and J.-J. E. Slotine, "Regret bounds for adaptive nonlinear control," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 471–483.

[39] K. B. Petersen, M. S. Pedersen *et al.*, "The matrix cookbook," *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[40] I. Ziemann, A. Tsiamis, B. Lee, Y. Jedra, N. Matni, and G. J. Pappas, "A tutorial on the non-asymptotic theory of system identification," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 8921–8939.

[41] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

VI. Outline for Proofs of Theorem III.1 and Theorem III.2

Our main results proceed by first defining a success events for which the certainty equivalent control scheme never aborts, and generates dynamics estimates $\begin{bmatrix}\hat{A}_k^{(h)} & \hat{B}_k^{(h)}\end{bmatrix}$ which are sufficiently close to the true dynamics $\begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}$ at all times. The success events are $\mathcal{E}_{\mathsf{success},1} = \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},1} \cap \mathcal{E}_{\mathsf{cont}}$ and $\mathcal{E}_{\mathsf{success},2} = \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},2} \cap \mathcal{E}_{\mathsf{cont}}$ for the settings where the task specific parameters are not easily identifiable and where they are, respectively. Here,

$$\mathcal{E}_{\mathsf{bound}} = \left\{\left\|x_t^{(h)}\right\|^2 \leq x_b^2 \log T \quad \forall t \in [T], \forall h \in [H]\right\} \cap \left\{\left\|\hat{K}_k^{(h)}\right\| \leq K_b \,, \forall k \in [k_{\mathsf{fin}}], \forall h \in [H]\right\},$$

$$\mathcal{E}_{\mathsf{est},1} = \left\{\left\|\begin{bmatrix}\hat{A}_k^{(h)} & \hat{B}_k^{(h)}\end{bmatrix} - \begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}\right\|_F^2 \leq C_{\mathsf{est},1} \frac{\sigma^2 d_\theta \left\|P_{K_0}^{(h)}\right\|}{\tau_k \sigma_k^2} \log(HT) + \frac{\beta_1 \log(HT) d(\hat{\Phi}_k, \Phi_\star)^2}{\sigma_k^2} \, \forall k \in [k_{\mathsf{fin}}], \forall h \in [H]\right\},$$

$$\mathcal{E}_{\mathsf{est},2} = \left\{\left\|\begin{bmatrix}\hat{A}_k^{(h)} & \hat{B}_k^{(h)}\end{bmatrix} - \begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}\right\|_F^2 \leq C_{\mathsf{est},2} \frac{\sigma^2 d_\theta}{\tau_k \alpha^2} \log(HT) + \beta_2 d(\hat{\Phi}_k, \Phi_\star)^2 \, \forall k \in [k_{\mathsf{fin}}], \forall h \in [H]\right\},$$

$$\mathcal{E}_{\mathsf{cont}} = \left\{d(\hat{\Phi}_k, \Phi_\star) \leq \rho^{kN} d(\hat{\Phi}_0, \Phi_\star) + \frac{C_{\mathsf{contract}} \sqrt{N} \log(HT)}{(1-\sqrt{2}\rho^N)\sqrt{H \tau_k \sigma_k^2}} \, \forall k \in [k_{\mathsf{fin}}]\right\},$$

and $C_{\mathsf{est},1}$ and $C_{\mathsf{est},2}$ are positive universal constants. We recall that

- x_b and K_b are the state and controller bounds triggering the abort procedure, see Assumption III.1.
- β_1 and β_2 are system theoretic constants defined in Assumption III.2 and Assumption III.4.
- k_{fin} is the total number of epochs run in Algorithm 1, and τ_k is the length of epoch k.
- α is the parameter defined in Assumption III.3 that quantifies the degree to which the initial and optimal controllers provide persistent excitation of the system specific parameters.
- σ_k^2 is the level of input exploration during epoch k.
- N is the number of descent steps run on the shared representation per epoch in Algorithm 3.
- ρ describes the radius of contraction for each iteration of Algorithm 3, while C_{contract} characterizes the numerator of the variance for each iteration; ρ is defined in Theorem II.1 and C_{contract} in Theorem VII.2.

With these events defined, the proofs for Theorem III.1 and Theorem III.2 consist of two steps:

1) In Section VIII we show that the success events $\mathcal{E}_{\mathsf{success},1}$ and $\mathcal{E}_{\mathsf{success},2}$ hold with high probability.
2) In Section IX, we decompose the expected regret into a component incurred under the success event and under the failure event. We show that the regret incurred under the failure event is small. The regret under the success event then dominates the overall regret, which is in turn bounded to obtain the expressions in Theorem III.1 and Theorem III.2.

Before doing so, we present formal versions of Theorem II.1 and Theorem II.2 in Section VII.

VII. Technical Preliminaries

To bound the probability of failure, we require two key components for our analysis: a high probability bound on the estimation error in terms of the level of misspecificiation, and a bound showing that the contraction event holds with high probability for any one epoch. The bound for the first step is provided in [34], and the bound on the second step is provided in [36]. We first describe the process characterizing the data collected during each epoch.

Consider a general estimation problem in which the system is excited by an arbitrary stabilizing controller K and excitation level defined by σ_u. In particular, we consider the evolution of the following system:

$$\begin{aligned} x_{t+1} &= A^\star x_t + B^\star u_t + w_t \\ u_t &= K x_t + \sigma_u g_t, \end{aligned} \tag{10}$$

where $g_t \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$, and w_t is a random variable with σ^2-sub-Gaussian entries satifying $\mathbf{E}[w_t w_t^\top] = I$. We assume that $\sigma_u^2 \leq 1$ and that x_1 is a random variable.

A. Least squares error

We first consider generating the estimates $\hat{\theta}, \Lambda = \mathtt{LS}(\hat{\Phi}, x_{1:t+1}, u_{1:t})$. We present a bound on the estimation error $\left\|\hat{\Phi}\hat{\theta} - \Phi_\star \theta^\star\right\|^2$ in terms of the true system parameters as well as the amount of data, t.

Theorem VII.1 (Misspecified LS Est. Error - Formal Version of Theorem II.2, Theorem 5 of [34]). *Let $\delta \in (0, 1/2)$. Suppose $t \geq c\tau_{\mathsf{ls}}(K, \left\|x_1\right\|^2, \delta)$ for*

$$\tau_{\mathsf{ls}}(K, \bar{x}, \delta) \triangleq \max\left\{\sigma^4 \left\|P_K\right\|^3 \Psi_{B^\star}^2 \left(d_{\mathsf{X}} + d_{\mathsf{U}} + \log\frac{1}{\delta}\right), \bar{x} \times \left\|P_K\right\| + 1\right\}$$

and a sufficiently large universal constant $c > 0$. There exists an event $\mathcal{E}_{\mathsf{ls}}$ which holds with probability at least $1-\delta$ under which the estimation error satisfies

$$\left\|\hat{\Phi}\hat{\theta} - \Phi_\star\theta^\star\right\|^2 \lesssim \frac{d_\theta\sigma^2}{t\lambda_{\min}\left(\bar{\Delta}^t(\sigma_u,K)\right)}\log\left(\frac{1}{\delta}\right) + \left(1 + \frac{\sigma^4\left\|P_K\right\|^7\Psi_{B^\star}^6\left(d_{\mathsf{X}} + d_{\mathsf{U}} + \log\frac{1}{\delta}\right)}{t\lambda_{\min}(\bar{\Delta}^t(\sigma_u,K))^2}\right)\frac{\left\|P_K\right\|^2\Psi_{B^\star}^2 d(\hat{\Phi},\Phi_\star)^2\left\|\theta^\star\right\|^2}{\lambda_{\min}(\bar{\Delta}^t(\sigma_u,K))}.$$

where

$$\bar{\Delta}^t(\sigma_u,K) \triangleq \hat{\Phi}^\top\left(\frac{1}{t}\sum_{s=0}^{t-2}\sum_{j=0}^{s}\begin{bmatrix} I \\ K\end{bmatrix} A_K^j(\sigma_u^2 B^\star(B^\star)^\top + I)\left(A_K^j\right)^\top\begin{bmatrix} I \\ K\end{bmatrix}^\top + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_u^2 I_{d_{\mathsf{U}}}\end{bmatrix}\right)\hat{\Phi}.$$

B. Representation Learning Guarantees from `DFW`

We now want to prove that applying `DFW` leads to the high-probability contraction guarantee previewed in Theorem II.1. Analogous to the original analysis provided in [36], to this end, we consider a general realizable regression setting, i.e. for each task h, the labels are generated by a ground truth mechanism

$$y_i^{(h)} = \mathsf{vec}^{-1}(\Phi_\star\theta_\star^{(h)})x_i^{(h)} + w_i^{(h)},$$

where $y_i^{(h)} \in \mathbb{R}^{d_{\mathsf{Y}}}$, $x_i^{(h)} \in \mathbb{R}^{d_{\mathsf{X}}}$. Note that our sysID setting simply follows by setting $y_i^{(h)} \triangleq x_{i+1}^{(h)}$, $x_i^{(h)} \triangleq \begin{bmatrix} x_i^{(h)\top} & u_i^{(h)\top}\end{bmatrix}$. For a given task h and current representation $\hat{\Phi}$ and task-specific weights $\hat{\theta}^{(h)}$, the representation gradient with respect to a given batch of data $\{(x_i^{(h)}, y_i^{(h)})\}_{i=1}^N$ can be expressed as [39]

$$\begin{aligned}
\nabla_\Phi^{(h)} &\triangleq \nabla_\Phi \frac{1}{2N}\sum_{i=1}^N\left\|y_i^{(h)} - \mathsf{vec}^{-1}(\hat{\Phi}\hat{\theta}^{(h)})x_i^{(h)}\right\|_2^2 \\
&= \frac{1}{N}\sum_{i=1}^N\left(X_i^{(h)\top}X_i^{(h)}\hat{\Phi}\hat{\theta}^{(h)}\hat{\theta}^{(h)\top} - X_i^{(h)\top}y_i^{(h)}\hat{\theta}^{(h)\top}\right) \\
&= \frac{1}{N}\sum_{i=1}^N X_i^{(h)\top}X_i^{(h)}\left(\hat{\Phi}\hat{\theta}^{(h)} - \Phi_\star\theta_\star^{(h)}\right)\hat{\theta}^{(h)\top} - \frac{1}{N}\sum_{i=1}^N X_i^{(h)\top}w_i^{(h)}\hat{\theta}^{(h)\top},
\end{aligned}$$

where $X_i^{(h)} \triangleq I_{d_{\mathsf{Y}}} \otimes x_i^{(h)\top}$. Recalling the definition of the orthogonal complement matrix $\Phi_{\star,\perp}$ and the subspace distance (Definition II.1), we note that $\left\|\Phi_{\star,\perp}^\top\hat{\Phi}\right\|$ is the subspace distance between $\Phi_\star$ and $\hat{\Phi}$, and $\Phi_{\star,\perp}^\top\Phi_\star = 0$. Noting these identities, the key insight in `DFW` [36] is to pre-multiply the representation gradient $\nabla_\Phi^{(h)}$ by the inverse sample-covariance $\left(\frac{1}{N}\sum X_i^{(h)\top}X_i^{(h)}\right)^{-1}$,

$$\begin{aligned}
\tilde{\nabla}_\Phi^{(h)} &\triangleq \left(\frac{1}{N}\sum X_i^{(h)\top}X_i^{(h)}\right)^{-1}\nabla_\Phi^{(h)} \\
&= \left(\hat{\Phi}\hat{\theta}^{(h)} - \Phi_\star\theta_\star^{(h)}\right)\hat{\theta}^{(h)\top} - \left(\sum_{i=1}^N X_i^{(h)\top}X_i^{(h)}\right)^{-1}\sum_{i=1}^N X_i^{(h)\top}w_i^{(h)}\hat{\theta}^{(h)\top}.
\end{aligned} \tag{11}$$

