Multi-Task Imitation Learning for Linear Dynamical Systems

Thomas T. Zhang*[1], Katie Kang*[2], Bruce D. Lee*[1], Claire Tomlin[2], Sergey Levine[2], Stephen Tu[3], and Nikolai Matni[1,3]

[1]University of Pennsylvania
[2]University of California, Berkeley
[3]Google Research, Brain Team

Abstract

We study representation learning for efficient imitation learning over linear systems. In particular, we consider a setting where learning is split into two phases: (a) a pre-training step where a shared k-dimensional representation is learned from H source policies, and (b) a target policy fine-tuning step where the learned representation is used to parameterize the policy class. We find that the imitation gap over trajectories generated by the learned target policy is bounded by $\tilde{O}\left(\frac{kn_x}{HN_{\mathrm{shared}}} + \frac{kn_u}{N_{\mathrm{target}}}\right)$, where $n_x > k$ is the state dimension, n_u is the input dimension, N_{shared} denotes the total amount of data collected for each policy during representation learning, and N_{target} is the amount of target task data. This result formalizes the intuition that aggregating data across related tasks to learn a representation can significantly improve the sample efficiency of learning a target task. The trends suggested by this bound are corroborated in simulation.

Keywords: Imitation learning, transfer learning, multi-task learning, representation learning

1 Introduction

Imitation learning (IL), which learns control policies by imitating expert demonstrations, has demonstrated success across a variety of domains including self-driving cars (Codevilla et al., 2018) and robotics (Schaal, 1999). However, using IL to learn a robust behavior policy may require a large amount of training data (Ross et al., 2011), and expert demonstrations are often expensive to collect. One remedy for this problem is multi-task learning: using data from other tasks (source tasks) in addition to from the task of interest (target task) to jointly learn a policy. We study the application of multi-task learning to IL over linear systems, and demonstrate improved sample efficiency when learning a controller via representation learning.

Our results expand on prior work that studies multi-task representation learning for supervised learning (Du et al., 2020; Tripuraneni et al., 2021), addressing the new challenges that arise in the imitation learning setting. First, the data for IL is temporally dependent, as it is generated from a dynamical system $x[t+1] = f(x[t], u[t], w[t])$. In contrast, the supervised learning setting assumes that both the train and test data are independent and identically distributed (i.i.d.) from the same underlying distribution. Furthermore, we are interested in the performance of the learned controller in closed-loop rather than its error on expert-controlled trajectories. Hence, bounds on excess risk, which corresponds to the one-step prediction error

*Authors contributed equally to this work.

of the learned controller under the expert distribution, are not immediately informative for the closed-loop performance. We instead focus our analysis on the tracking error between the learned and expert policies, which requires us to account for the distribution shift between the learned and expert controllers.

We address these challenges in the setting of IL for linear systems. The following statement captures the benefits of multi-task representation learning on sample complexity:

Theorem 1.1 (main result, informal) *Suppose that the source task controllers are sufficiently related to the target task controller. Then, the tracking error between the learned target controller and the corresponding expert is bounded with high probability by:*

$$\textit{tracking error} \lesssim \frac{\textit{rep. dimension } \times \textit{ state dimension}}{\textit{\# source task datapoints}} + \frac{\textit{rep. dimension } \times \textit{ input dimension}}{\textit{\# target task datapoints}}.$$

The first term in this bound corresponds to the error from learning a common representation, and the second term the error in fitting the remaining weights of the target task controller. The key upshot of this result is that the numerator of the second term (rep. dimension $\times$ input dimension) is smaller than the number of parameters (input dimension $\times$ state dimension) in the target controller. This demonstrates an improvement in sample complexity of multi-task IL over direct IL, where the error scales as $\frac{\#\text{parameters}}{\#\text{datapoints}}$. Furthermore, we note that the error in learning the representation decays along all axes of the data: # of tasks $\times$ # of trajs $\times$ traj length for source tasks, and # of trajs $\times$ traj length for the target task. It is non-trivial to demonstrate that the error decays with the trajectory length, and doing so requires tools that handle causally dependent data in our analysis.

The remainder of the paper formulates the multi-task IL problem, and the assumptions required to prove Theorem 1.1. The main contributions may be summarized as follows:

- We provide novel interpretable notions of source task overlap with the target task (§2 and §3).
- We bound the imitation gap achieved by multi-task IL as in Theorem 1.1 (§3).
- We empirically show the efficacy of multi-task IL when the assumptions are satisfied (§4).

1.1 Related Work

Multi-task imitation and reinforcement learning: Multi-task RL and IL methods seek to represent policies solving different tasks with shared parameters, enabling the transfer of knowledge across related tasks (Teh et al., 2017; Espeholt et al., 2018; Hessel et al., 2018; Singh et al., 2020; Deisenroth et al., 2014), and rapid test-time adaptation to new tasks (Finn et al., 2017; Rakelly et al., 2019; Duan et al., 2016; Yu et al., 2021; Yang and Nachum, 2021). There also exists a body of work which theoretically analyses the sample complexity of representation learning in multi-task RL and IL (Lu et al., 2021; Cheng et al., 2022; Lu et al., 2022; Xu et al., 2020; Maurer et al., 2015; Arora et al., 2020). While this line of work considers a more general MDP setting compared with the linear dynamical systems we consider, the specific results are often stated with incompatible assumptions (such as bounded states/cost functions and discrete action spaces), and/or do not reflect how system-theoretic properties such as closed-loop task stability affect the final rates.

Multi-task system identification and adaptive control: Recent work has also considered applications of multi-task learning where the dynamics change between tasks, and the goal is to perform adaptive control (Harrison et al., 2018; Richards et al., 2021, 2022; Shi et al., 2021; Muthirayan et al., 2022) or dynamics forecasting (Wang et al., 2021). Multi-task system identification (Modi et al., 2021) and stabilization using data from related systems (Li et al., 2022) have also been considered. Our work instead studies the problem of learning to imitate different expert controllers while the system remains the same, and demonstrates bounds on the tracking error between the learned controller and its corresponding expert.

Sample complexity of multi-task learning: Numerous works have studied the sample efficiency gains of multi-task learning for regression and classification under various task similarity assumptions (Baxter, 1995; Crammer et al., 2008; Maurer et al., 2016; Tripuraneni et al., 2020; Chua et al., 2021). Most closely related to our results are Du et al. (2020) and Tripuraneni et al. (2021), both of which show multi-task representation learning sample complexity bounds in the linear regression setting in which the error from learning the representation decays with the total number of source training samples. Our work leverages these results to tackle the setting of linear imitation learning, where have the additional challenges of non-i.i.d. data and test time distribution shift.

2 Problem Formulation

2.1 Multi-Task Imitation Learning

Imitation learning uses state/action pairs $(x, u) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$ of expert demonstrations to learn a controller $\hat{\pi} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$, by matching the learned controller actions to the expert actions. In particular, if $\mathcal{D}$ is the training set of expert state/action pairs, then $\hat{\pi} \in \operatorname{argmin}_\pi \sum_{(x,u)\in\mathcal{D}} \|\pi(x) - u\|^2$.

We are interested in the problem of *multi-task* imitation learning, where we consider $H + 1$ different expert controllers. We call the first H controllers *source controllers* and the $(H + 1)^{\text{st}}$ controller the *target controller*. We assume that we have access to N_1 trajectories for each source task, and $N_2 \leq N_1$ trajectories for the target task. For simplicity, we assume all trajectories are of the same length T. In particular, for each source task $h \in \{1, \dots, H\}$, our source data consists of $\left\{\left\{(x_i^{(h)}[t], u_i^{(h)}[t])\right\}_{t=0}^{T-1}\right\}_{i=1}^{N_1}$, while our target data consists of $\left\{\left\{(x_i^{(H+1)}[t], u_i^{(H+1)}[t])\right\}_{t=0}^{T-1}\right\}_{i=1}^{N_2}$. Our goal is to learn a controller which effectively imitates the target controller. However, because we only have access to a small number (N_2) of target expert trajectories, we leverage the HN_1 expert trajectories from the source controllers to accelerate the learning of the target controller.

To do so, we break our training into two stages: a pre-training stage which learns from the combined source task data, and a target training stage which only learns from the target task data. In the pre-training stage, we extract a common, low dimensional representation for the source controllers, which is used later in the target training stage. More specifically, we learn a common, low dimensional representation mapping $\hat{\phi} : \mathbb{R}^{n_x} \to \mathbb{R}^k$, where $k < n_x$ is the dimension of the representation, and linear predictors $\hat{F}^{(h)} \in \mathbb{R}^{n_u \times k}$ unique to each task:

$$\hat{\phi}, \hat{F}^{(1)}, \dots, \hat{F}^{(H)} \in \operatorname*{argmin}_{\phi, F^{(1)}, \dots, F^{(H)}} \sum_{h=1}^{H} \sum_{i=1}^{N_1} \sum_{t=0}^{T-1} \left\| F^{(h)} \phi(x_i^{(h)}[t]) - u_i^{(h)}[t] \right\|^2. \tag{1}$$

We do not address the details of solving the empirical risk minimization problem, and instead perform our analysis assuming (1) can be solved to optimality. Note however that Tripuraneni et al. (2021) demonstrate in the linear regression setting that a method-of-moments-based algorithm can efficiently find approximate empirical risk minimizers.

Once a common representation $\hat{\phi}$ is obtained, we move on to target task training. During target task training, we use the common representation mapping $\hat{\phi}$ learned from the pre-training step to map the states into the lower dimensional representation, and learn an additional linear predictor $\hat{F}^{(H+1)}$ unique to the target task to model the target controller:

$$\hat{F}^{(H+1)} = \underset{F}{\operatorname{argmin}} \sum_{i=1}^{N_2} \sum_{t=0}^{T-1} \left\| F\hat{\phi}\left(x_i^{(H+1)}[t]\right) - u_i^{(H+1)}[t] \right\|^2. \tag{2}$$

Since the representation $\hat{\phi}$ is fixed from pre-training, (2) is an ordinary least squares problem.

2.2 System and Data Assumptions

We focus our analysis on a linear systems setting, with state $x[t] \in \mathbb{R}^{n_x}$, input $u[t] \in \mathbb{R}^{n_u}$, and Gaussian process noise $w[t] \in \mathbb{R}^{n_x}$ obeying dynamics

$$x[t+1] = Ax[t] + Bu[t] + w[t]. \tag{3}$$

Let each expert controller be of the form $u[t] = K^{(h)}x[t] + z[t]$, where $z[t] \in \mathbb{R}^{n_u}$ is Gaussian actuator noise.[1] We assume the system matrices (A, B) remain the same between tasks, but the process noise covariance and the controllers may change.[2] In particular, we have $w^{(h)}[t] \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_w^{(h)})$ and $z^{(h)}[t] \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_z^2 I)$ with $\Sigma_w^{(h)} \succ 0$ for all $h \in [H+1]$ and $\sigma_z^2 > 0$.

We assume all of the expert controllers $K^{(h)}$ are stabilizing, i.e., the spectral radii of $A + BK^{(h)}$ are less than one. Note that this implies that (A, B) is stabilizable. No other assumptions on (A, B) are required. The state distribution of the system under each expert controller will converge to the stationary distribution $\mathcal{N}(0, \Sigma_x^{(h)})$, where $\Sigma_x^{(h)}$ solves the following discrete Lyapunov equation:

$$\Sigma_x^{(h)} = (A + BK^{(h)})\Sigma_x^{(h)}(A + BK^{(h)})^\top + \sigma_z^2 BB^\top + \Sigma_w^{(h)}.$$

For simplicity, we assume that the initial states of the expert demonstrations in our datasets are sampled $x_i^{(h)}[0] \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_x^{(h)})$. Thus at all times, the marginal state distributions of the expert demonstrations are equal to $\mathcal{N}(0, \Sigma_x^{(h)})$.[3]

Finally, we assume that the expert controllers share a low dimensional representation. Specifically, there exists some $\Phi_\star \in \mathbb{R}^{k \times n_x}$ with $2k \leq n_x$[4] and weights $F_\star^{(1)}, F_\star^{(2)}, \ldots, F_\star^{(H+1)} \in \mathbb{R}^{n_u \times k}$ such that for all $h \in [H+1]$, $K^{(h)} = F_\star^{(h)}\Phi_\star$, and the action taken at time t for trajectory i is:[5]

$$u_i^{(h)}[t] = F_\star^{(h)}\Phi_\star x_i^{(h)}[t] + z_i^{(h)}[t].$$

Under this assumption, the learned common representation $\hat{\phi}$ in Section 2.1 can be restricted to linear representations, i.e., $\hat{\phi}(x) = \hat{\Phi}x$, where $\hat{\Phi} \in \mathbb{R}^{k \times n_x}$. Note that solving Problem (2) with $\hat{\Phi}$ fixed involves solving

[1] As the control actions are the labels in IL, actuator noise corresponds to label noise in supervised learning. In the absence of such noise, the controller is recovered by n_x linearly independent states and corresponding expert inputs.

[2] This could be the case, for instance, if different controllers are designed for different levels of noise.

[3] This assumption is not restrictive, as stable systems exponentially converge to stationarity.

[4] While this assumption is more stringent than the intuitive $k < n_x$ assumption, it arises from the fact that the residual of the stacked source controllers may be of rank $2k$.

[5] An example of a setting where expert controllers satisfy this assumption is when the system has high dimensional states which exhibit low dimensional structure, e.g. when A and B can be decomposed into $A = \Phi_\star^\dagger \tilde{A}\Phi_\star$ and $B = \Phi_\star^\dagger \tilde{B}$, where $\tilde{A} \in \mathbb{R}^{k \times k}$ and $\tilde{B} \in \mathbb{R}^{k \times n_u}$. Here, linear policies K which optimize some objective in terms of the low dimensional features of the system can be decomposed into $K = \tilde{K}\Phi_\star$, where $\tilde{K} \in \mathbb{R}^{n_u \times k}$, mirroring the assumptions of our expert controllers. We provide a concrete example in Section 4.

for only kn_u parameters, which is smaller than the $n_u n_x$ unknown parameters when learning from scratch. In particular, by representing the controller as $F^{(H+1)}\Phi$, we have $k(n_u + n_x)$ unknown parameters: kn_x of the parameters are, however, learned using the source task data, leaving only kn_u parameters to learn with target task data.

2.3 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 $\rho(A)$. We use $\dagger$ to denote the Moore-Penrose pseudo-inverse. For a square matrix A with $\rho(A) < 1$, define $\mathcal{J}(A) = \sum_{t\geq 0}\left\|A^t\right\| < \infty$. A symmetric, positive semi-definite (psd) matrix $A = A^\top$ is denoted $A \succeq 0$. Similarly $A \succeq B$ denotes that $A - B$ is positive semidefinite. The condition number of a positive definite matrix A is denoted $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$, where $\lambda_{\max}$ and $\lambda_{\min}$ denote the maximum and minimum eigenvalues, respectively. Similarly, $\sigma_i(A)$ denote the singular values of A. We denote the normal distribution with mean μ and covariance Σ by $\mathcal{N}(\mu, \Sigma)$. We use standard $\mathcal{O}(\cdot)$, $\Theta(\cdot)$ and $\Omega(\cdot)$ to omit universal constant factors, and $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Theta}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to also omit polylog factors. We also use $a \lesssim b$ to denote $a = O(b)$. We use the indexing shorthand $[K] := \{1, \dots, K\}$. For a given task $h \in [H+1]$, the matrix of stacked states is defined as

$$\mathbf{X}^{(h)} = \begin{bmatrix} x_1^{(h)}[0] & \dots & x_1^{(h)}[T-1] & \dots & x_{N_1}^{(h)}[0] & \dots & x_{N_1}^{(h)}[T-1] \end{bmatrix}^\top \in \mathbb{R}^{N_1 T \times n_x}. \tag{4}$$

Lastly, let $\bar{\lambda} = \max_{1\leq h\leq H} \lambda_{\max}(\Sigma_x^{(h)})$ and $\underline{\lambda} = \min_{1\leq h\leq H} \lambda_{\min}(\Sigma_x^{(h)})$.

3 Sample Complexity of Multi-Task Imitation Learning

In order to derive any useful information from source tasks for a downstream task, the source tasks must satisfy some notion of *task diversity* that sufficiently covers the downstream task. To that end, we introduce the following notions of source tasks covering the target task.

Definition 3.1 (target task covariance coverage (Du et al., 2020)) *Define the constant c as:*

$$c := \min_{h\in[H]} \lambda_{\min}\big((\Sigma_x^{(H+1)})^{-1/2}\Sigma_x^{(h)}(\Sigma_x^{(H+1)})^{-1/2}\big). \tag{5}$$

Note that c is well-defined and positive by our assumption that $\Sigma_w^{(h)} \succ 0$ for all $h \in [H+1]$.

Definition 3.1 captures the degree to which the closed-loop distribution of states for each source task aligns with that of the target task. We then introduce the following notion of task similarity between the source and target task weights, which generalizes the well-conditioning assumptions in Du et al. (2020) and Tripuraneni et al. (2021).

Assumption 3.1 (diverse source controllers) *We assume the target task weights $F_\star^{(H+1)}$ and the stacked source task weights $\mathbf{F}_\star := \begin{bmatrix} \left(F_\star^{(1)}\right)^\top & \cdots & \left(F_\star^{(H)}\right)^\top \end{bmatrix} \in \mathbb{R}^{k\times n_u H}$ satisfy*

$$\left\|\left(\mathbf{F}_\star^\dagger\right)^\top F^{(H+1)}\right\|^2 \leq \mathcal{O}\left(\frac{1}{H}\right). \tag{6}$$

Assumption 3.1 states that the alignment and loadings of the singular spaces between the stacked source task weights and target task weights closely match along the low-dimensional representation dimension. For example, if $F_\star^{(h)} = F_\star^{(H+1)}$ for each $h \in [H]$, the RHS of (6) is $1/H$. We note that this assumption subsumes and is more geometrically informative than a direct bound on the ratio of singular values, e.g. $\sigma^2_{\max}(F_\star^{(H+1)})/\sigma_k^2(\mathbf{F}_\star) \leq \mathcal{O}(1/H)$, which would follow by naively extending the well-conditioning assumptions in Du et al. (2020) and Tripuraneni et al. (2021). Notably, such a condition might not be satisfied even if $F_\star^{(h)} = F_\star^{(H+1)}, \forall h \in [H]$, e.g., if $F_\star^{(H+1)}$ is rank-deficient.

3.1 Excess Risk Bound: Generalization Along Expert Target Task Trajectories

First we show that learning controllers through multi-task representation learning leads to favorable generalization bounds on the excess risk of the learned controller inputs on the expert target task state distribution, analogous to the bounds on multi-task linear regression in Du et al. (2020); Tripuraneni et al. (2021). However, a key complicating factor in our setting is the fact that the input noise $z^{(h)}[t]$ enters the process, and thus the data $x^{(h)}[t]$ is causally dependent on the "label noise". In order to overcome this issue and preserve our statistical gains along time T, we leverage the theory of self-normalized martingales, in particular generalizing tools from Abbasi-Yadkori et al. (2011) to the matrix-valued setting. The full argument is detailed in Appendix A and Appendix B. This culminates in the following target task excess risk bound.