Therefore, performing a descent step with the adjusted gradient $\tilde{\nabla}_\Phi^{(h)}$ and averaging the resulting updated representations across tasks h yields

$$\overline{\Phi}_+ = \frac{1}{H}\sum_{h=1}^H\left(\hat{\Phi} - \eta\tilde{\nabla}_\Phi^{(h)}\right) \tag{12}$$

$$= \hat{\Phi}\left(I - \frac{\eta}{H}\sum_{h=1}^H\hat{\theta}^{(h)}\hat{\theta}^{(h)\top}\right) + \Phi_\star\left(\frac{\eta}{H}\sum_{h=1}^H\theta_\star^{(h)}\hat{\theta}^{(h)\top}\right) + \frac{\eta}{H}\sum_{h=1}^H\left(\sum_{i=1}^N X_i^{(h)\top}X_i^{(h)}\right)^{-1}\sum_{i=1}^N X_i^{(h)\top}w_i^{(h)}\hat{\theta}^{(h)\top}. \tag{13}$$

Pulling out the orthonormalization factor and left-multiplying the above by $\Phi_{\star,\perp}^\top$ yields

$$\Phi_{\star,\perp}^\top\hat{\Phi}_+R = \Phi_{\star,\perp}^\top\hat{\Phi}\left(I - \frac{\eta}{H}\sum_{h=1}^H\hat{\theta}^{(h)}\hat{\theta}^{(h)\top}\right) + \Phi_{\star,\perp}\frac{\eta}{H}\sum_{h=1}^H\left(\sum_{i=1}^N X_i^{(h)\top}X_i^{(h)}\right)^{-1}\sum_{i=1}^N X_i^{(h)\top}w_i^{(h)}\hat{\theta}^{(h)\top}.$$

Thus, as long as the orthonormalization factor R is sufficiently well-conditioned, by taking the spectral norm on both sides of the above, we get the following decomposition

$$d(\hat{\Phi}_+, \Phi_\star) \leq d(\hat{\Phi}, \Phi_\star) \left\| I - \frac{\eta}{H} \sum_{h=1}^{H} \hat{\theta}^{(h)} \hat{\theta}^{(h)\top} \right\| \left\| R^{-1} \right\| + \left\| \frac{\eta}{H} \sum_{h=1}^{H} (\mathbf{X}^{(h)} \mathbf{X}^{(h)\top})^{-1} \mathbf{X}^{(h)} \mathbf{W}^{(h)\top} \hat{\theta}^{(h)\top} \right\| \left\| R^{-1} \right\| \tag{14}$$

As proposed in [36], bounding the improvement from $\hat{\Phi}$ to $\hat{\Phi}_+$ essentially reduces to establishing that $\left\| I - \frac{\eta}{H} \sum_{h=1}^{H} \hat{\theta}^{(h)} \hat{\theta}^{(h)\top} \right\|$ is a contraction with high-probability and analyzing the noise term $\frac{1}{H} \sum_{h=1}^{H} (\mathbf{X}^{(h)} \mathbf{X}^{(h)\top})^{-1} \mathbf{X}^{(h)} \mathbf{W}^{(h)\top}$ as an *average of self-normalized martingales*. The following largely adapts the analysis from [36], with minor alterations due to the slightly modified setting.

Contraction and Orthonormalization Factor

As aforementioned, bounding the "contraction rate" of the representation toward optimality amounts to bounding

$$\left\| I - \frac{\eta}{H} \sum_{h=1}^{H} \hat{\theta}^{(h)} \hat{\theta}^{(h)\top} \right\| \left\| R^{-1} \right\|.$$

The proof of this largely follows from the analysis in [36], by lower bounding the minimum eigenvalue of $\sum_{h=1}^{H} \hat{\theta}^{(h)} \hat{\theta}^{(h)\top}$, as in section A.2 of [36], and by $\left\| R^{-1} \right\|$ by $1/\left(1 - \frac{4c\eta}{H} \sum_{h=1}^{H} \theta_\star^{(h)} \theta_\star^{(h)\top}\right)$ for a universal constant $c > 0$. Substituting the value of this c, and apprpriately defining the step size results in the contraction.

Noise Term

We now consider bounding the noise term in (14):

$$\left\| \frac{\eta}{H} \sum_{h=1}^{H} (\mathbf{X}^{(h)} \mathbf{X}^{(h)\top})^{-1} \mathbf{X}^{(h)} \mathbf{W}^{(h)\top} \hat{\theta}^{(h)\top} \right\| \left\| R^{-1} \right\|.$$

This also largely follows from analysis in [36]; however, we note that the shape of the representation with respect to the input and output dimensions introduces a couple discrepancies we must address. In [36] and other related work, the representation enters as: $y = F_\star \Phi_\star x + w$, where x is a (stationary) subgaussian process, and w is a conditionally-subgaussian zero-mean noise process, whereas our setting includes a vectorization operator $y = \mathsf{vec}^{-1}(\Phi_\star \theta_\star) x + w$, making analogous assumptions on x, w. The analysis in [36] relies on the *self-normalized martingale* toolbox [24, 40], which provide tools to control the process

$$\sum_{i=1}^{N} w_i x_i \left(\sum_{i=1}^{N} x_i x_i^\top \right)^{-1}.$$

In our setting, $\mathbf{X}^{(h)}$ is determined by the Kronecker product with identity. In light of this, the corresponding step of the proof may be replaced with application of Theorem 7 of Lee et al. [34].

We now present a bound showing the improvement of the subspace distance $d\left(\hat{\Phi}_N, \Phi_\star\right)$ after running Algorithm 3.

We want to show that `DFW` (Algorithm 3) contracts the subspace distance between the representation estimate $\hat{\Phi}$ and the optimal basis $\Phi_\star$. Before proceeding to representation error guarantee, we state the burn-in sample requirements for running `DFW`, which recur for our ensuing regret analysis.

Definition VII.1. *For given stabilizing controllers K^h, $h \in [H]$, i.e. $\rho(A_\star^{(h)} + B_\star^{(h)} K^{(h)}) < 1$, define $\Gamma_K^\vee > 0$ and $\mu_K^\vee \in (0,1)$ as constants such that for all $h \in [H]$, $\|(A_\star^{(h)} + B_\star^{(h)} K^{(h)})^t\| \leq \Gamma_K^\vee (\mu_K^\vee)^t$ for any $t \geq 0$.*[5]

Assumption VII.1 (`DFW` burn-in, redux)**.** *Consider running Algorithm 3 on data generated from arbitrary initial states $x_1^{(1)}, \ldots, x_1^{(H)}$, norm-bounded by $x_b \sqrt{\log(T)}$ (see Line 7), by closed loop systems under stabilizing controllers $K^{(1)}, \ldots, K^{(H)}$ with exploratory noise $\sigma_u g_t$, $g_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_{\mathsf{U}}})$, and representation $\hat{\Phi}$. For a given failure probability $\delta \in (0,1)$, let the following hold on the epoch length and systems $h \in [H]$:*

$$\operatorname{rank}\left(\sum_{h=1}^{H} \theta_\star^{(h)} \theta_\star^{(h)\top} \right) = d_\theta$$

[5] Such constants are guaranteed to exist by, e.g. Gelfand's Formula [41].

$$
\begin{aligned}
t/\tau_{\mathsf{mix}} \gtrsim \max_h \Big\{ & \sigma^4 \|P_{K^{(h)}}\|^3 \|\psi_{B_\star^{(h)}}\|^2 (d_\theta + \log(H/\delta)), \\
& \frac{\sigma^2 \sigma_u^2}{\|\theta_\star^{(h)}\|^2 \big(1 + \|K^{(h)}\|^2 + \sigma_u^2\big)} (d_{\mathsf{X}} + d_\theta + \log(H/\delta)) \Big\} \\
d(\hat{\Phi}, \Phi_\star) \le \frac{4}{25} & \min_h \kappa\Big(\Sigma^t(K^{(h)}, \sigma_u, x_1^{(h)})\Big)^{-1} \\
& \cdot \kappa(\sum_{h=1}^{H} \theta_\star^{(h)} \theta_\star^{(h)\top})^{-1},
\end{aligned}
$$

where $\tau_{\mathsf{mix}} \triangleq \frac{1}{1-\mu_K^\vee} \log\Big(\frac{t\Gamma_K^\vee}{\delta}\sqrt{x_b^2 \log(T) + \frac{d_{\mathsf{X}}}{1-(\mu_K^\vee)^2}}\Big)$ *and* $\Sigma^t(K, \sigma_u, x_1)$ *denotes the* t*-horizon population covariance matrix of* (x, u) *initialized at* x_1 *and under feedback controller* K *and exploratory signal level* σ_u*.*

Assumption VII.2 (`DFW` burn-in, redux)**.** *Consider running Algorithm 3 on data generated from arbitrary initial states* $x_1^{(1)}, \dots, x_1^{(H)}$*, norm-bounded by* $x_b\sqrt{\log(T)}$ *(see Line 7), by closed loop systems under stabilizing controllers* $K^{(1)}, \dots, K^{(H)}$ *with exploratory noise* $\sigma_u g_t$*,* $g_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_{\mathsf{U}}})$*, and representation* $\hat{\Phi}$*. Fixing a failure probability* $\delta \in (0, 1)$*, let the following hold:*

$$
\begin{aligned}
&\operatorname{rank}\left(\sum_{h=1}^{H} \theta_\star^{(h)} \theta_\star^{(h)\top}\right) = d_\theta \\
&t/N \ge \tilde{\mathcal{O}}(d_{\mathsf{X}} + d_\theta + \log(H/\delta)) \\
&d(\hat{\Phi}, \Phi_\star) \le d_{\mathsf{init}}.
\end{aligned}
$$

In the context of our problem, Assumption VII.2 in short requires that the optimal weights $\theta_\star^{(h)}$ span R^{d_θ}, which is necessary to identify the parameter space in all dimensions, and that the (initial) epoch length t is long enough for certain quantities to be well-defined. We note the assumption on the initial subspace distance is largely technical, as discussed in prior work [14, 36], and can be satisfied by an appropriate initialization scheme, which we do not discuss here. We now state a bound on the improvement of the subspace distance after running Algorithm 3.

Theorem VII.2 (`DFW` guarantee)**.** *Let Assumption II.1 and Assumption VII.2 hold and fix* $\delta \in (0, 1)$*. Then, with probability at least* $1 - \delta$ *running Algorithm 3 yields the following guarantee on the updated representation* $\hat{\Phi} \to \hat{\Phi}_N$*:*

$$
d(\hat{\Phi}_N, \Phi_\star) \le \rho^N d(\hat{\Phi}, \Phi^\star) + C_{\mathsf{contract}} \frac{\sqrt{N}}{\sigma_u \sqrt{tH}},
$$

where

$$
\begin{aligned}
C_{\mathsf{contract}} &= \frac{\overline{K}_{\mathsf{avg}}}{1 - \sqrt{2}\rho^N} \\
\rho &= \frac{1}{4}\kappa\left(\sum_{h=1}^{H} \theta_\star^{(h)} \theta_\star^{(h)\top}\right)^{-1} \\
\overline{K}_{\mathsf{avg}} &= \sqrt{\frac{1}{H}\sum_{h=1}^{H} \sigma^2 \|\theta_\star^{(h)}\|^2 (1 + \|K^{(h)}\|^2 + \sigma_u^2)} \\
&\quad \cdot \operatorname{polylog}(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, H, 1/\delta).
\end{aligned}
$$

VIII. High probability Bounds on the Success events

We begin by presenting several auxiliary lemmas from prior work.