Theorem 3.1 (target task excess risk bound) *Given $\delta \in (0,1)$, suppose that*

$$N_1 T \gtrsim \max_{h \in [H]} \mathcal{J}\left(A + BK^{(h)}\right)^2 \kappa\left(\Sigma_x^{(h)}\right)(n_x + \log(H/\delta)),$$
$$N_2 T \gtrsim \mathcal{J}\left(A + BK^{(H+1)}\right)^2 \kappa\left(\Sigma_x^{(H+1)}\right)(k + \log(1/\delta)).$$

Define $\mathcal{P}_{0:T-1}^{(H+1)}$ as the distribution over target task trajectories $(x^{(H+1)}[0], \cdots, x^{(H+1)}[T-1])$. Then with probability at least $1-\delta$, the excess risk of the learned representation $\hat{\Phi}$ and target task weights $\hat{F}^{(H+1)}$ is bounded by

$$\begin{aligned}\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) &:= \frac{1}{2T}\mathbb{E}_{\mathcal{P}_{0:T-1}^{(H+1)}}\left[\sum_{t=0}^{T-1}\left\|(F_\star^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi})x^{(H+1)}[t]\right\|^2\right] \\ &\lesssim \sigma_z^2\left(\frac{kn_x \log\left(N_1 T\frac{\bar{\lambda}}{\underline{\lambda}}\right)}{cN_1TH} + \frac{kn_u + \log(\frac{1}{\delta})}{N_2 T}\right). \end{aligned} \tag{7}$$

Note that when we are operating in the setting where we have much more source data than target data, the second term limits the excess risk bound in (7). The second term scales with kn_u, which is smaller than the number of total parameters in the controller $n_u n_x$, or $k(n_u + n_x)$ under the assumption of a low rank (rank-k) controller. Therefore, the benefit of multi-task learning exhibited by this bound is most clear in the setting of underactuation, i.e., when $n_u \leq n_x$. It should also be noted that the quantity kn_x in the numerator of the first term will only be smaller than the number of source controller parameters ($n_x n_u H$) if k is much smaller than $n_u H$. This is reasonable, because if $k \geq n_u H$, an optimal representation could simply contain all of the source task controllers.

3.2 Closed-Loop Guarantees: Tackling Distribution Shift

We show that using multi-task representation learning leads to favorable generalization bounds of the performance of the learned target controller in closed-loop. As we are studying the pure offline imitation learning ("behavioral cloning") setting, we do not assume that the expert controllers are optimizing any particular objective. Therefore, to quantify the performance of the controller, we bound the deviation of states generated by the learned and expert target controller run in closed-loop, i.e., the *tracking error*, which implies general expected-cost bounds.

In order to transfer a bound on the excess risk of the target task $\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})$ into a bound on the tracking error, we must account for the fundamental distribution shift between the expert trajectories seen during training and the trajectories generated by running the learned controller in closed-loop. We leverage the recent framework of Pfrommer et al. (2022) to bound the tracking error, making the necessary modifications to handle stochasticity. Our bound formalizes the notion that "low training error implies low test error," even under the aforementioned distribution shift. A detailed exposition can be found in Appendix C.

Let us define the following coupling of the states of the expert versus learned target task closed-loop systems: given a learned controller $\hat{K} = \hat{F}^{(H+1)}\hat{\Phi}$ from solving the pre-training and fine-tuning optimization problems (1) and (2), for a realization of process randomness $x[0] \sim \mathcal{N}(0, \Sigma_x^{(H+1)})$ and $z[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_z^2 I)$, $w[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w^{(H+1)})$ for $t = 0, \dots, T-1$, we write

$$\begin{aligned} x_\star[t+1] &= (A + BK^{(H+1)})x_\star[t] + Bz[t] + w[t], \quad x_\star[0] = x[0], \\ \hat{x}[t+1] &= (A + B\hat{K})\hat{x}[t] + Bz[t] + w[t], \quad \hat{x}[0] = x[0]. \end{aligned}$$

Thus $\hat{x}[t]$ and $x_\star[t]$ are the states visited by the learned and expert target task systems with the *same* draw of process randomness. We show a high probability bound on the closed-loop tracking error $\|x_\star[t] - \hat{x}[t]\|$ that scales with the excess risk of the learned controller. Denote by $\mathcal{P}^\star_{1:T}$ and $\hat{\mathcal{P}}_{1:T}$ the distributions of trajectories $\{x_\star[t]\}_{t=1}^T$ and $\{\hat{x}[t]\}_{t=1}^T$.

Theorem 3.2 (Target task tracking error bound) *Let $(\hat{\Phi}, \hat{F}^{(H+1)})$ denote the learned representation and target task weights, and $\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})$ denote the corresponding excess risk. Define $A_{\mathsf{cl}} := A + BK^{(H+1)}$. Assume that the excess risk satisfies:*

$$\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) \lesssim \frac{\lambda_{\min}\left(\Sigma_x^{(H+1)}\right)}{\mathcal{J}(A_{\mathsf{cl}})^2 \|B\|^2}. \tag{8}$$

Then with probability greater than $1-\delta$, for a new target task trajectory sampled with process randomness $x[0] \sim \mathcal{N}(0, \Sigma_x^{(H+1)})$ and $z[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_z^2 I)$, $w[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w^{(H+1)})$ for $t = 0, \dots, T-1$, the tracking error satisfies

$$\max_{1 \le t \le T} \|\hat{x}[t] - x_\star[t]\|^2 \lesssim \mathcal{J}(A_{\mathsf{cl}})^2 \|B\|^2 \log\left(\frac{T}{\delta}\right) \mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}). \tag{9}$$

Furthermore, for any cost function $h(\cdot)$ that is L-Lipschitz with respect to the trajectory-wise metric $d(\vec{x}_{1:T}, \vec{y}_{1:T}) = \max_{1 \le t \le T} \|x[t] - y[t]\|$, we have the following bound on the expected cost gap

$$\left| \mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right] - \mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})] \right| \lesssim L\mathcal{J}(A_{\mathsf{cl}}) \|B\| \sqrt{\log T} \sqrt{\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})} \tag{10}$$

By invoking the bound on the excess risk from Theorem 3.1, condition (8) is satisfied with probability at least $1-\delta'$ if we have sufficiently many samples H, T, N_1, N_2 such that

$$\sigma_z^2\left(\frac{kn_x\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right)}{cN_1TH}+\frac{kn_u+\log(\frac{1}{\delta'})}{N_2T}\right)\lesssim\frac{\lambda_{\min}\left(\Sigma_x^{(H+1)}\right)}{\mathcal{J}(A_{\mathsf{cl}})^2\left\|B\right\|^2}.$$

The bound on excess risk from Theorem 3.1 may also be substituted into the tracking error bound in (9) to find that with probability at least $1-\delta-\delta'$, the tracking error satisfies

$$\max_{1\leq t\leq T}\left\|\hat{x}[t]-x_\star[t]\right\|^2\lesssim\mathcal{J}(A_{\mathsf{cl}})^2\left\|B\right\|^2\log\left(\frac{T}{\delta}\right)\sigma_z^2\left(\frac{kn_x\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right)}{cN_1TH}+\frac{kn_u+\log(\frac{1}{\delta'})}{N_2T}\right).$$

The above inequality provides the informal statement of the main result in Theorem 1.1 by hiding log terms as well as the terms dependent on system parameters. A bound for the expected cost gap $\left|\mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right]-\mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})]\right|$ can be similarly instantiated.

Remark 3.1 *The dependence of the tracking error bound in* (9) *on the stability of the target-task closed-loop system through $\mathcal{J}(A_{\mathsf{cl}})$ is tight (see Appendix C). Intuitively, less stable systems exacerbate the input errors from the learned controller.*

Remark 3.2 *Some examples of $h(\cdot)$ include LQR state costs $h(\vec{x}_{1:T})=\max_t\left\|Q^{1/2}x[t]\right\|$ and regularized tracking costs $h(\vec{x}_{1:T})=\max_t\left\|x[t]-x_{\text{goal}}[t]\right\|+\lambda\left\|Rx[t]\right\|$. Since $\frac{1}{T}\sum_{t=1}^T\left\|x[t]-y[t]\right\|\leq\max_{1\leq t\leq T}\left\|x[t]-y[t]\right\|$,* (10) *holds with no modification for time-averaged costs $h(\cdot)$.*

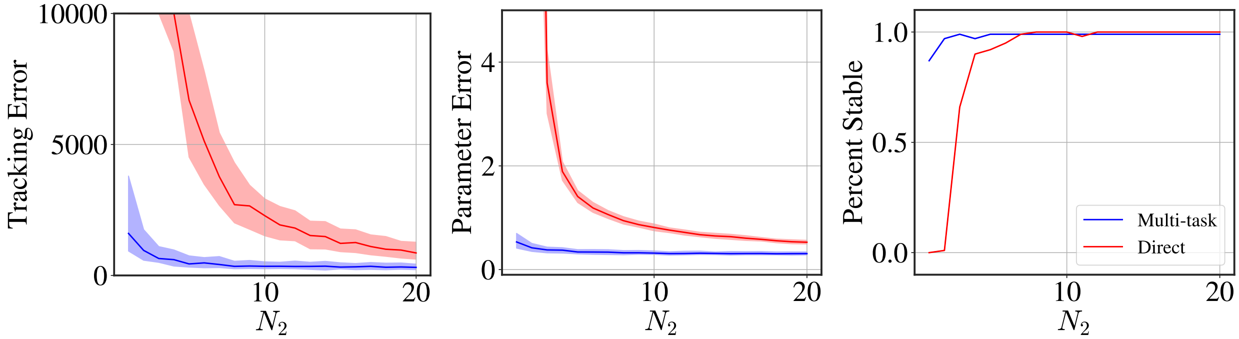
4 Numerical Results

We consider a simple system with $n_x=4$ and $n_u=2$ from Hong et al. (2021). In particular, let

$$x[t+1]=\begin{bmatrix}.99 & .03 & -.02 & -.32\\ .01 & .47 & 4.7 & .00\\ .02 & -.06 & .40 & .00\\ .01 & -.04 & .72 & .99\end{bmatrix}x[t]+\begin{bmatrix}.01 & .99\\ -3.44 & 1.66\\ -.83 & .44\\ -.47 & .25\end{bmatrix}u[t]=:Ax[t]+Bu[t].$$

We generate a collection of stabilizing controllers $K^{(1)},K^{(2)},\dots,K^{(H+1)}$ as LQR controllers with different cost matrices. Specifically, let $R=I_2$, and $Q^{(h)}=\alpha^{(h)}I_4$ for $\alpha^{(h)}\in\texttt{logspace}(-2,2,H+1)$, where $H=9$. The controllers $K^{(h)}$ are then given by $K^{(h)}=-(B^\top P^{(h)}B+R)^{-1}B^\top P^{(h)}A$, where $P^{(h)}$ solves the following Discrete Algebraic Riccati equation: $P^{(h)}=A^\top P^{(h)}A+A^\top P^{(h)}B(B^\top P^{(h)}B+R)^{-1}B^\top P^{(h)}A+Q^{(h)}$.

Next, assume that rather than directly observing the state, we obtain a high dimensional observation given by an injective linear function of the state: such an observation model can be viewed as a linear "perceptual sensor" or camera. In particular, we suppose that $y_t=Gx_t$, where $G\in\mathbb{R}^{50\times4}$. For simplicity, we select the elements of G i.i.d. from $\mathcal{N}(0,1)$, which ensures that G is injective almost surely. The dynamics of the observations may be written $y[t+1]=GAx_t+GBu[t]=GAG^\dagger y[t]+GBu[t]$, with the



(a) Tracking Error (b) Parameter Error (c) Percent Stable

Figure 1: We plot the tracking error between trajectories from the expert and learned controllers, $\max_{1 \le t \le T_{\text{test}}} \|\hat{y}[t] - y_\star[t]\|^2$, the parameter error, $\left(\left\|\hat{F}^{(H+1)}\hat{\Phi} - \bar{K}^{(H+1)}\right\|_F\right)$, and the percent of stable closed-loop systems for varying amounts of target task data to compare multi-task IL to directly learning the controller from target task data only. All metrics are plotted with respect to the lifted system in Equation (11). Multi-task IL demonstrates a significant benefit over direct IL in all metrics, especially when there is limited target task data.

input $u[t] = K^{(h)}x[t] = K^{(h)}G^\dagger y[t]$. Define $\bar{A} = GAG^\dagger$ and $\bar{B} = GB$, and $\bar{K}^{(h)} = K^{(h)}G^\dagger$. Consider the dynamics in the face of process noise $w[t] \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{50})$, along with inputs corrupted by noise $z[t] \overset{i.i.d.}{\sim} \mathcal{N}(0, I_2)$:

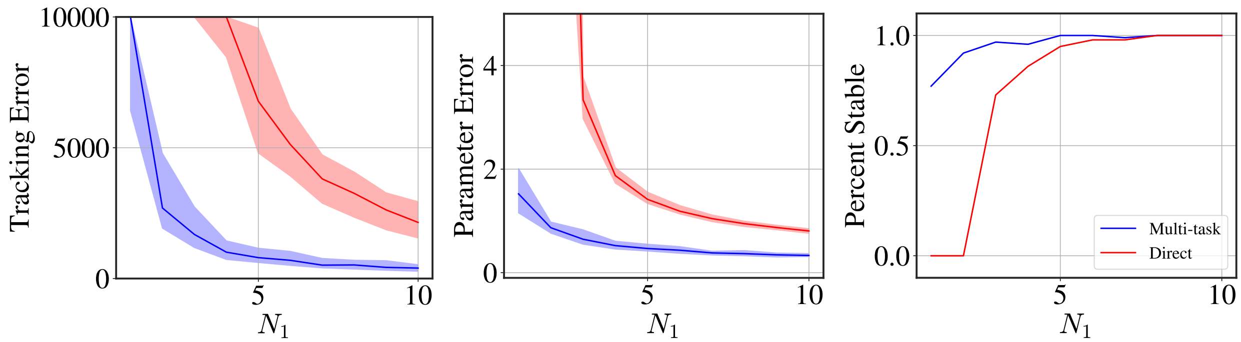
$$y[t+1] = (\bar{A} + \bar{B}\bar{K}^{(h)})y[t] + \bar{B}z[t] + w[t], \quad u[t] = \bar{K}^{(h)}y[t] + z[t]. \tag{11}$$

For the first H controllers, we collect N_1 trajectories of length $T = 20$ to get the pairs $\left\{\left\{\{(y_i^h[t], u_i^h[t])\}_{t=0}^{T-1}\right\}_{i=1}^{N_1}\right\}_{h=1}^{H}$. For the last controller, we collect N_2 length $T = 20$ trajectories to get the dataset $\left\{\left\{(y_i^{H+1}[t], u_i^{H+1}[t])\right\}_{t=0}^{T-1}\right\}_{i=1}^{N_2}$. Our goal is to learn the controller $\bar{K}^{(H+1)}$ from the collected state measurements and inputs. We compare the following ways of doing so:

- **Multi-task Imitation Learning:** We observe that the data generating mechanism ensures the existence of a low dimensional representation. In particular, one possible $\Phi_\star$ is $G^\dagger$. Therefore, the stage is set for the two step approach outlined in Section 2. In particular, we assume that the true underlying state dimension is known, and we set the low dimensional representation dimension to $k = 4$, and jointly optimize over Φ, $F^{(1)}, \dots, F^{(H)}$ in Problem (1). We approximately solve this problem with 10000 steps of alternating gradient descent using the `adam` optimizer (Kingma and Ba, 2014) in `optax` (Babuschkin et al., 2020) with a learning rate of 0.0001. The learned representation is then fixed, and the target training data is used to optimize $F^{(H+1)}$.
- **Direct Imitation Learning:** We compare multi-task learning to direct learning, which does not leverage the source data. In particular, given the target data, direct learning solves the problem $\text{minimize}_{F^{(H+1)}} \sum_{i=1}^{N_1} \sum_{t=0}^{T-1} \left\| F^{(H+1)} y_i^{(H+1)}[t] - u_i^{(H+1)}[t] \right\|^2$ [6]. In this setting, we let $\hat{\Phi} = I_{50}$.

Note that a 2×50 controller has $n_u \times n_x = 100$ parameters to learn from the target data. Meanwhile, multi-task imitation learning needs to learn a total of $k \times n_u + k \times n_x = 208$ parameters for the target

[6] Note that another baseline leverages the fact that a k dimensional representation of the state exists, and learns it using target task data only by solving $\text{minimize}_{F^{(H+1)}, \Phi} \sum_{i=1}^{N_1} \sum_{t=0}^{T-1} \left\| F^{(H+1)} \Phi y_i^{(H+1)}[t] - u_i^{(H+1)}[t] \right\|^2$. For the current example, however, $n_u < k$, so this approach is less efficient than the direct learning approach proposed.



(a) Tracking Error (b) Parameter Error (c) Percent Stable

Figure 2: We plot the tracking error between trajectories from the expert and learned controllers, $\max_{1 \le t \le T_{\text{test}}} \|\hat{y}[t] - y_\star[t]\|^2$, the parameter error, $\left(\left\|\hat{F}^{(H+1)}\hat{\Phi} - \bar{K}^{(H+1)}\right\|_F\right)$, and the percent of stable closed-loop systems for varying amounts of target task data to compare multi-task IL to directly learning the controller from target task data only. Similar to the setting of multi-task IL for transfer to a new task, leveraging all source data to learn the controller for a single source task provides a significant benefit in all three metrics over direct IL.

controller, but the $k \times n_u$ parameters are learned using source task data. This leaves only 8 parameters to learn using target task data.

Figure 1 plots three metrics that provide insight into the efficacy of these approaches: the imitation gap given by $\max_{1 \le t \le T_{\text{test}}} \|\hat{y}[t] - y_\star[t]\|^2$ for length $T_{\text{test}} = 100$ observation trajectories $\hat{y}$ and $y_\star$ rolled out under the learned controller and expert controller, respectively, with the same noise realizations; the parameter error, $\left\|\hat{F}^{(H+1)}\hat{\Phi} - \bar{K}^{(H+1)}\right\|_F$; and the percentage of trials where the learned controller is stabilizing. The trials are over ten realizations of G, as well as ten realizations of the noise, for a total of 100 trials. For each trial, $N_1 = 10$, while N_2 sweeps values in $[20]$. The medians for the imitation gap and parameter error are shown, with the $20\% - 80\%$ quantiles shaded.

In Figure 2, we additionally plot these metrics on one of the H source training tasks (arbitrarily chosen as $h = 7$) for varying amounts of training data to demonstrate the efficacy of the approach for multi-task learning. Here, N_1 ranges from 1 to 10. We compare the controller $\hat{F}^{(h)}\hat{\Phi}$ resulting from the shared training in Problem (1) with the controller from directly training a controller on this task without leveraging source task data. We note that our theoretical results, with mild modification, also support the efficacy of this simultaneous training of a representation and task weights.