A. Auxillary Lemmas

Lemma VIII.1. *(Noise bound (Lemma 13 of [34])) Let* $\delta \in (0, 1)$*. For any task* $h \in [H]$*, it holds that*

$$
\max_{1 \le t \le T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\| \le 4\sigma\sqrt{(d_{\mathsf{X}} + d_{\mathsf{U}}) \log\frac{T}{\delta}},
$$

with probability at least $1 - \delta$*.*

For any task $h \in [H]$, we define the empirical covariance matrix conditioned on the initial state $x_1^{(h)}$ as follows:

$$\Sigma_h^t(K^{(h)}, \sigma_u, x_1^{(h)}) \triangleq \mathbf{E}\left[\frac{1}{t}\sum_{s=1}^{t}\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix}\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix}^\top |x_1^{(h)}\right] \text{ and}$$

$$\bar{\Sigma}_h^t(K^{(h)}, \sigma_u, x_1^{(h)}) \triangleq \mathbf{E}\left[\frac{1}{t}\sum_{s=1}^{t}\left(\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix} - \mathbf{E}\left[\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix}|x_1^{(h)}\right]\right)\left(\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix} - \mathbf{E}\left[\begin{bmatrix} x_s^{(h)} \\ u_s^{(h)} \end{bmatrix}|x_1^{(h)}\right]\right)^\top\right].$$

where $\bar{\Sigma}_h^t(K^{(h)}, \sigma_u, x_1^{(h)})$ denotes the centered empirical covariance matrix from rolling out system h under control inputs $u_s^{(h)} = K^{(h)}x_s^{(h)} + \sigma_u g_s^{(h)}$ for t steps starting from an arbitrary initial state $x_1^{(h)}$.

Lemma VIII.2. *(Epoch-wise covariance bounds (Lemma 2 of [34])) For $t \geq 2$ and task $h \in [H]$, where we denote $K^{(h)} = K$, $A^{(h)} = A$, $B^{(h)} = B$, $\Sigma_h^t(K^{(h)}, \sigma_u, x_1^{(h)}) = \Sigma^t(K^{(h)}, \sigma_u, x_1^{(h)})$, and $\bar{\Sigma}_h^t(K^{(h)}, \sigma_u, x_1^{(h)}) = \bar{\Sigma}^t(K^{(h)}, \sigma_u, x_1^{(h)})$ we have*

1) $\bar{\Sigma}^t(K, \sigma_u, x_1) = \frac{1}{t}\sum_{s=0}^{t-2}\sum_{j=0}^{s}\begin{bmatrix} I \\ K \end{bmatrix} A_K^j(\sigma_u^2 B^\star(B^\star)^\top + I)\left(A_K^j\right)^\top\begin{bmatrix} I \\ K \end{bmatrix}^\top + \begin{bmatrix} 0 & 0 \\ 0 & \sigma_u^2 I_{d_{\mathsf{U}}} \end{bmatrix}$

2) $\Sigma^t(K, \sigma_u, x_1) = \bar{\Sigma}_k + \frac{1}{t}\sum_{s=0}^{t-1}\begin{bmatrix} I \\ K \end{bmatrix} A_K^s x_1 x_1^\top (A_K^s)^\top \begin{bmatrix} I \\ K \end{bmatrix}^\top$

3) $\Sigma^t(K, \sigma_u, x_1) \succeq \bar{\Sigma}^t(K, \sigma_u, x_1) \succeq \frac{\sigma_u^2}{2(1+2\|K\|^2+\sigma_u^2)} I$

4) $\Sigma^t(K, \sigma_u, x_1) \preceq (1 + \|P_K\| \frac{\|x_1^2\|}{t-1})\bar{\Sigma}^t(K, \sigma_u, x_1)$

5) $\left\|\bar{\Sigma}^t(K, \sigma_u, x_1)\right\| \leq 5\|P_K\|^2 \Psi_{B^\star}^2$.

Lemma VIII.3. *(State bounds (Lemma 15 of [34])) Consider rolling out the system $x_{s+1} = A^\star x_s + B^\star u_s + w_s$ from initial state $x_1^{(h)}$ for t time-steps under the control action $u_s = Kx_s + \sigma_u g_s$ where K is stabilizing and $\sigma_u \leq 1$. Suppose*

- $\|x_1\| \leq 16\|P_{K_0}\|^{3/2}\Psi_{B^\star}\max_{1\leq t\leq T}\left\|\begin{bmatrix} w_t \\ g_t \end{bmatrix}\right\|$
- $\|P_K\| \leq 2\|P_{K_0}\|$
- $t \geq \log_{\left(1-\frac{1}{\|P_K\|}\right)}\left(\frac{1}{4\|P_K\|}\right) + 1$.

Then for $s = 1, \ldots, t$

$$\|x_s\| \leq 40\|P_{K_0}\|^2 \Psi_{B^\star}\max_{1\leq t\leq T}\left\|\begin{bmatrix} w_t \\ g_t \end{bmatrix}\right\|.$$

Furthermore,

$$\|x_t\| \leq 16\|P_{K_0}\|^{3/2} \Psi_{B^\star}\max_{1\leq t\leq T}\left\|\begin{bmatrix} w_t \\ g_t \end{bmatrix}\right\|.$$

Theorem VIII.1. *(Theorem 3 of [28]) Define $\varepsilon^{(h)} \triangleq \frac{1}{2916\left\|P_\star^{(h)}\right\|}$ for any task $h \in [H]$. As long as*

$$\left\|\begin{bmatrix}\hat{A}^{(h)} & \hat{B}^{(h)}\end{bmatrix} - \begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}\right\|_F^2 \leq \varepsilon,$$

we have that $P_{\hat{K}}^{(h)} \preceq \frac{21}{20}P_\star^{(h)}$, $\left\|\hat{K}^{(h)} - K_\star^{(h)}\right\| \leq \frac{1}{6\left\|P_\star^{(h)}\right\|^{3/2}}$, and

$$\mathcal{J}^{(h)}(\hat{K}^{(h)}) - \mathcal{J}^{(h)}(K_\star^{(h)}) \leq 142\left\|P_\star^{(h)}\right\|^8\left\|\begin{bmatrix}\hat{A}^{(h)} & \hat{B}^{(h)}\end{bmatrix} - \begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}\right\|_F^2.$$

Using the above lemmas and theorems, we can mirror the arguments from Appendix C of [34] to show that the events of success $\mathcal{E}_{\mathsf{success},1}$ and $\mathcal{E}_{\mathsf{success},2}$ hold under high probability.

B. High Probability Bound on Success Event 1 (Hard to identify parameters)

Lemma VIII.4. *Running Algorithm 1 with the arguments defined in Theorem III.1, the event $\mathcal{E}_{\mathsf{success},1}$ holds with probability at least $1 - T^{-2}$.*

Proof. To show that the success event $\mathcal{E}_{\mathsf{success},1}$ holds under probability $1 - T^{-2}$ we can use an induction approach. For this purpose, we show, with high probability, that for every epoch $k \in [k_{\mathsf{fin}}]$, Algorithm 1 does not abort, i.e., the state and controller bounds are satisfied, the least-square estimation error is maintained small and scales according to the bound in $\mathcal{E}_{\mathsf{est},1}$, and the learned common representation contracts towards its optimal as in $\mathcal{E}_{\mathsf{cont}}$. We begin our analysis by studying the first epoch.

Base case: We consider the first epoch $k = 1$ as the base case of the induction approach. For convenience we assume that $x_1^{(h)} = 0$, for all tasks $h \in [H]$. However, it is worth noting that the proof below can be readily extended to bounded non-zero initial states.

- **The bounds on $\|x_t^{(h)}\|^2$ for $t = \{0, 1, \dots, \tau_1\}$ and $K_0^{(h)}$ are not violated:** We first show that, with high probability, the state and controller bounds are not violated during the first epoch. To do so we have to bound the worst-case behavior of the process and exploratory noises, which can be accomplished by using Lemma VIII.1 to obtain

$$\max_{1 \leq t \leq T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\| \leq 4\sigma \sqrt{3(d_{\mathsf{X}} + d_{\mathsf{U}}) \log(3HT)}. \tag{15}$$

with probability $1 - \frac{1}{3}T^{-2}$, for all tasks $h \in [H]$. Then, since the initial state norm (i.e., $\|x_1^{(h)}\| = 0$) satisfy

$$\left\| x_1^{(h)} \right\| \leq 16(P_0^{\vee})^{3/2} \Psi_B^{\vee} \max_{1 \leq t \leq T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\|,$$

and the initial epoch length can selected according to $\tau_1 \geq \frac{c \log \frac{1}{P_\star^{\wedge}}}{\log\left(1 - \frac{1}{P_\star^{\wedge}}\right)}$, for a sufficiently large constant c. We then may use Lemma VIII.3 to write

$$\left\| x_t^{(h)} \right\| \leq 40(P_0^{\vee})^2 \Psi_B^{\vee} \max_{1 \leq t \leq T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\|, \quad \forall t = \{0, 1, \dots, \tau_1\}, \tag{16}$$

and by using (15) in (16) we have

$$\left\| x_t^{(h)} \right\|^2 \leq 76800(P_0^{\vee})^4 (\Psi_B^{\vee})^2 \sigma^2 (d_{\mathsf{X}} + d_{\mathsf{U}}) \log(3HT), \quad \forall t = \{0, 1, \dots, \tau_1\}$$

with probability $1 - \frac{1}{3}T^{-3}$, $\forall h \in [H]$, which implies that $\left\| x_t^{(h)} \right\|^2 \leq x_b^2 \log T$. For the controller bound, we can notice that $\|K_0^{(h)}\|^2 \leq P_0^{\vee} \leq 2P_0^{\vee}$, which leads to $\|K_0^{(h)}\| \leq K_b$. Therefore, we define the event where the state and controller bounds are satisfied for the first epoch and obtain that $\mathcal{E}_{\text{bound},1}$ holds under high probability $1 - \frac{1}{3}T^{-2}$.