5 Conclusion and Future Work

In this work, we study the sample complexity of multi-task imitation learning for linear systems. We find that if the different sets of expert demonstrations share a low dimensional representation, and the demonstrations are sufficiently diverse, then doing multi-task representation learning will lead to a smaller tracking error when deploying the learned policy in closed-loop, compared to learning a policy only from target task data. Our results are a first step towards understanding how the performance of a controller trained on multi-task data relates to the characteristics of the multi-task training data and the system being controlled. Some exciting directions for future work would be to extend our analysis to nonlinear systems and nonlinear representation functions, as well as other types of learning algorithms such as model-based and model-free

RL.

Acknowledgements

Katie Kang is supported by a NSF Graduate Research Fellowship. Bruce D. Lee is supported by the DoD through a National Defense Science & Engineering Fellowship. Sergey Levine is supported in part by the DARPA Assured Autonomy program. Nikolai Matni and Thomas Zhang are supported in part by NSF awards CPS-2038873, CAREER award ECCS-2045834, and a Google Research Scholar award.

References

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE, 2018.

Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6): 233–242, 1999.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

Yee Whye Teh, Victor Bapst, Wojciech Marian Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning, 2017. URL https://arxiv.org/abs/1707.04175.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures, 2018. URL https://arxiv.org/abs/1802.01561.

Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart, 2018. URL https://arxiv.org/abs/1809.04474.

Avi Singh, Eric Jang, Alexander Irpan, Daniel Kappler, Murtaza Dalal, Sergey Levine, Mohi Khansari, and Chelsea Finn. Scalable multi-task imitation learning with autonomous improvement, 2020. URL https://arxiv.org/abs/2003.02636.

Marc Peter Deisenroth, Peter Englert, Jan Peters, and Dieter Fox. Multi-task policy search for robotics. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3876–3881, 2014. doi: 10.1109/ICRA.2014.6907421.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. URL https://arxiv.org/abs/1703.03400.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables, 2019. URL https://arxiv.org/abs/1903.08254.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. Rl^2: Fast reinforcement learning via slow reinforcement learning, 2016. URL https://arxiv.org/abs/1611.02779.

Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Sergey Levine, and Chelsea Finn. Conservative data sharing for multi-task offline reinforcement learning, 2021. URL https://arxiv.org/abs/2109.08128.

Mengjiao Yang and Ofir Nachum. Representation matters: Offline pretraining for sequential decision making, 2021. URL https://arxiv.org/abs/2102.05815.

Rui Lu, Gao Huang, and Simon S. Du. On the power of multitask representation learning in linear mdp, 2021. URL https://arxiv.org/abs/2106.08053.

Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning, 2022. URL https://arxiv.org/abs/2206.05900.

Rui Lu, Andrew Zhao, Simon S. Du, and Gao Huang. Provable general function class representation learning in multitask bandits and mdps, 2022. URL https://arxiv.org/abs/2205.15701.

Zhiyuan Xu, Kun Wu, Zhengping Che, Jian Tang, and Jieping Ye. Knowledge transfer in multi-task deep reinforcement learning for continuous control, 2020. URL https://arxiv.org/abs/2010.07494.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning, 2015. URL https://arxiv.org/abs/1505.06279.

Sanjeev Arora, Simon S. Du, Sham Kakade, Yuping Luo, and Nikunj Saunshi. Provable representation learning for imitation learning via bi-level optimization, 2020. URL https://arxiv.org/abs/2002.10544.

James Harrison, Apoorva Sharma, Roberto Calandra, and Marco Pavone. Control adaptation via meta-learning dynamics. In *Workshop on Meta-Learning at NeurIPS*, volume 2018, 2018.

Spencer M Richards, Navid Azizan, Jean-Jacques Slotine, and Marco Pavone. Adaptive-control-oriented meta-learning for nonlinear systems. *arXiv preprint arXiv:2103.04490*, 2021.

Spencer M Richards, Navid Azizan, Jean-Jacques Slotine, and Marco Pavone. Control-oriented meta-learning. *arXiv preprint arXiv:2204.06716*, 2022.

Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. *Advances in Neural Information Processing Systems*, 34: 10013–10025, 2021.

Deepan Muthirayan, Dileep Kalathil, and Pramod P Khargonekar. Meta-learning online control for linear dynamical systems. *arXiv preprint arXiv:2208.10259*, 2022.

Rui Wang, Robin Walters, and Rose Yu. Meta-learning dynamics forecasting using task inference. *arXiv preprint arXiv:2102.10271*, 2021.

Aditya Modi, Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Joint learning of linear time-invariant dynamical systems. *arXiv preprint arXiv:2112.10955*, 2021.

Lidong Li, Claudio De Persis, Pietro Tesi, and Nima Monshizadeh. Data-based transfer stabilization in linear systems. *arXiv preprint arXiv:2211.05536*, 2022.

Jonathan Baxter. Learning internal representations. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 311–320, 1995.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from multiple sources. *Journal of Machine Learning Research*, 9(8), 2008.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33:7852–7862, 2020.

Kurtland Chua, Qi Lei, and Jason D Lee. How fine-tuning allows for effective meta-learning. *Advances in Neural Information Processing Systems*, 34:8871–8884, 2021.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Daniel Pfrommer, Thomas TCK Zhang, Stephen Tu, and Nikolai Matni. Tasil: Taylor series imitation learning. *arXiv preprint arXiv:2205.14812*, 2022.

J Hong, N Moehle, and S Boyd. Lecture notes in “introduction to matrix methods”, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, John Quan, George Papamakarios, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Luyu Wang, Wojciech Stokowiec, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL http://github.com/deepmind.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Yassir Jedra and Alexandre Proutiere. Finite-time identification of stable linear systems optimality of the least-squares estimator. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 996–1001, 2020. doi: 10.1109/CDC42340.2020.9304362.

Yan Mei Chen, Xiao Shan Chen, and Wen Li. On perturbation bounds for orthogonal projections. *Numerical Algorithms*, 73(2):433–444, 2016.

Xuefeng Xu. On the perturbation of an l2-orthogonal projection. *Journal of Computational and Applied Mathematics*, 368:112327, 2020.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

H.K. Khalil. *Nonlinear Systems*. Prentice Hall, 1996. ISBN 9780132280242. URL https://books.google.com/books?id=qiBuQgAACAAJ.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

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

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Additional Notation The set of $n \times m$ matrices ($n \geq m$) with orthonormal columns is denoted by $\mathcal{O}_{n,m}$. The set of d dimensional unit vectors is denoted by $\mathcal{S}^{d-1}$. We let $P_A = A(A^\top A)^\dagger A^\top$ denote the projection onto $\mathbf{span}(A)$, and let $P_A^\perp = I - P_A$ denote the projection onto $\mathbf{span}(A)^\perp$. The Kronecker product of a matrix A with a matrix B is denoted $A \otimes B$. The vectorization of a matrix A with columns $A_1, \dots, A_n$ is denoted $\mathrm{Vec}(A) = \begin{bmatrix} A_1^\top & A_2^\top & \dots & A_n^\top \end{bmatrix}^\top$. We denote the chi-squared distribution with k degrees of freedom by $\chi^2(k)$. Finally, we define the following $N_1 T \times n_u$ data matrices:

$$\begin{aligned} \mathbf{U}^{(h)} &= \begin{bmatrix} u_1^{(h)}[0] & \dots & u_1^{(h)}[T-1] & \dots & u_{N_1}^{(h)}[0] & \dots & u_{N_1}^{(h)}[T-1] \end{bmatrix}^\top, \\ \mathbf{Z}^{(h)} &= \begin{bmatrix} z_1^{(h)}[0] & \dots & z_1^{(h)}[T-1] & \dots & z_{N_1}^{(h)}[0] & \dots & z_{N_1}^{(h)}[T-1] \end{bmatrix}^\top. \end{aligned}$$

A Bounding the covariance concentration

We introduce the following version of the Hanson-Wright inequality (Vershynin, 2018, Theorem 6.3.2).

Proposition A.1 (Hanson-Wright for Gaussian Chaos) *Let $z \sim \mathcal{N}(0, I_n)$ be a random Gaussian vector, and $R \in \mathbb{R}^{m \times n}$ is an arbitrary fixed matrix. Then the following concentration inequalities hold for all $\varepsilon > 0$:*

$$\mathbb{P}\left[\|Rz\|^2 \geq (1+\varepsilon)\|R\|_F^2\right] \leq \exp\left(-\frac{1}{\sqrt{3}} \min\left\{\frac{9\varepsilon^2}{64}, \frac{3\varepsilon}{8}\right\} \frac{\|R\|_F^2}{\|R\|^2}\right), \tag{12}$$

and

$$\mathbb{P}\left[\left|\|Rz\|^2 - \|R\|_F^2\right| \geq \varepsilon \|R\|_F^2\right] \leq 2\exp\left(-\frac{1}{\sqrt{3}} \min\left\{\frac{9\varepsilon^2}{64}, \frac{3\varepsilon}{8}\right\} \frac{\|R\|_F^2}{\|R\|^2}\right), \tag{13}$$

Proof This result is a Gaussian specialization of Vershynin (2018, Theorem 6.3.2). To get explicit values of the absolute constants, we observe that the sub-Gaussian norm of $z \sim \mathcal{N}(0, I)$ can explicitly calculated via the MGF of a $\chi^2(1)$ random variable

$$K := \|z\|_{\psi_2} := \sup_{v \in \mathcal{S}^{n-1}} \inf\left\{t \geq 0 : \mathbb{E}\left[\exp((v^\top z)^2/t^2)\right] \leq 2\right\} = \inf\left\{t \geq 0 : \frac{t}{\sqrt{t^2-2}} \leq 2\right\} = \sqrt{\frac{8}{3}},$$

and setting $c = 1/\sqrt{3}$ satisfies the inequality

$$\mathbb{E}\left[\exp\left(\lambda^2 \|Rz\|^2\right)\right] \leq \exp\left(3\lambda^2 \|R\|_F^2\right), \quad |\lambda| \leq \frac{c}{\|R\|}. \tag{14}$$

These instantiate the constants contained in the proof of Vershynin (2018, Theorem 6.2.1) ∎

The following ε-net argument is standard, and a proof may be found in Chapter 4 of Vershynin (2018).

Lemma A.1 *Let W be and $d \times d$ random matrix, and $\varepsilon \in [0, 1/2)$. Furthermore, let $\mathcal{N}$ be an ε-net of $\mathcal{S}^{d-1}$ with minimal cardinality. Then for all $\rho > 0$,*

$$\mathbb{P}[\|W\| > \rho] \leq \left(\frac{2}{\varepsilon} + 1\right)^d \max_{x \in \mathcal{N}} \mathbb{P}\left[\left|x^\top W x\right| > (1-2\epsilon)\rho\right].$$

We now prove a generalized result on the concentration of covariances from Jedra and Proutiere (2020) in which the covariates are pre and post multiplied by a matrix with orthonormal rows and its transpose, respectively. This demonstrates error bounds scaling with the lower dimension of the orthonormal matrix.

Lemma A.2 (Modified Lemma 2 of Jedra and Proutiere (2020)) *Let $A \in \mathbb{R}^{n\times n}$ satisfy $\rho(A) < 1$ and let $\Sigma_w \succ 0$ with dimension $n \times n$. Let Σ_x solve the discrete Lyapunov equation:*

$$\Sigma_x = A\Sigma_x A^\top + \Sigma_w.$$

Consider drawing N_1 trajectories of length T from a linear system $x_i[t+1] = Ax_i[t] + w_i[t]$, where $w_i[t] \overset{i.i.d}{\sim} \mathcal{N}(0, \Sigma_w)$ for $i = 1, \dots N_1$ and $t = 0, \dots T-1$, and $x_i[0] \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_x)$ for $i = 1, \dots, N_1$. Let $\mathbf{X}$ be the matrix of stacked state data as in Equation (4). Let $U \in \mathcal{O}_{n,d}$ with $d \le n$ be independent of $\mathbf{X}$. Define $M = \left(N_1 T U^\top \Sigma_x U\right)^{-1/2}$. Then

$$\left\|(\mathbf{X}UM)^\top \mathbf{X}UM - I\right\| > \max\{\varepsilon, \varepsilon^2\}$$

holds with probability at most

$$2\exp\left(-c_1\varepsilon^2 \frac{N_1 T}{\kappa(\Sigma_x)\mathcal{J}(A)^2} + c_2 d\right)$$

for some absolute constants c_1 and c_2.

Proof

Note that we can write

$$\mathbb{E}\Big[(\mathbf{X}U)^\top \mathbf{X}U\Big] = \mathbb{E}\left[U^\top \sum_{i=1}^{N_1}\sum_{t=0}^{T-1} x_i[t]x_i[t]^\top U\right] = N_1 T U^\top \Sigma_x U.$$

Then

$$\begin{aligned}
\left\|(\mathbf{X}UM)^\top(\mathbf{X}UM) - I\right\| &= \sup_{v\in\mathcal{S}^{d-1}} \left|v^\top\Big((\mathbf{X}UM)^\top\mathbf{X}UM - I\Big)v\right| \\
&= \sup_{v\in\mathcal{S}^{d-1}} \left|v^\top\Big((\mathbf{X}UM)^\top\mathbf{X}UM\Big)v - \mathbb{E}v^\top\Big((\mathbf{X}UM)^\top\mathbf{X}UM\Big)v\right| \\
&= \sup_{v\in\mathcal{S}^{d-1}} \left|\|\mathbf{X}UMv\|^2 - \mathbb{E}\,\|\mathbf{X}UMv\|^2\right| \\
&\overset{(i)}{=} \sup_{v\in\mathcal{S}^{d-1}} \left|\left\|\sigma_{Mv}^\top\tilde{\Gamma}\xi\right\|^2 - \mathbb{E}\left\|\sigma_{Mv}^\top\tilde{\Gamma}\xi\right\|^2\right| \\
&= \sup_{v\in\mathcal{S}^{d-1}} \left|\left\|\sigma_{Mv}^\top\tilde{\Gamma}\xi\right\|^2 - \left\|\sigma_{Mv}^\top\tilde{\Gamma}\right\|_F^2\right|,
\end{aligned}$$

where (i) follows by defining

$$\Gamma = I_{N_1} \otimes \begin{bmatrix} \Sigma_x^{1/2} & & & \\ A\Sigma_x^{1/2} & \Sigma_w^{1/2} & & 0 \\ & & \ddots & \\ A^{T-1}\Sigma_x^{1/2} & \dots & A\Sigma_w^{1/2} & \Sigma_w^{1/2} \end{bmatrix},$$

$$\xi = \mathrm{Vec}\big(\begin{bmatrix}\eta_0[-1] & \dots & \eta_0[T-1] & \dots & \eta_{N_1}[0] & \dots & \eta_{N_1}[T-2]\end{bmatrix}\big),$$

where $\eta_i[t] \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ for $t \in [-1, T-2]$, and

$$\sigma_{Mv} := I_{N_1 T} \otimes (Mv), \qquad \tilde{\Gamma} := (I_{N_1 T} \otimes U^\top)\Gamma,$$

and recalling that $\mathrm{Vec}(U^\top \mathbf{X}^\top) = (I_{N_1 T} \otimes U^\top)\Gamma\xi = \tilde{\Gamma}\xi$. Next oberserve that for any $v \in \mathcal{S}^{d-1}$, we can apply Proposition A.1 to show that

$$\mathbb{P}\left[\left|\left\|\sigma_{Mv}^\top \tilde{\Gamma}\xi\right\|^2 - \left\|\sigma_{Mv}\tilde{\Gamma}\right\|_F^2\right| > \rho\left\|\sigma_{Mv}^\top\tilde{\Gamma}\right\|_F^2\right] \le 2\exp\left(-C\min\{\rho^2, \rho\}\frac{\left\|\sigma_{Mv}^\top\tilde{\Gamma}\right\|_F^2}{\left\|\sigma_{Mv}^\top\tilde{\Gamma}\right\|^2}\right)$$

for some positive universal constant C. Next observe that $\left\|\sigma_{Mv}\tilde{\Gamma}\right\|_F^2 = 1$, and $\left\|\sigma_{Mv}^\top\tilde{\Gamma}\right\| \le \|M\|\left\|\tilde{\Gamma}\right\|$, and thus the right hand side reduces to

$$2\exp\left(-\frac{C\min\{\rho^2, \rho\}}{\|M\|^2\left\|\tilde{\Gamma}\right\|^2}\right).$$

Applying Lemma A.1 with $\varepsilon = \frac{1}{4}$, we have that

$$\mathbb{P}\left[\left\|(\mathbf{X}UM)^\top\mathbf{X}UM - I\right\| > \rho\right] \le 2\cdot 9^d\exp\left(-\frac{C\min\{\rho^2/4, \rho/2\}}{\|M\|^2\left\|\tilde{\Gamma}\right\|^2}\right).$$

Setting $\rho = \max\{\varepsilon, \varepsilon^2\}$ and rearranging constants provides

$$\mathbb{P}\left[\left\|(\mathbf{X}UM)^\top\mathbf{X}UM - I\right\| > \max\{\varepsilon, \varepsilon^2\}\right] \le 2\exp\left(-\frac{c_1\varepsilon^2}{\|M\|^2\|\Gamma\|^2} + c_2 d\right).$$

Lastly, note that

$$\|M\| = \frac{\left\|(U^\top\Sigma_x U)^{-1/2}\right\|}{\sqrt{N_1 T}} = \frac{1}{\sqrt{N_1 T}}\frac{1}{\lambda_d(U^\top\Sigma_x U)^{1/2}} \le \frac{1}{\sqrt{N_1 T}}\frac{1}{\lambda_{\min}(\Sigma_x)^{1/2}}$$

and

$$\begin{aligned}
\left\|\tilde{\Gamma}\right\| &\le \|\Gamma\| \\
&= \left\|\begin{bmatrix} \Sigma_x^{1/2} & & & \\ A\Sigma_x^{1/2} & \Sigma_w^{1/2} & & 0 \\ & & \ddots & \\ A^{T-1}\Sigma_x^{1/2} & \dots & A\Sigma_w^{1/2} & \Sigma_w^{1/2}\end{bmatrix}\right\| \\
&\le \left\|\begin{bmatrix}\Sigma_x^{1/2} \\ \vdots \\ A^{T-1}\Sigma_x^{1/2}\end{bmatrix}\right\| + \left\|\begin{bmatrix}\Sigma_w^{1/2} & & 0 \\ & & \ddots & \\ A^{T-2}\Sigma_w^{1/2} & \dots & A\Sigma_w^{1/2} & \Sigma_w^{1/2}\end{bmatrix}\right\| \\
&\le \sum_{s=0}^{T-1}\|A^s\|\|\Sigma_x\|^{1/2} + \left\|\begin{bmatrix} I & & 0 \\ & & \ddots & \\ A^{T-2} & \dots & A & I\end{bmatrix}\right\|\|\Sigma_w\|^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \sum_{s \geq 0} \|A^s\| \, \|\Sigma_x\|^{1/2} \\
&= 2\mathcal{J}(A) \, \|\Sigma_x\|^{1/2}.
\end{aligned}
$$

where the final inequality follows by applying Lemma 5 of Jedra and Proutiere (2020) and the fact that $\Sigma_w \preceq \Sigma_x$. The lemma then follows from the fact that $\kappa(\Sigma_x) = \frac{\|\Sigma_x\|}{\lambda_{\min}(\Sigma_x)}$. ■

The previous lemma can now be applied to show concentration of the empirical covariance for both the source and target task.