- **Controlling the least-square estimation error:** To control the estimation error at the first epoch, one may exploit Theorem VII.1. Note that a condition $\tau_{\mathsf{warm\_up}} \geq \sigma^4 P_0^{\vee} (\Psi_B^{\vee})^2 (d_{\mathsf{X}} + d_{\mathsf{U}})$ implies that $\tau_1 \geq c\tau_{\text{ls}}(K_0, 0, \frac{1}{3}T^{-3})$, for a sufficiently large constant c, which satisfy the condition of Theorem VII.1 to obtain

$$\begin{aligned} \left\| \begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 &\lesssim \frac{d_\theta \sigma^2 \log(HT)}{\tau_1 \min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)} \\ &+ \left(1 + \frac{\sigma^4 (P_0^{\vee})^7 (\Psi_B^{\vee})^6 (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_1 \min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)^2} \right) \frac{(P_0^{\vee})^2 (\Psi_B^{\vee})^2 d(\hat{\Phi}_1, \Phi_\star)^2 (\theta^{\vee})^2}{\min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)}. \end{aligned} \tag{17}$$

with probability $1 - \frac{1}{3}T^{-3}$, for all tasks $h \in [H]$. We note that the rate of the decay in the estimation error is controlled by the minimum eigenvalue of the input-state covariance matrix. Then, we may use the third point of Lemma VIII.2 to obtain

$$\min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1) \geq \frac{\sigma_1^2}{2(2 + 2(K_0^{\vee})^2)} \geq \frac{\sigma_1^2}{8P_0^{\vee}}, \tag{18}$$

where $K_0^{\vee} = \max_{h=1,\dots,H} \left\| K_0^{(h)} \right\|$ and the final inequality follows from the fact that $2 + 2(K_0^{\vee})^2 \leq 2 + 2P_0^{\vee} \leq 4P_0^{\vee}$ and $(P_0^{\vee}) \geq 1$. We then use (18) in (17) to obtain

$$\begin{aligned} \left\| \begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 &\lesssim \frac{d_\theta \sigma^2 (P_0^{\vee})}{\tau_1 \sigma_1^2} \log(HT) \\ &+ \left(1 + \frac{\sigma^4 (P_0^{\vee})^9 (\Psi_B^{\vee}) (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_1 \sigma_1^4} \right) \frac{(P_0^{\vee})^3 (\Psi_B^{\vee})^2 d(\hat{\Phi}_1, \Phi_\star)^2 (\theta^{\vee})^2}{\sigma_1^2}. \end{aligned}$$

and from $\sigma_1^2 \geq \frac{\sqrt{d_\theta / d_{\mathsf{U}}}}{\sqrt{\tau_1}}$ we have that

$$\left\| \begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \lesssim \frac{d_\theta \sigma^2 (P_0^{\vee})}{\tau_1 \sigma_1^2} \log(HT)$$

$$+ \sigma^4 \frac{d_{\mathsf{U}}}{d_\theta} (P_0^\vee)^{12} (\Psi_B^\vee)^8 (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))(\theta^\vee)^2 \frac{d^2(\hat{\Phi}_1, \Phi_\star)}{\sigma_1^2}.$$

Then, by defining $\beta_1 \triangleq C_{\mathsf{bias},1} \sigma^4 (P_0^\vee)^{12} (\Psi_B^\vee)^8 (\theta^\vee)^2 (d_{\mathsf{X}} + d_{\mathsf{U}}) \frac{d_{\mathsf{U}}}{d_\theta}$ we obtain

$$\left\| \begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq C_{\text{est},1} \frac{d_\theta \sigma^2 (P_0^\vee)}{\tau_1 \sigma_1^2} \log(HT) + \frac{\beta_1 \log(HT) d^2(\hat{\Phi}_1, \Phi_\star)}{\sigma_1^2}.$$

Therefore, by defining the event $\mathcal{E}_{\text{ls},1}$ where the above least-square estimation error at the first epoch holds, we have that $\mathcal{E}_{\text{ls},1}$ holds under probability $1 - \frac{1}{3} T^{-3}$, for all tasks $h \in [H]$.

- **Controlling the error in the learned representation:** For the first epoch, we initialize the representation as $\hat{\Phi}_0$. Then, Algorithm 1 play $K_0^{(h)}$ for all tasks $h \in [H]$ to collect a multi-task dataset that is leveraged to compute $\hat{\Phi}_1$ via Algorithm 3. Therefore, we can set $\tau_1 \geq c \tau_{\mathsf{dfw}}$, for a sufficiently large constant c, to use Theorem VII.2 to obtain

$$d\left(\hat{\Phi}_+, \Phi_\star\right) \leq \rho d\left(\hat{\Phi}, \Phi_\star\right) + \frac{C_{\mathsf{contract}} \sqrt{N} \log(HT)}{\sqrt{H \tau_1 \sigma_1^2}},$$

with probability $1 - \frac{1}{3} T^{-3}$. Then, by unrolling the above expression for N iterations of Algorithm 3, we have that

$$d\left(\hat{\Phi}_1, \Phi_\star\right) \leq \rho^N d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{C_{\mathsf{contract}} \sqrt{N}}{1 - \sqrt{2} \rho^N} \frac{\log(HT)}{\sqrt{H \tau_1 \sigma_1^2}},$$

and we denote $\mathcal{E}_{\text{c},1}$ as the event where the above bound holds under probability $1 - \frac{1}{3} T^{-3}$, for the first epoch.

Induction step: We now introduce an induction step to extend our analysis for every epoch. For this purpose, based on the first epoch one may establish the following inductive hypothesis:

$$\textbf{Bounded state}: \;\; \left\| x_{\tau_k}^{(h)} \right\| \leq 16 (P_0^\vee)^{3/2} (\Psi_B^\vee) \max_{1 \leq t \leq T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\|, \tag{19}$$

$$\textbf{Least-square error}: \;\; \left\| \begin{bmatrix} \hat{A}_k^{(h)} & \hat{B}_k^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq C_{\text{est},1} \frac{d_\theta \sigma^2 (P_0^\vee)}{\tau_k \sigma_k^2} \log(HT) + \frac{\beta_1 \log(HT) d^2(\hat{\Phi}_k, \Phi_\star)}{\sigma_k^2}, \tag{20}$$

and

$$\textbf{Representation error}: \;\; d(\hat{\Phi}_k, \Phi_\star) \leq \rho^{kN} d(\hat{\Phi}_0, \Phi_\star) + \frac{C_{\mathsf{contract}} \sqrt{N} \log(HT)}{1 - \sqrt{2} \rho^N \sqrt{H \tau_k \sigma_k^2}}, \tag{21}$$

- **Controlling the least-square estimation error:** To control the estimation error along the epochs, we first need to control the variance term in (21) and obtain $d(\hat{\Phi}_k, \Phi_\star) \leq d(\hat{\Phi}_0, \Phi_\star)$, for all $k \in [k_{\text{fin}}\,]$. We can set $\tau_1 \geq 8 \frac{C_{\mathsf{contract}}^3 N^{3/2} \log^3(HT)}{(1 - \sqrt{2} \rho^N)^3 H^{3/4} d(\hat{\Phi}_0, \Phi_\star)^3}$, and use the condition on the exploratory sequence $\sigma_k^2 \geq \tau_k^{-1/3} H^{-1/2}$, to obtain $d(\hat{\Phi}_k, \Phi_\star) \leq d(\hat{\Phi}_0, \Phi_\star)$. Moreover, we can use a condition on the first epoch length such that $\tau_k \geq \tau_1 \geq c\left(\sigma^2 (P_0^\vee) \frac{\sqrt{d_\theta d_{\mathsf{U}}}}{\varepsilon^\wedge} \log T\right)^2$, for a sufficiently large constant c, along with the condition on the exploratory sequence $\sigma_k^2 \geq \sqrt{\frac{d_{\mathsf{U}} d_\theta}{\tau_k}}$, and initial representation error $d(\hat{\Phi}_0, \Phi_\star) \leq \sqrt{\frac{\varepsilon^\wedge}{2 \beta_1 \log(HT)}}$ to obtain $\left\| \begin{bmatrix} \hat{A}_k^{(h)} & \hat{B}_k^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq \varepsilon^\wedge \leq \varepsilon^{(h)}$. Therefore, the conditions of Lemma VIII.1 are satisfied and we may write

$$\tau_{\text{ls}}(\hat{K}_{k+1}^{(h)}, x_b^2 \log T, \frac{1}{3} T^{-3}) \leq 2 \tau_{\text{ls}}(K_\star^{(h)}, x_b^2 \log T, \frac{1}{3} T^{-3}) \text{ and } \left\| P_{\hat{K}_{k+1}}^{(h)} \right\| \leq 1.05 (P_0^\vee) \leq 2 (P_0^\vee).$$

where the first is true since the lower bound on τ_{ls} scales with $\|P_K^{(h)}\|$. Therefore, by selecting the initial epoch lengh according to $\tau_1 \geq c \tau_{\text{ls}}(K_\star^{(h)}, x_b^2 \log T, \frac{1}{2} T^{-3})$, for a sufficiently large constant c, we can use Theorem VII.1 to obtain, with probability $1 - \frac{1}{3} T^{-3}$, for all tasks $h \in [H]$, the following

$$\left\| \begin{bmatrix} \hat{A}_{k+1}^{(h)} & \hat{B}_{k+1}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \lesssim \frac{d_\theta \sigma^2 (P_0^\vee)}{\tau_{k+1} \sigma_{k+1}^2} \log(HT)$$
$$+ \left(1 + \frac{\sigma^4 (P_0^\vee)^9 (\Psi_B^\vee)(d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_{k+1} \sigma_{k+1}^4}\right) \frac{(P_0^\vee)^3 (\Psi_B^\vee)^2 d(\hat{\Phi}_{k+1}, \Phi_\star)^2 (\theta^\vee)^2}{\sigma_{k+1}^2},$$

where we can use $\left\|P^{(h)}_{\hat{K}_{k+1}}\right\| \leq 2(P_0^{\vee})$ and control the minimum eigenvalue of the input-state covariance matrix as follows

$$\min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_{k+1}^{\top}\left(\bar{\Sigma}^{\tau_{k+1}}(\hat{K}^{(h)}_{k+1}, \sigma_{k+1}, x^{(h)}_{k+1}) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_{k+1}) \geq \frac{\sigma_{k+1}^2}{8(P_0^{\vee})},$$

which implies that from the condition $\sigma_{k+1}^2 \geq \frac{\sqrt{d_\theta/d_{\mathsf{U}}}}{\sqrt{\tau_{k+1}}}$ and the definition of β_1, we obtain

$$\left\|\begin{bmatrix}\hat{A}^{(h)}_{k+1} & \hat{B}^{(h)}_{k+1}\end{bmatrix} - \begin{bmatrix}A^{(h)}_{\star} & B^{(h)}_{\star}\end{bmatrix}\right\|_F^2 \leq C_{\text{est},1}\frac{d_\theta \sigma^2(P_0^{\vee})}{\tau_{k+1}\sigma_{k+1}^2}\log(HT) + \frac{\beta_1 \log(HT) d^2(\hat{\Phi}_{k+1}, \Phi_\star)}{\sigma_{k+1}^2}.$$

Therefore, we proved that since $\mathcal{E}_{\text{ls},k}$ holds under high probability, then $\mathcal{E}_{\text{ls},k}$ also holds under probability $1-\frac{1}{3}T^{-3}$. By union bounding for all the epochs we have $\mathcal{E}_{\mathsf{est},1} \subseteq \mathcal{E}_{\text{ls},1} \cap \dots \cap \mathcal{E}_{\text{ls},k_{\mathsf{fin}}}$ holds under probability of at least $1-\frac{1}{3}T^{-2}$.

- **The bounds on $\|x_t^{(h)}\|^2$ for $t = \{\tau_k+1, \dots, \tau_{k+1}\}$ and $K_0^{(h)}$ are not violated:** By following our inductive hypothesis, we have

$$\left\|x^{(h)}_{\tau_k}\right\| \leq 16(P_0^{\vee})^{3/2}(\Psi_B^{\vee}) \max_{1\leq t\leq T}\left\|\begin{bmatrix} w_t^{(h)} \\ g_t^{(h)}\end{bmatrix}\right\|,$$

which combined with $\left\|P^{(h)}_{\hat{K}_{k+1}}\right\| \leq 2(P_0^{\vee})$ and $\tau_1 \geq c\frac{\log\frac{1}{P_\star^{\wedge}}}{\log\left(1-\frac{1}{P_\star^{\wedge}}\right)}$, for a sufficiently large constant c, we can exploit Lemma VIII.3 to write

$$\left\|x^{(h)}_{t}\right\| \leq 40(P_0^{\vee})^{2}(\Psi_B^{\vee}) \max_{1\leq t\leq T}\left\|\begin{bmatrix} w_t^{(h)} \\ g_t^{(h)}\end{bmatrix}\right\|, \quad \forall t = \{\tau_k+1, \dots, \tau_{k+1}\}, \tag{22}$$

and by using (15) in (22), the state bound satisfies $\left\|x_t^{(h)}\right\|^2 \leq x_b^2 \log T$ with probability $1-\frac{1}{3}T^{-2}$, for all tasks $h \in [H]$. Moreover, the controller bound is satisfied since $\left\|\hat{K}^{(h)}_{k+1}\right\|^2 \leq \left\|P_{\hat{K}_{k+1}}\right\| \leq 2(P_0^{\vee})$, which implies that $\left\|\hat{K}^{(h)}_{k+1}\right\| \leq K_b$. Therefore, $\mathcal{E}_{\mathsf{bound},k+1}$ holds under probability $1-\frac{1}{3}T^{-2}$, which implies that $\mathcal{E}_{\mathsf{bound}}$ holds under probability of at least $1-\frac{1}{3}T^{-2}$.

- **Controlling the error in the learned representation:** Following our inductive hypothesis on the contraction of the learned representation and the condition on initial epoch length $\tau_1 \geq c\tau_{\mathsf{dfw}}$, for a sufficiently large constant c, we can use Theorem VII.2 to obtain

$$d\left(\hat{\Phi}_{k+1}, \Phi_\star\right) \leq \rho^N d\left(\hat{\Phi}_{k}, \Phi_\star\right) + \frac{C_{\mathsf{contract}}\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}}, \tag{23}$$

with probability $1-\frac{1}{3}T^{-3}$. Therefore, by applying (21) to (23) we have

$$\begin{aligned}
d\left(\hat{\Phi}_{k+1}, \Phi_\star\right) &\leq \rho^N \rho^{kN} d\left(\hat{\Phi}_{0}, \Phi_\star\right) + \frac{\rho^N C_{\mathsf{contract}}\sqrt{N}}{1-\sqrt{2}\rho^N}\frac{\log(HT)}{\sqrt{H\tau_k\sigma_k^2}} + \frac{C_{\mathsf{contract}}\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&\overset{(i)}{\leq} \rho^{(k+1)N} d\left(\hat{\Phi}_{0}, \Phi_\star\right) + \frac{\sqrt{2}\rho^N C_{\mathsf{contract}}\sqrt{N}}{1-\sqrt{2}\rho^N}\frac{\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} + \frac{C_{\mathsf{contract}}\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&= \rho^{(k+1)N} d\left(\hat{\Phi}_{0}, \Phi_\star\right) + \left(1 + \frac{\sqrt{2}\rho^N}{1-\sqrt{2}\rho^N}\right)\frac{C_{\mathsf{contract}}\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&= \rho^{(k+1)N} d\left(\hat{\Phi}_{0}, \Phi_\star\right) + \frac{C_{\mathsf{contract}}\sqrt{N}}{1-\sqrt{2}\rho^N}\frac{\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}},
\end{aligned}$$

where (i) follows from the fact that $\tau_k\sigma_k^2 \geq \frac{1}{2}\tau_{k+1}\sigma_{k+1}^2$. Therefore, we conclude that since $\mathcal{E}_{\text{c},k}$ holds under probability $1-\frac{1}{3}T^{-3}$, then $\mathcal{E}_{\text{c},k+1}$ also holds under at least the same probability. Then, by union bounding for all the epochs, we have that $\mathcal{E}_{\mathsf{cont}} \subseteq \mathcal{E}_{\text{c},1} \cap \dots \cap \mathcal{E}_{\text{c},k_{\mathsf{fin}}}$ holds under probability of at least $1-\frac{1}{3}T^{-2}$.

We complete the proof by union bounding the events $\mathcal{E}_{\mathsf{bound}}$, $\mathcal{E}_{\mathsf{est},1}$, and $\mathcal{E}_{\mathsf{cont}}$. We then have that $\mathcal{E}_{\mathsf{success},1} \subseteq \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},1} \cap \mathcal{E}_{\mathsf{cont}}$ holds under probability of at least $1-T^{-2}$.

□

C. High Probability Bound on Success Event 2 (Easy to identify parameters)

Lemma VIII.5. *Running Algorithm 1 with the arguments defined in Theorem III.2, the event $\mathcal{E}_{\mathsf{success},2}$ holds with probability at least $1 - T^{-2}$.*

Proof. Analogous to the probability of success event $\mathcal{E}_{\text{success}}$, we show that $\mathcal{E}_{\text{success},2}$ holds with probability $1 - T^{-2}$ by induction. To do so, we show, with high probability, that for every epoch $k \in [k_{\text{fin}}\,]$, Algorithm 1 does not abort, i.e., the state and controller bounds are satisfied, the least-square estimation error is maintained small and scales according to the bound in $\mathcal{E}_{\text{est},2}$, and the learned common representation contracts towards its optimal as in $\mathcal{E}_{\text{cont}}$. We begin our analysis by studying the first epoch.

Base case: We consider the first epoch $k = 1$ as the base case of the induction approach. For convenience we assume that $x_1^{(h)} = 0$, for all tasks $h \in [H]$. However, it is worth noting that our proofs can be readily extended to bounded non-zero initial states.

- **The bounds on $\|x_t^{(h)}\|^2$ for $t = \{0, 1, \dots, \tau_1\}$ and $K_0^{(h)}$ are not violated:** We begin our the analysis, by showing with high probability that the state and controller bounds are not violated. In order to ensure that the bounds on the state and controller are not violated, we first bound the worst-case behavior of the process and exploratory noises. For this purpose, we use Lemma VIII.1 to obtain
$$\max_{1 \le t \le T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\| \le 4\sigma \sqrt{3(d_{\mathsf{X}} + d_{\mathsf{U}}) \log(3HT)}. \tag{24}$$
with probability $1 - \frac{1}{3}T^{-2}$, for all tasks $h \in [H]$. Therefore, since the initial state satisfy
$$\left\| x_1^{(h)} \right\| \le 16 (P_0^{\vee})^{3/2} (\Psi_B^{\vee}) \max_{1 \le t \le T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\|,$$
and the initial epoch length can be selected such that $\tau_1 \ge \frac{c \log \frac{1}{8\sqrt{P_\star^{\wedge}}}}{\log\left(1 - \frac{1}{2P_\star^{\wedge}}\right)}$, for a sufficiently large constant c, respectively. We use Lemma VIII.3 to write
$$\left\| x_t^{(h)} \right\| \le 40 (P_0^{\vee})^2 (\Psi_B^{\vee}) \max_{1 \le t \le T} \left\| \begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix} \right\|, \quad \forall t = \{0, 1, \dots, \tau_1\} \tag{25}$$
Therefore, by using (24) in (25) we have
$$\left\| x_t^{(h)} \right\|^2 \le 76800 (P_0^{\vee})^4 (\Psi_B^{\vee})^2 \sigma^2 (d_{\mathsf{X}} + d_{\mathsf{U}}) \log(3HT), \quad \forall t = \{0, 1, \dots, \tau_1\}$$
with probability $1 - \frac{1}{3}T^{-2}$, for all tasks $h \in [H]$. This implies that the state bound is satisfied, i.e., $\left\| x_t^{(h)} \right\|^2 \le x_b^2 \log T$. On the other hand, for the controller bound, we note that $\|K_0^{(h)}\|^2 \le P_0^{\vee} \le 2P_0^{\vee}$, which implies that $\|K_0^{(h)}\| \le K_b$. Therefore, $\mathcal{E}_{\text{bound},1}$ (i.e., the event where the state and controller bounds are satisfied at the first epoch) holds with probability $1 - \frac{1}{3}T^{-2}$.

- **Controlling the least-square estimation error:** To control the estimation error at the first epoch, we can use Theorem VII.1. In addition, a condition $\tau_{\mathsf{warm\_up}} \ge \sigma^4 (P_0^{\vee})^3 (\Psi_B^{\vee})^2 (d_{\mathsf{X}} + d_{\mathsf{U}})$ implies that $\tau_1 \ge c\tau_{\mathsf{ls}}(K_0, 0, \frac{1}{3}T^{-3})$, for a sufficiently large constant c. Then, from Theorem VII.1, we have
$$\begin{aligned} \left\| \begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 &\lesssim \frac{d_\theta \sigma^2 \log(HT)}{\tau_1 \min\limits_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1^2, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)} \\ &+ \left(1 + \frac{\sigma^4 (P_0^{\vee})^7 (\Psi_B^{\vee})^6 (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_1 \min\limits_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1^2, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)^2} \right) \frac{(P_0^{\vee})^2 (\Psi_B^{\vee})^2 d(\hat{\Phi}_1, \Phi_\star)^2 (\theta^{\vee})^2}{\min\limits_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1^2, 0) \otimes I_{d_{\mathsf{X}}} \right) \hat{\Phi}_1)}. \end{aligned}$$
with probability $1 - \frac{1}{3}T^{-3}$, for all tasks $h \in [H]$. The rate of the decay in the estimation error is controlled by the minimum eigenvalue of the input-state covariance matrix. The main difference between this proof to the one for $\mathcal{E}_{\text{success},2}$

is on the lower bound of minimum eigenvalue of the input-state covariance matrix. Here, we exploit Assumption III.3 to write

$$\begin{aligned}\min_{h=1,\dots,H} v^\top \hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K_0^{(h)}, \sigma_1^2, 0) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_1 v &\geq \min_{h=1,\dots,H} \frac{1}{2} v\hat{\Phi}_1^\top \left(\begin{bmatrix} I \\ K_0^{(h)} \end{bmatrix}\begin{bmatrix} I \\ K_0^{(h)} \end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_1 v \\ &= \min_{h=1,\dots,H} \frac{1}{2}\left(\sum_{i=1}^{d_\theta} v_i \hat{\Phi}_{1,i}\right)^\top \left(\begin{bmatrix} I \\ K_0^{(h)} \end{bmatrix}\begin{bmatrix} I \\ K_0^{(h)} \end{bmatrix}^\top \otimes I_{d_{\mathsf{X}}}\right)\left(\sum_{i=1}^{d_\theta} v_i \hat{\Phi}_{1,i}\right) \\ &= \min_{h=1,\dots,H} \frac{1}{2}\left\|\sum_{i=1}^{d_\theta} v_i \,\mathsf{vec}^{-1}\left(\hat{\Phi}_{1,i}\right)\begin{bmatrix} I \\ K_0^{(h)} \end{bmatrix}\right\|_F^2 \\ &= \min_{h=1,\dots,H} \frac{1}{2}\left\|\sum_{i=1}^{d_\theta} v_i \left[\hat{\Phi}_{1,i}^{A^{(h)}} + \hat{\Phi}_{1,i}^{B^{(h)}} K_0^{(h)}\right]\right\|_F^2 \\ &\geq \min_{h=1,\dots,H} \min_{v, \|v\|=1} \frac{1}{2}\left\|\sum_{i=1}^{d_\theta} v_i \left[\hat{\Phi}_{1,i}^{A^{(h)}} + \hat{\Phi}_{1,i}^{B^{(h)}} K_0^{(h)}\right]\right\|_F^2 \geq \frac{\alpha^2}{2},\end{aligned}$$