Lemma A.3 *Suppose $N_1 T \geq C \max_h \mathcal{J}\left(A + BK^{(h)}\right)^2 \kappa\left(\Sigma_x^{(h)}\right)(n_x + \log(H/\delta))$ for $\delta \in (0,1)$, where C is a universal numerical constant. Then with probability at least $1 - \frac{\delta}{10}$ over the states $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(H)}$ in the source tasks, we have*

$$
0.9\Sigma_x^{(h)} \preceq \frac{1}{N_1 T}\left(\mathbf{X}^{(h)}\right)^\top \mathbf{X}^{(h)} \preceq 1.1\Sigma_x^{(h)}, \qquad \forall h \in [H].
$$

Proof Applying Lemma A.2 with $U = I$ and $\varepsilon = 0.1$, as long as $C \geq 100\frac{\max\{c_2, \log(1/20)\}}{c_1}$, then for any task $h \in [H]$, we have that

$$
\left\|(\mathbf{X}^{(h)} M)^\top \mathbf{X}^{(h)} M - I\right\| \leq 0.1,
$$

with probability at least $1 - 2\exp\left(-\frac{0.01 c_1 N_1 T}{\mathcal{J}(A + BK^{(h)})^2 \kappa(\Sigma_x^{(h)})} + c_2 n_x\right) \geq 1 - \frac{\delta}{10H}$. This may be written equivalently as

$$
0.9\Sigma_x^{(h)} = 0.9\frac{M^{-2}}{N_1 T} \preceq \frac{\mathbf{X}^{(h)\top}\mathbf{X}^{(h)}}{N_1 T} \preceq 1.1\frac{M^{-2}}{N_1 T} = 1.1\Sigma_x^{(h)}.
$$

Taking a union bound over the H tasks provides the desired result. ■

Lemma A.4 *Suppose $N_2 T \geq C\mathcal{J}\left(A + BK^{(H+1)}\right)^2 \kappa\left(\Sigma_x^{(H+1)}\right)(k + \log(1/\delta))$ for $\delta \in (0,1)$, where C is a universal numerical constant. Then for any given matrix $\Phi \in \mathbb{R}^{n_x \times 2k}$ independent of $\mathbf{X}^{(H+1)}$, with probability at least $1 - \frac{\delta}{10}$ over $\mathbf{X}^{(H+1)}$ we have*

$$
0.9\Phi^\top \Sigma_x^{(H+1)} \Phi \preceq \frac{1}{N_2 T}\Phi^\top \left(\mathbf{X}^{(H+1)}\right)^\top \mathbf{X}^{(H+1)} \Phi \preceq 1.1\Phi^\top \Sigma_x^{(H+1)} \Phi.
$$

Proof Let $\Phi = USV^\top$ be the singular value decomposition of Φ with $U \in \mathbb{R}^{n_x \times 2k}$. Applying Lemma A.2 with $\varepsilon = 0.1$, we have that

$$
0.9U^\top \Sigma_x^{(H+1)} U = 0.9\frac{M^{-2}}{N_2 T} \preceq \frac{U^\top \mathbf{X}^{(H+1)\top}\mathbf{X}^{(H+1)} U}{N_2 T} \preceq 1.1\frac{M^{-2}}{N_2 T} = 1.1U^\top \Sigma_x^{(H+1)} U \tag{15}
$$

with probability at least $1 - 2\exp\left(-0.01 c_1 \frac{N_2 T}{\mathcal{J}(A + BK^{(H+1)})^2 \kappa(\Sigma_x^{(H+1)})} + 2c_2 k\right) \geq 1 - \frac{\delta}{10}$ as long as $C \geq 100\frac{\max\{c_2, \log(1/20)\}}{c_1}$. Left and right multiplying Equation (15) by $V^\top S$ and SV respectively results in

$$
0.9\Phi^\top \Sigma_x^{(H+1)} \Phi \preceq \frac{\Phi^\top \mathbf{X}^{(H+1)\top}\mathbf{X}^{(H+1)} \Phi}{N_2 T} \preceq 1.1\Phi^\top \Sigma_x^{(H+1)} \Phi
$$

with probability at least $1 - \frac{\delta}{10}$. ■

B Data Guarantees

We will now use the covariance concentration results to show concentration of the source and target controllers to the optimal controllers. We begin by recalling Lemma A.5 of Du et al. (2020).

Lemma B.1 *There exists a subset $\mathcal{N} \subset \mathcal{O}_{d_1,d_2}$ that is an ε-net of $\mathcal{O}_{d_1,d_2}$ in Frobenius norm such that $|\mathcal{N}| \le \left(\frac{6\sqrt{d_2}}{\varepsilon}\right)^{d_1 d_2}$.*

Also, note the following fact.

Fact B.1 *For size conforming Z, X, with X full column rank,*

$$\sup_F \left[4\langle Z, XF^\top\rangle - \left\|XF^\top\right\|_F^2\right] = 4\left\|(X^\top X)^{-1/2}X^\top Z\right\|_F^2 = 4\left\|P_X Z\right\|_F^2.$$

The following result on perturbation of projection matrices may be found in Chen et al. (2016); Xu (2020).

Lemma B.2 (Perturbation of projection matrices) *Let A, B be size conforming matrices both of the same rank. We have:*

$$\|P_A - P_B\| \le \min\left\{\left\|A^\dagger\right\|, \left\|B^\dagger\right\|\right\}\|A - B\|.$$

Lemma B.3 *Let $(x_t)_{t\ge1}$ be a $\mathbb{R}^d$-valued process adapted to a filtration $(\mathcal{F}_t)_{t\ge1}$. Let $(\eta_t)_{t\ge1}$ be a $\mathbb{R}^m$-valued process adapted to $(\mathcal{F}_t)_{t\ge2}$. Suppose that $(\eta_t)_{t\ge1}$ is a σ-sub-Gaussian martingale difference sequence, i.e.,:*

$$\mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \tag{16}$$

$$\mathbb{E}[\exp(\lambda\langle v, \eta_t\rangle) \mid \mathcal{F}_t] \le \exp\left(\frac{\lambda^2\sigma^2\|v\|^2}{2}\right) \;\; \forall \mathcal{F}_t\text{-measurable } \lambda \in \mathbb{R}, v \in \mathbb{R}^m. \tag{17}$$

For $\Lambda \in \mathbb{R}^{m\times d}$, let $(M_t(\Lambda))_{t\ge1}$ be the $\mathbb{R}$-valued process:

$$M_t(\Lambda) = \exp\left(\frac{1}{\sigma}\sum_{i=1}^t \langle \Lambda x_i, \eta_i\rangle - \frac{1}{2}\sum_{i=1}^t \|\Lambda x_i\|^2\right). \tag{18}$$

The process $(M_t(\Lambda))_{t\ge1}$ satisfies $\mathbb{E}[M_t(\Lambda)] \le 1$ for all $t \ge 1$.

Proof Let $M_0(\Lambda) := 1$. Fix $t \ge 1$. Observe:

$$\begin{aligned}\mathbb{E}[M_t(\Lambda)] = \mathbb{E}[\mathbb{E}[M_t(\Lambda) \mid \mathcal{F}_t]] &= \mathbb{E}\left[M_{t-1}(\Lambda)\mathbb{E}\left[\exp\left(\frac{1}{\sigma}\langle \Lambda x_t, \eta_t\rangle\right) \,\middle|\, \mathcal{F}_t\right]\exp\left(-\frac{1}{2}\|\Lambda x_t\|^2\right)\right] \\ &\le \mathbb{E}[M_{t-1}(\Lambda)].\end{aligned}$$

■

The following result generalizes the self-normalized martingale inequality from Abbasi-Yadkori et al. (2011) to handle multiple matrix valued self-normalized martingales.

Proposition B.1 (Generalized self-normalized martingale inequality) *Fix $H \in \mathbb{N}_+$. For $h \in [H]$, let $(x_t^h, \eta_t^h)_{t\geq 1}$ be a $\mathbb{R}^d \times \mathbb{R}^m$-valued process and $(\mathcal{F}_t^h)_{t\geq 1}$ be a filtration such that $(x_t^h)_{t\geq 1}$ is adapted to $(\mathcal{F}_t^h)_{t\geq 1}$, $(\eta_t^h)_{t\geq 1}$ is adapted to $(\mathcal{F}_t^h)_{t\geq 2}$, and $(\eta_t^h)_{t\geq 1}$ is a σ-sub-Gaussian martingale difference sequence. Suppose that for all $h_1 \neq h_2$, the process $(x_t^{h_1}, \eta_t^{h_1})$ is independent of $(x_t^{h_2}, \eta_t^{h_2})$. Fix (non-random) positive definite matrices $\{V^h\}_{h=1}^H$. For $t \geq 1$ and $h \in [H]$, define:*

$$\bar{V}_t^h := V^h + V_t^h, \;\; V_t^h := \sum_{i=1}^t x_i x_i^\top, \;\; S_t^h := \sum_{i=1}^t x_i \eta_i^\top. \tag{19}$$

For any fixed $T \in \mathbb{N}_+$, with probability at least $1-\delta$:

$$\sum_{h=1}^H \left\| (\bar{V}_T^h)^{-1/2} S_T^h \right\|_F^2 \leq 2\sigma^2 \left[\sum_{h=1}^H \log\det((\bar{V}_T^h)^{m/2}(V^h)^{-m/2}) + \log(1/\delta) \right]. \tag{20}$$

Proof By rescaling, we assume that $\sigma = 1$ without loss of generality. For $h \in [H]$, define:

$$M_T^h(\Lambda) := \exp\left(\sum_{t=1}^T \langle \Lambda x_t^h, \eta_t^h \rangle - \frac{1}{2} \sum_{t=1}^T \left\| \Lambda x_t^h \right\|^2 \right), \;\; \Lambda \in \mathbb{R}^{m\times d}. \tag{21}$$