then, since $\|v\| = 1$, we have that

$$\min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}_1^\top \left(\bar{\Sigma}_h^{\tau_1}(K^{(h)}, 0, 0) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_1) \geq \frac{\alpha^2}{2},$$

which implies that

$$\begin{aligned}\left\|\begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix}\right\|_F^2 &\lesssim \frac{d_\theta \sigma^2}{\tau_1 \alpha^2}\log(HT) \\ &+ \left(1 + \frac{\sigma^4 (P_0^\vee)^7 (\Psi_B^\vee)^6 (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_1 \alpha^4}\right)\frac{(P_0^\vee)^2 (\Psi_B^\vee)^2 d(\hat{\Phi}_1, \Phi_\star)^2 (\theta^\vee)^2}{\alpha^2}.\end{aligned}$$

and by using a condition on the initial epoch length $\tau_1 \geq \frac{c\sigma^2 d_\theta \log(HT)}{2\varepsilon^\wedge \alpha^2}$, for a sufficiently large constant c, we have

$$\left\|\begin{bmatrix} \hat{A}_1^{(h)} & \hat{B}_1^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix}\right\|_F^2 \leq C_{\mathsf{est},2}\frac{\sigma^2 d_\theta \log(HT)}{\tau_1 \alpha^2} + \beta_2 d(\hat{\Phi}_1, \Phi_\star)^2,$$

where $\beta_2 \triangleq C_{\mathsf{bias},2}\frac{\sigma^2 \varepsilon^\wedge (\Psi_B^\vee)^8 (\theta^\vee)^2 (d_{\mathsf{X}} + d_{\mathsf{U}})}{d_\theta \min\{\alpha^2, \alpha^4\}}$. Then, by defining the event $\mathcal{E}_{\mathrm{ls},1}$ where the above estimation error bound holds for the first epoch, we have that $\mathcal{E}_{\mathrm{ls},1}$ holds with probability $1 - \frac{1}{3}T^{-3}$ for all tasks $h \in [H]$.

- **Controlling the contraction in the learned representation:** For the first epoch, we initialize the representation as $\hat{\Phi}_0$. Then, Algorithm 1 play $K_0^{(h)}$ for all tasks $h \in [H]$ to collect a multi-task dataset that is leveraged to update the representation $\hat{\Phi}_1$ with N iterations of Algorithm 3. Then, since $\tau_1 \geq c\tau_{\mathrm{dfw}}$, for a sufficiently large constant c, we can use Theorem VII.2 to obtain with probability $1 - \frac{1}{3}T^{-3}$, the following

$$d\left(\hat{\Phi}_1, \Phi_\star\right) \leq \rho^N d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{C_{\mathsf{contract}}\sqrt{N}}{1 - \sqrt{2}\rho^N}\frac{\log(HT)}{\sqrt{H\tau_1\sigma_1^2}},$$

and we denote $\mathcal{E}_{\mathrm{c},1}$ as the event where the above bound holds with probability $1 - \frac{1}{3}T^{-3}$, for the first epoch.

Induction step: We now use an induction step with the following inductive hypothesis:

$$\textbf{Bounded state}: \left\|x_{\tau_k}^{(h)}\right\| \leq 16(P_0^\vee)^{3/2}(\Psi_B^\vee) \max_{1 \leq t \leq T}\left\|\begin{bmatrix} w_t^{(h)} \\ g_t^{(h)} \end{bmatrix}\right\|, \tag{26}$$

$$\textbf{Least-square error}: \left\|\begin{bmatrix} \hat{A}_k^{(h)} \hat{B}_k^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} B_\star^{(h)} \end{bmatrix}\right\|_F^2 \leq C_{\mathsf{est},2}\frac{\sigma^2 d_\theta \log(HT)}{\tau_k \alpha^2} + \beta_2 d(\hat{\Phi}_k, \Phi_\star)^2, \tag{27}$$

and

$$\textbf{Representation error}: d\left(\hat{\Phi}_k, \Phi_\star\right) \leq \rho^{kN} d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{C_{\mathsf{contract}}\sqrt{N}}{1 - \sqrt{2}\rho^N}\frac{\log(HT)}{\sqrt{H\tau_k\sigma_k^2}}, \tag{28}$$

- **Controlling the least-square estimation error:** To control the estimation error throughout the epochs, we first note that since $\tau_1 \geq 8\frac{C^2_{\mathsf{contract}} N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 H^{1/2} d(\hat{\Phi}_0, \Phi_\star)^2}$, we then have $d\left(\hat{\Phi}_k, \Phi_\star\right) \leq d\left(\hat{\Phi}_0, \Phi_\star\right)$ for all $k \in [k_{\mathsf{fin}}]$. Moreover, we can select the first epoch length as $\tau_1 \geq \frac{c\sigma^2 d_\theta \log(HT)}{2\varepsilon^\wedge \alpha^2}$, for a sufficiently large constant c, along with the initial representation error $d(\hat{\Phi}_0, \Phi_\star) \leq \sqrt{\frac{\varepsilon^\wedge}{2\beta_2}}$ to obtain $\left\| \begin{bmatrix} \hat{A}^{(h)}_k & \hat{B}^{(h)}_k \end{bmatrix} - \begin{bmatrix} A^{(h)}_\star & B^{(h)}_\star \end{bmatrix} \right\|^2_F \leq \varepsilon^\wedge \leq \varepsilon^{(h)}$. Therefore, the conditions of Lemma VIII.1 are satisfied and we can write

$$\tau_{\mathsf{ls}}(\hat{K}^{(h)}_{k+1}, x_b^2 \log T, \frac{1}{3}T^{-3}) \leq 2\tau_{\mathsf{ls}}(K^{(h)}_\star, x_b^2 \log T, \frac{1}{3}T^{-3}) \text{ and } \left\| P^{(h)}_{\hat{K}_{k+1}} \right\| \leq 1.05(P^\vee_0) \leq 2(P^\vee_0).$$

where the first is due to the fact that the lower bound on τ_{ls} scales with $\|P^{(h)}_K\|$. Therefore, by setting the first epoch length such that $\tau_1 \geq c\tau_{\mathsf{ls}}(K^{(h)}_\star, x_b^2 \log T, \frac{1}{2}T^{-3})$, for a sufficiently large universal constant c, we use Theorem VII.1 to obtain

$$\begin{aligned} \left\| \begin{bmatrix} \hat{A}^{(h)}_{k+1} & \hat{B}^{(h)}_{k+1} \end{bmatrix} - \begin{bmatrix} A^{(h)}_\star & B^{(h)}_\star \end{bmatrix} \right\|^2_F &\lesssim \frac{d_\theta \sigma^2 \log(HT)}{\tau_{k+1} \min\limits_{h=1,\dots,H} \lambda_{\min}\left(\hat{\Phi}^\top_{k+1}\left(\bar{\Sigma}^{\tau_{k+1}}(K^{(h)}_0, \sigma^2_{k+1}, x^{(h)}_{\tau_{k+1}}) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_{k+1}\right)} \\ &+ \left(1 + \frac{\sigma^4 (P^\vee_0)^7 (\Psi^\vee_B)^6 (d_{\mathsf{X}} + d_{\mathsf{U}} + \log(HT))}{\tau_{k+1} \min\limits_{h=1,\dots,H} \lambda_{\min}\left(\hat{\Phi}^\top_{k+1}\left(\bar{\Sigma}^{\tau_{k+1}}(K^{(h)}_0, \sigma^2_{k+1}, x^{(h)}_{\tau_{k+1}}) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_{k+1}\right)^2}\right) \\ &\times \frac{(P^\vee_0)^2 (\Psi^\vee_B)^2 d(\hat{\Phi}_1, \Phi_\star)^2 (\theta^\vee)^2}{\min\limits_{h=1,\dots,H} \lambda_{\min}\left(\hat{\Phi}^\top_{k+1}\left(\bar{\Sigma}^{\tau_{k+1}}(K^{(h)}_0, \sigma^2_{k+1}, x^{(h)}_{\tau_{k+1}}) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_{k+1}\right)}, \end{aligned}$$

with probability $1 - \frac{1}{3}T^{-3}$ for all tasks $h \in [H]$. In the above expression we also use $\left\| P^{(h)}_{\hat{K}_{k+1}} \right\| \leq 2(P^\vee_0)$. We control the minimum eigenvalue of the input-state covariance matrix as follows

$$\min_{h=1,\dots,H} \lambda_{\min}(\hat{\Phi}^\top_{k+1}\left(\bar{\Sigma}^{\tau_{k+1}}(\hat{K}^{(h)}_{k+1}, \sigma^2_{k+1}, x^{(h)}_{\tau_{k+1}}) \otimes I_{d_{\mathsf{X}}}\right)\hat{\Phi}_{k+1}) \geq \frac{\alpha^2}{8},$$

since

$$\begin{aligned} \min_{h=1,\dots,H} \left\| \sum_{i=1}^{d_\theta} v_i \left(\hat{\Phi}^{A^{(h)}}_i + \hat{\Phi}^{B^{(h)}}_i K^{(h)\star}\right) + \sum_{i=1}^{d_\theta} v_i \hat{\Phi}^{B^{(h)}}_i (\hat{K}^{(h)}_{k+1} - K^{(h)\star}) \right\|_F &\geq \alpha - \left\| \hat{K}^{(h)}_{k+1} - K^{(h)\star} \right\| \\ &\geq \alpha - \frac{1}{6(P^\vee_0)^{3/2}} \geq \frac{\alpha}{2}. \end{aligned}$$

which implies that

$$\left\| \begin{bmatrix} \hat{A}^{(h)}_{k+1} & \hat{B}^{(h)}_{k+1} \end{bmatrix} - \begin{bmatrix} A^{(h)}_\star & B^{(h)}_\star \end{bmatrix} \right\|^2_F \leq C_{\mathsf{est},2} \frac{\sigma^2 d_\theta \log(HT)}{\tau_{k+1}\alpha^2} + \beta_2 d(\hat{\Phi}_{k+1}, \Phi_\star)^2,$$

Therefore, since $\mathcal{E}_{\mathsf{ls},k}$ holds with high probability, then $\mathcal{E}_{\mathsf{ls},k}$ also holds with probability $1 - \frac{1}{3}T^{-3}$. This implies that $\mathcal{E}_{\mathsf{est},2} \subseteq \mathcal{E}_{\mathsf{ls},1} \cap \cdots \cap \mathcal{E}_{\mathsf{ls},k_{\mathsf{fin}}}$ holds with probability of at least $1 - \frac{1}{3}T^{-2}$ for all tasks $h \in [H]$.

- **The bounds on $\|x^{(h)}_t\|^2$ for $t = \{\tau_k + 1, \dots, \tau_{k+1}\}$ and $K^{(h)}_0$ are not violated:** By following our inductive hypothesis, we have

$$\left\| x^{(h)}_{\tau_k} \right\| \leq 16(P^\vee_0)^{3/2}(\Psi^\vee_B) \max_{1 \leq t \leq T} \left\| \begin{bmatrix} w^{(h)}_t \\ g^{(h)}_t \end{bmatrix} \right\|,$$

which combined with $\left\| P^{(h)}_{\hat{K}_{k+1}} \right\| \leq 2(P^\vee_0)$ and $\tau_1 \geq c\frac{\log \frac{1}{8\sqrt{P^\wedge_\star}}}{\log\left(1 - \frac{1}{2P^\wedge_\star}\right)}$, for a sufficiently large constant c, we can use Lemma VIII.3 to write

$$\left\| x^{(h)}_t \right\| \leq 40(P^\vee_0)^2(\Psi^\vee_B) \max_{1 \leq t \leq T} \left\| \begin{bmatrix} w^{(h)}_t \\ g^{(h)}_t \end{bmatrix} \right\|, \quad \forall t = \{\tau_k + 1, \dots, \tau_{k+1}\}, \tag{29}$$

and by using (24) in (29), the state bound is satisfied , i.e., $\left\| x^{(h)}_t \right\|^2 \leq x_b^2 \log T$, with probability $1 - \frac{1}{3}T^{-2}$, for all tasks $h \in [H]$. Moreover, the controller bound is verified since $\left\| \hat{K}^{(h)}_{k+1} \right\|^2 \leq \left\| P_{\hat{K}^{(h)}_{k+1}} \right\| \leq 2(P^\vee_0)$, which implies that

$\left\|\hat{K}_{k+1}^{(h)}\right\| \le K_b$. Then, $\mathcal{E}_{\mathsf{bound},k+1}$ holds with probability $1 - \frac{1}{3}T^{-2}$, which implies that $\mathcal{E}_{\mathsf{bound}}$ holds with probability of at least $1 - \frac{1}{3}T^{-2}$, for all tasks $h \in [H]$.

- **Controlling the error in the learned representation:** Following our inductive hypothesis on the contraction of the learned representation and the condition on initial epoch length $\tau_1 \ge c\tau_{\text{dfw}}$, for a sufficiently large constant c, we can use Theorem VII.2 to obtain

$$d\left(\hat{\Phi}_{k+1}, \Phi_\star\right) \le \rho^N d\left(\hat{\Phi}_k, \Phi_\star\right) + C_{\mathsf{contract}} \frac{\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}}, \tag{30}$$

with probability $1 - \frac{1}{3}T^{-3}$. Therefore, by applying (28) to (30) we have

$$\begin{aligned}
d\left(\hat{\Phi}_{k+1}, \Phi_\star\right) &\le \rho^N \rho^{kN} d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{\rho^N C_{\mathsf{contract}}\sqrt{N}}{1-\sqrt{2}\rho^N} \frac{\log(HT)}{\sqrt{H\tau_k\sigma_k^2}} + C_{\mathsf{contract}} \frac{\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&\overset{(i)}{\le} \rho^{(k+1)N} d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{\sqrt{2}\rho^N C_{\mathsf{contract}}\sqrt{N}}{1-\sqrt{2}\rho^N} \frac{\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} + C_{\mathsf{contract}} \frac{\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&= \rho^{(k+1)N} d\left(\hat{\Phi}_0, \Phi_\star\right) + \left(1 + \frac{\sqrt{2}\rho^N}{1-\sqrt{2}\rho^N}\right) C_{\mathsf{contract}} \frac{\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}} \\
&= \rho^{(k+1)N} d\left(\hat{\Phi}_0, \Phi_\star\right) + \frac{C_{\mathsf{contract}}}{1-\sqrt{2}\rho^N} \frac{\sqrt{N}\log(HT)}{\sqrt{H\tau_{k+1}\sigma_{k+1}^2}},
\end{aligned}$$

where (i) follows from the fact that $\tau_k\sigma_k^2 \ge \frac{1}{2}\tau_{k+1}\sigma_{k+1}^2$. Therefore, we conclude that since $\mathcal{E}_{\mathsf{c},k}$ holds with probability $1 - \frac{1}{3}T^{-3}$, then $\mathcal{E}_{\mathsf{c},k+1}$ also holds with at least the same probability. Then, by union bounding for all the epochs, we have that $\mathcal{E}_{\mathsf{cont}} \subseteq \mathcal{E}_{\mathsf{c},1} \cap \cdots \cap \mathcal{E}_{\mathsf{c},k_{\mathsf{fin}}}$ holds under probability of at least $1 - \frac{1}{3}T^{-2}$.

We complete the proof by union bounding the events $\mathcal{E}_{\mathsf{bound}}$, $\mathcal{E}_{\mathsf{est},1}$, and $\mathcal{E}_{\mathsf{cont}}$. Then, we have that $\mathcal{E}_{\mathsf{success},2} \subseteq \mathcal{E}_{\mathsf{bound}} \cap \mathcal{E}_{\mathsf{est},2} \cap \mathcal{E}_{\mathsf{cont}}$ holds under probability of at least $1 - T^{-2}$.

□

IX. Synthesizing the regret bounds

We use the success events to decompose the expected regret as in [29]: $\mathbf{E}\left[\mathcal{R}_T^{(h)}\right] = R_1^{(h)} + R_2^{(h)} + R_3^{(h)} - T\mathcal{J}(K_\star^{(h)})$, where for $\mathcal{E}_{\mathsf{success}} = \mathcal{E}_{\mathsf{success},1}$ or $\mathcal{E}_{\mathsf{success}} = \mathcal{E}_{\mathsf{success},2}$,

$$R_1^{(h)} = \mathbf{E}\left[\mathbf{1}(\mathcal{E}_{\mathsf{success}}) \sum_{k=2}^{k_{\mathsf{fin}}} J_k^{(h)}\right], \quad R_2^{(h)} = \mathbf{E}\left[\mathbf{1}(\mathcal{E}_{\mathsf{success}}^c) \sum_{t=\tau_1+1}^{T} c_t^{(h)}\right], \quad \text{and} \quad R_3^{(h)} = \mathbf{E}\left[\sum_{t=1}^{\tau_1} c_t^{(h)}\right]. \tag{31}$$

Here, $J_k^{(h)} = \sum_{t=\tau_k}^{\tau_{k+1}-1} c_t^{(h)}$ is the epoch cost and $c_t^{(h)} = x_t^{(h)\top} Q x_t^{(h)} + u_t^{(h)\top} R u_t^{(h)}$ is the stage cost.

The terms $R_2^{(h)}$ and $R_3^{(h)}$ may be bounded directly by invoking Lemmas 20 and 22 of [34] along with the high probability bounds of Lemma VIII.4 and Lemma VIII.5. It therefore remains to bound $R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)})$. This is done for the settings of Theorem III.1 and Theorem III.2 in the following two lemmas.

Lemma IX.1. *In the setting of Theorem III.1, we have*

$$\begin{aligned}
R_1^{(h)} &\le \mathtt{poly}\left(d_{\mathsf{X}}, d_{\mathsf{U}}, P_0^\vee, \Psi_B^\vee, \tau_{\mathsf{warm\,up}}, x_b, \frac{1}{1-\sqrt{2}\rho^N}, d(\hat{\Phi}_0, \Phi_\star)\right) \log^2 T \\
&\quad + \mathtt{poly}(P_0^\vee, \Psi_B^\vee, \sigma)\sqrt{d_\theta d_{\mathsf{U}}}\sqrt{T}\log T \\
&\quad + \mathtt{poly}\left(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, P_0^\vee, \Psi_B^\vee, \theta^\vee, \sigma, \frac{1}{1-\sqrt{2}\rho^N}, N\right) \frac{T^{2/3}}{\sqrt{H}} \log^2(TH).
\end{aligned}$$

Proof. We may invoke Lemma 22 of [34] to show that

$$\begin{aligned}
R_1^{(h)} \le \sum_{k=2}^{k_{\mathsf{fin}}} 142 \left\|P_\star^{(h)}\right\|^8 (\tau_k - \tau_{k-1}) \mathbf{E}\left[\mathbf{1}\left[E_k^{(h)}\right] \left\|\begin{bmatrix}\hat{A}_{k-1}^{(h)} & \hat{B}_{k-1}^{(h)}\end{bmatrix} - \begin{bmatrix}A_\star^{(h)} & B_\star^{(h)}\end{bmatrix}\right\|^2\right] \\
+ (\tau_k - \tau_{k-1})\mathcal{J}^{(h)}(K_\star^{(h)}) + 4(\tau_k - \tau_{k-1})d_{\mathsf{U}} \left\|P_\star^{(h)}\right\| \sigma_k^2 \Psi_{B_\star^{(h)}}^2 + 2x_b \log T \left\|P_\star^{(h)}\right\|.
\end{aligned} \tag{32}$$

where

$$E_k^{(h)} = \left\{ d(\hat{\Phi}_{k-1}, \Phi_\star) \leq \rho^{(k-1)N} d(\hat{\Phi}_0, \Phi_\star) + \frac{C_{\textsf{contract}} \sqrt{P_0^\vee}\sqrt{N}\log(HT)}{(1-\sqrt{2}\rho^N)\sqrt{H\tau_k\sigma_k^2}} \right\}$$
$$\cap \left\{ \left\| \begin{bmatrix} \hat{A}_{k-1}^{(h)} & \hat{B}_{k-1}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq C_{\textsf{est},1} \frac{\sigma^2 d_\theta \left\| P_{K_0}^{(h)} \right\|}{\tau_{k-1}\sigma_{k-1}^2} \log T + \frac{\beta_1 d(\hat{\Phi}_{k-1}, \Phi_\star)^2}{\sigma_k^2} \right\}$$

is the event bounding the norm of the dynamics error in terms of the misspecification as well as the misspecification in terms of the amount of data. Under the event $E_k^{(h)}$, we have

$$\mathbf{E}\left[\mathbf{1}\left[E_k^{(h)}\right] \left\| \begin{bmatrix} \hat{A}_{k-1}^{(h)} & \hat{B}_{k-1}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star(h) \end{bmatrix} \right\|^2 \right]$$
$$\leq C_{\textsf{est},1} \frac{\sigma^2 d_\theta \left\| P_{K_0}^{(h)} \right\|}{\tau_{k-1}\sigma_{k-1}^2} \log T + \frac{2\beta_1 \rho^{2(k-1)N} d(\hat{\Phi}_0, \Phi_\star)^2}{\sigma_{k-1}^2} + 2\beta_1 \frac{C_{\textsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 H \tau_{k-1}\sigma_{k-1}^4}.$$

Substituting the above inequality into (32), we find

$$R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)})$$
$$\lesssim \sum_{k=2}^{k_{\textsf{fin}}} \left(\left\| P_\star^{(h)} \right\|^8 \tau_{k-1} \left(\frac{\sigma^2 d_\theta \left\| P_{K_0}^{(h)} \right\|}{\tau_{k-1}\sigma_{k-1}^2} \log T + \frac{\beta_1 \rho^{2(k-1)N} d(\hat{\Phi}_0, \Phi_\star)^2}{\sigma_{k-1}^2} + \beta_1 \frac{C_{\textsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 H \tau_{k-1}\sigma_{k-1}^4} \right) \right)$$
$$+ \tau_{k-1} d_{\textsf{U}} \left\| P_\star^{(h)} \right\| \sigma_k^2 \Psi_{B_\star^{(h)}}^2 + x_b \log T \left\| P_\star^{(h)} \right\|.$$

Substituting in our choice of σ_k^2 from (7), we find

$$R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)})$$
$$\lesssim \sum_{k=2}^{k_{\textsf{fin}}} \sigma^2 \sqrt{d_\theta d_{\textsf{U}}} \left\| P_{K_0^{(h)}}^{(h)} \right\|^9 \Psi_{B_\star^{(h)}}^2 \sqrt{\tau_{k-1}} \log T + \frac{\left(\left\| P_{K_0^{(h)}}^{(h)} \right\|^8 \beta_1 C_{\textsf{contract}}^2 P_0^\vee N \log^2(HT) + d_{\textsf{U}} \left\| P_{K_0^{(h)}}^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right)}{(1-\sqrt{2}\rho^N)\sqrt{H}} \tau_{k-1}^{2/3}$$
$$+ \left(\beta_1 \left\| P_{K_0^{(h)}}^{(h)} \right\|^8 + d_{\textsf{U}} \left\| P_{K_0^{(h)}}^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right) \tau_{k-1} \rho^{(k-1)N} d(\Phi_0, \Phi_\star) + x_b \log T \left\| P_\star^{(h)} \right\|$$
$$\lesssim \sigma^2 \sqrt{d_\theta d_{\textsf{U}}} \left\| P_{K_0^{(h)}}^{(h)} \right\|^9 \Psi_{B_\star^{(h)}}^2 \sqrt{T} \log T + \frac{\left(\beta_1 \left\| P_{K_0^{(h)}}^{(h)} \right\|^8 C_{\textsf{contract}}^2 P_0^\vee N \log^2(HT) + d_{\textsf{U}} \left\| P_{K_0^{(h)}}^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right)}{(1-\sqrt{2}\rho^N)} \frac{T^{2/3}}{\sqrt{H}}$$
$$+ \frac{\left(\beta_1 \left\| P_{K_0^{(h)}}^{(h)} \right\|^8 + d_{\textsf{U}} \left\| P_{K_0^{(h)}}^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right) \tau_1}{1-\sqrt{2}\rho^N} d(\Phi_0, \Phi_\star) + x_b \log^2 T \left\| P_{K_0^{(h)}}^{(h)} \right\|.$$

The result now follows by substituting in the definition of τ_1 from Theorem III.1, of β_1 from Assumption III.2, and of $C_{\textsf{contract}}$ from Theorem VII.2. □

Lemma IX.2. *In the setting of Theorem III.2, we have*

$$R_1^{(h)} \leq \texttt{poly}\left(\sigma, d_\theta, d_{\textsf{U}}, d_{\textsf{X}}, \frac{1}{\alpha}, \frac{1}{1-\sqrt{2}\rho^N}, \Psi_B^\vee, P_0^\vee, x_b, \tau_{\textsf{warm up}}, d(\hat{\Phi}_0, \Phi_\star) \right) \log^2 T$$
$$+ \texttt{poly}\left(\sigma, d_\theta, d_{\textsf{U}}, d_{\textsf{X}}, \frac{1}{\alpha}, \frac{1}{1-\sqrt{2}\rho^N}, \Psi_B^\vee, P_0^\vee, x_b, N \right) \frac{\sqrt{T}}{\sqrt{H}} \log^2(TH).$$

Proof. We again invoke Lemma 22 of [34] to show that (32) holds in this setting, where the event bounding the norm of the dynamics error is now given by

$$E_k^{(h)} = \left\{ d(\hat{\Phi}_{k-1}, \Phi_\star) \leq \rho^{(k-1)N} d(\hat{\Phi}_0, \Phi_\star) + \frac{C_{\textsf{contract}} \sqrt{P_0^\vee}\sqrt{N}\log(HT)}{(1-\sqrt{2}\rho^N)\sqrt{H\tau_k\sigma_k^2}} \right\}$$

$$\cap \left\{ \left\| \begin{bmatrix} \hat{A}_k^{(h)} & \hat{B}_k^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star^{(h)} \end{bmatrix} \right\|_F^2 \leq C_{\mathsf{est},2} \frac{\sigma^2 d_\theta}{\tau_k \alpha^2} \log T + \beta_2 d(\hat{\Phi}_k, \Phi_\star)^2 \right\}.$$

Under this event, we have

$$\begin{aligned}
&\mathbf{E}\left[\mathbf{1}\left[E_k^{(h)}\right] \left\| \begin{bmatrix} \hat{A}_{k-1}^{(h)} & \hat{B}_{k-1}^{(h)} \end{bmatrix} - \begin{bmatrix} A_\star^{(h)} & B_\star(h) \end{bmatrix} \right\|^2 \right] \\
&\leq C_{\mathsf{est},2} \frac{\sigma^2 d_\theta}{\tau_{k-1} \alpha^2} \log T + \frac{2\beta_2 \rho^{2(k-1)N} d(\hat{\Phi}_0, \Phi_\star)^2}{\sigma_{k-1}^2} + 2\beta_2 \frac{C_{\mathsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 H \tau_{k-1} \sigma_{k-1}^2}.
\end{aligned}$$

Substituting the above inequality into (32), we have

$$\begin{aligned}
&R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)}) \\
&\lesssim \sum_{k=2}^{k_{\mathsf{fin}}} \left\| P_\star^{(h)} \right\|^8 \tau_{k-1} \left(\frac{\sigma^2 d_\theta}{\tau_{k-1}\alpha^2} \log T + \frac{\beta_2 \rho^{2(k-1)N} d(\hat{\Phi}_0, \Phi_\star)^2}{\sigma_{k-1}^2} + \beta_2 \frac{C_{\mathsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 H \tau_{k-1} \sigma_{k-1}^2} \right) \\
&+ \tau_{k-1} d_{\mathsf{U}} \left\| P_\star^{(h)} \right\| \sigma_k^2 \Psi_{B_\star^{(h)}}^2 + x_b \log T \left\| P_\star^{(h)} \right\|.
\end{aligned}$$

Substituting our choice of σ_k^2 from (8), we have

$$\begin{aligned}
R_1^{(h)} - T\mathcal{J}(K_\star^{(h)}) &\lesssim \sum_{k=2}^{k_{\mathsf{fin}}} \frac{\left\| P_\star^{(h)} \right\|^8 \sigma^2 d_\theta}{\alpha^2} \log T + \left(\beta_2 \left\| P_\star^{(h)} \right\|^8 + d_{\mathsf{U}} \left\| P_\star^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right) \tau_{k-1} \rho^{(k-1)N} d(\hat{\Phi}_0, \Phi_\star) \\
&+ \left(\beta_2 \left\| P_\star^{(h)} \right\|^8 \frac{C_{\mathsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2 \sqrt{H}} + \frac{d_{\mathsf{U}} \left\| P_\star^{(h)} \right\| \Psi_{B_\star^{(h)}}^2}{\sqrt{H}} \right) \sqrt{\tau_{k-1}} \\
&+ x_b \log T \left\| P_\star^{(h)} \right\| \\
&\lesssim \frac{\left\| P_\star^{(h)} \right\|^8 \sigma^2 d_\theta}{\alpha^2} \log^2 T + \frac{\tau_1 \left(\beta_2 \left\| P_\star^{(h)} \right\|^8 + d_{\mathsf{U}} \left\| P_\star^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right) d(\hat{\Phi}_0, \Phi_\star)}{1-\sqrt{2}\rho^N} + x_b \log^2 T \left\| P_\star^{(h)} \right\| \\
&+ \left(\beta_2 \left\| P_\star^{(h)} \right\|^8 \frac{C_{\mathsf{contract}}^2 P_0^\vee N \log^2(HT)}{(1-\sqrt{2}\rho^N)^2} + d_{\mathsf{U}} \left\| P_\star^{(h)} \right\| \Psi_{B_\star^{(h)}}^2 \right) \frac{\sqrt{T}}{\sqrt{H}}.
\end{aligned}$$

We conclude by substituting β_2 from Assumption III.4, C_{contract} from Theorem VII.2, and τ_1 from Theorem III.2. □

With these lemmas in hand, we are now ready to prove the main results.

1) Proof of Theorem III.1:

Proof. It follows from Lemma 19 of [34] that

$$R_3^{(h)} \leq 3\tau_1 \max\{d_{\mathsf{X}}, d_{\mathsf{U}}\} \left\| P_{K_0^{(h)}} \right\| \Psi_{B_\star^{(h)}}^2. \tag{33}$$

The second term, $R_2^{(h)}$ may be bounded by using the fact that the state is bounded up until a failure situation is reached, and after that failure situation, the initial stabilizing controller is played. In the probability $1 - T^{-2}$, we have from Lemma 20 of [34] that

$$R_2^{(h)} \leq T^{-1} \log(T) \mathtt{poly}(\sigma, d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, x_b, K_b, \|Q\|, \theta^\vee, P_0^\vee, \Psi_B^\vee) + \sum_{k=1}^{k_{\mathsf{fin}}} 2(\tau_k - \tau_{k-1}) d_{\mathsf{U}} \sigma_k^2. \tag{34}$$

By substituting the choice of σ_k^2 from Theorem III.1 into the above inequality, and invoking Lemma IX.1, we find that

$$\begin{aligned}
\mathcal{R}_T^{(h)} &= R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)}) + R_2^{(h)} + R_3^{(h)} \\
&\leq \mathtt{poly}\left(\sigma, d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, x_b, K_b, \|Q\|, \theta^\vee, P_0^\vee, \Psi_B^\vee, \tau_{\mathsf{warm\ up}}, x_b, \frac{1}{1-\sqrt{2}\rho^N}, d(\hat{\Phi}_0, \Phi_\star) \right) \log^2 T \\
&+ \mathtt{poly}(P_0^\vee, \Psi_B^\vee, \sigma) \sqrt{d_\theta d_{\mathsf{U}}} \sqrt{T} \log^2 T \\
&+ \mathtt{poly}\left(d_{\mathsf{X}}, d_{\mathsf{U}}, d_\theta, P_0^\vee, \Psi_B^\vee, \theta^\vee, \sigma, \frac{1}{1-\sqrt{2}\rho^N}, N \right) \frac{T^{2/3}}{\sqrt{H}} \log^2(TH).
\end{aligned}$$

□

2) Proof of Theorem III.2:

Proof. We may again invoke Lemma 19 and 20 of [34] to show that (34) and (33) hold. Subsituting the choice of σ_k^2 from Theorem III.2 into (34), and invoking Lemma IX.2, we find

$$\begin{aligned}\mathcal{R}_T^{(h)} &= R_1^{(h)} - T\mathcal{J}^{(h)}(K_\star^{(h)}) + R_2^{(h)} + R_3^{(h)} \\ &\leq \texttt{poly}\left(\sigma, d_\theta, d_{\mathsf{U}}, d_{\mathsf{X}}, \frac{1}{\alpha}, \frac{1}{1-\sqrt{2}\rho^N}, \Psi_B^\vee, P_0^\vee, x_b, K_b, \theta^\vee, \left\|Q\right\|, \tau_{\textsf{warm up}}, d(\hat{\Phi}_0, \Phi_\star)\right) \log^2 T \\ &+ \texttt{poly}\left(\sigma, d_\theta, d_{\mathsf{U}}, d_{\mathsf{X}}, \frac{1}{\alpha}, \frac{1}{1-\sqrt{2}\rho^N}, \Psi_B^\vee, P_0^\vee, x_b, N\right) \frac{\sqrt{T}}{\sqrt{H}} \log^2(TH).\end{aligned}$$

□