By Lemma B.3, $\mathbb{E}[M_T^h(\Lambda)] \leq 1$ for any fixed Λ. Now, we use the method of mixtures argument from Abbasi-Yadkori et al. (2011). Let $\Gamma \in \mathbb{R}^{m\times d}$ be a random matrix with i.i.d. $N(0,1)$ entries, independent of everything else. By the tower property, $\mathbb{E}[M_T^h(\Gamma)] = \mathbb{E}[\mathbb{E}[M_T^h(\Gamma)] \mid \Gamma] \leq 1$. On the other hand, let us compute $\mathbb{E}[M_T^h(\Gamma) \mid \mathcal{F}_\infty^h]$. Let $p(\Lambda)$ denote the density of Γ, and also let $p(g)$ denote the density of a d-dimensional isotropic normal vector. Let us momentarily drop the superscript h since the computation is identical for every h. First, we note that the proof of Theorem 3 in Abbasi-Yadkori et al. (2011) shows the following identity for any fixed $u \in \mathbb{R}^d$, $V \in \mathbb{R}^{m\times m}$ positive semidefinite, and $V_0 \in \mathbb{R}^{m\times m}$ positive definite:

$$\int \exp\left(\langle g, u \rangle - \frac{1}{2} \|g\|_V^2 \right) p(g)\, \mathrm{d}g = \left(\frac{\det(V_0)}{\det(V_0 + V)} \right)^{1/2} \exp\left(\frac{1}{2} \|u\|_{(V+V_0)^{-1}}^2 \right).$$

With this identity and independence of the rows of Γ:

$$\begin{aligned}
\mathbb{E}[M_T(\Gamma) \mid \mathcal{F}_\infty] &= \int \exp\left(\langle \Lambda^\top, S_T \rangle - \frac{1}{2} \left\| \Lambda^\top \right\|_{V_t}^2 \right) p(\Lambda)\, \mathrm{d}\Lambda \\
&= \int \exp\left(\sum_{i=1}^m \langle \Lambda^\top e_i, S_T e_i \rangle - \frac{1}{2} \left\| \Lambda^\top e_i \right\|_{V_t}^2 \right) p(\Lambda)\, \mathrm{d}\Lambda \\
&= \prod_{i=1}^m \int \exp\left(\langle g, S_T e_i \rangle - \frac{1}{2} \|g\|_{V_t}^2 \right) p(g)\, \mathrm{d}g \\
&= \left(\frac{\det(V)}{\det(V+V_T)} \right)^{m/2} \exp\left(\frac{1}{2} \sum_{i=1}^m \|S_T e_i\|_{(V+V_T)^{-1}}^2 \right) \\
&= \left(\frac{\det(V)}{\det(V+V_T)} \right)^{m/2} \exp\left(\frac{1}{2} \left\| (V+V_T)^{-1/2} S_T \right\|_F^2 \right).
\end{aligned}$$

That is, for every $h \in [H]$, we have:

$$\mathbb{E}\left[\left(\frac{\det(V^h)}{\det(\bar{V}_T^h)} \right)^{m/2} \exp\left(\frac{1}{2} \left\| (\bar{V}_T^h)^{-1/2} S_T \right\|_F^2 \right) \right] \leq 1.$$

Now by Markov's inequality and independence of the processes across h:

$$
\begin{aligned}
&\mathbb{P}\left(\sum_{h=1}^{H}\left\|(\bar{V}_T^h)^{-1/2}S_T\right\|_F^2 > 2\left[\sum_{h=1}^{H}\log\det((\bar{V}_T^h)^{m/2}(V^h)^{-m/2}) + \log(1/\delta)\right]\right)\\
&= \mathbb{P}\left(\exp\left(\frac{1}{2}\sum_{h=1}^{H}\left\|(\bar{V}_T^h)^{-1/2}S_T\right\|_F^2\right) > \delta^{-1}\prod_{h=1}^{H}\left(\frac{\det(\bar{V}_T^h)}{\det(V^h)}\right)^{m/2}\right)\\
&= \mathbb{P}\left(\prod_{h=1}^{H}\left(\frac{\det(V^h)}{\det(\bar{V}_T^h)}\right)^{m/2}\exp\left(\frac{1}{2}\left\|(\bar{V}_T^h)^{-1/2}S_T\right\|_F^2\right) > \delta^{-1}\right)\\
&\leq \delta\mathbb{E}\left[\prod_{h=1}^{H}\left(\frac{\det(V^h)}{\det(\bar{V}_T^h)}\right)^{m/2}\exp\left(\frac{1}{2}\left\|(\bar{V}_T^h)^{-1/2}S_T\right\|_F^2\right)\right]\\
&= \delta\prod_{h=1}^{H}\mathbb{E}\left[\left(\frac{\det(V^h)}{\det(\bar{V}_T^h)}\right)^{m/2}\exp\left(\frac{1}{2}\left\|(\bar{V}_T^h)^{-1/2}S_T\right\|_F^2\right)\right]\\
&\leq \delta.
\end{aligned}
$$

■

Remark B.1 *The original self-normalized martingale bound from Abbasi-Yadkori et al. (2011) holds for an arbitrary stopping time, and consequently* uniformly *for all time $t \in \mathbb{N}$ by a simple reduction. In contrast, Proposition B.1 holds for a fixed $t \in \mathbb{N}$, which suffices for our purposes.*

Lemma B.4 (source controller concentration) *Under the setting and data assumptions of Theorem 3.1, with probability at least $1 - \frac{\delta}{5}$,*

$$
\sum_{h=1}^{H}\left\|\mathbf{X}^{(h)}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \lesssim \sigma_z^2\left(kmH + kn\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right)\right).
$$

Proof Before we start the proof, we first define two subsets:

$$
\begin{aligned}
\Theta_1 &:= \{(F_1\Phi, \dots, F_h\Phi) \mid F_i \in \mathbb{R}^{n_u \times k},\ \ \Phi \in \mathbb{R}^{k \times n_x}\},\\
\Theta_2(a,b) &:= \{(F_1\Phi, \dots, F_h\Phi) \mid F_i \in \mathbb{R}^{n_u \times b},\ \ \Phi^\top \in \mathcal{O}_{a,b}\},\ \ a \geq b.
\end{aligned}
$$

It is easy to see that $\Theta_1 = \Theta_2(n_x, k)$. Furthermore, a simple argument shows that

$$
\Theta_2(n_x, k) - \Theta_2(n_x, k) \subset \Theta_2(n_x, 2k),
$$

where the minus sign indicates the Minkowski difference of two sets.

Throughout this proof, we will condition on the following event, which holds with probability at least $1 - \frac{\delta}{10}$ by Lemma A.3:

$$
0.9\Sigma_x^{(h)} \preceq \frac{1}{N_1T}\left(\mathbf{X}^{(h)}\right)^\top\mathbf{X}^{(h)} \preceq 1.1\Sigma_x^{(h)}, \qquad \forall h \in [H]. \tag{22}
$$

By optimality of $\hat{F}^{(1)}, \hat{F}^{(2)}, \ldots, \hat{F}^{(H)}$ and $\hat{\Phi}$ for Problem (1), we have that

$$\sum_{h=1}^{H} \left\| \mathbf{U}^{(h)} - \mathbf{X}^{(h)} (\hat{F}^{(h)} \hat{\Phi})^\top \right\|_F^2 \leq \sum_{h=1}^{H} \left\| \mathbf{U}^{(h)} - \mathbf{X}^{(h)} (F_\star^{(h)} \Phi_\star)^\top \right\|_F^2.$$

Substituting in $\mathbf{U}^{(h)} = \mathbf{X}^{(h)} (F_\star^{(h)} \Phi_\star)^\top + \mathbf{Z}^{(h)}$ and re-arranging yields the basic inequality:

$$\sum_{h=1}^{H} \left\| \mathbf{X}^{(h)} (\hat{F}^{(h)} \hat{\Phi} - F_\star^{(h)} \Phi_\star)^\top \right\|_F^2 \leq 2 \sum_{h=1}^{H} \operatorname{Tr}\Big(\Big(\mathbf{Z}^{(h)}\Big)^\top \mathbf{X}^{(h)} (F_\star^{(h)} \Phi_\star - \hat{F}^{(h)} \hat{\Phi})^\top \Big). \tag{23}$$

Let $\Delta^{(h)} := \hat{F}^{(h)} \hat{\Phi} - F_\star^{(h)} \Phi_\star$. Multiplying the basic inequality above by two and re-arranging again, we obtain the *offset basic inequality* (Liang et al., 2015):

$$\begin{aligned}
\sum_{h=1}^{H} \left\| \mathbf{X}^{(h)} (\Delta^{(h)})^\top \right\|_F^2 &\leq \sum_{h=1}^{H} 4 \operatorname{Tr}\Big((\mathbf{Z}^{(h)})^\top \mathbf{X}^{(h)} (\Delta^{(h)})^\top \Big) - \left\| \mathbf{X}^{(h)} (\Delta^h(h))^\top \right\|_F^2 \\
&\leq \sup_{\{\Delta^{(h)}\}_{h=1}^H \in \Theta_1 - \Theta_1} \sum_{h=1}^{H} 4 \operatorname{Tr}\Big((\mathbf{Z}^{(h)})^\top \mathbf{X}^{(h)} (\Delta^{(h)})^\top \Big) - \left\| \mathbf{X}^{(h)} (\Delta^h(h))^\top \right\|_F^2 \\
&= \sup_{\{\Delta^{(h)}\}_{h=1}^H \in \Theta_2(n_x,k) - \Theta_2(n_x,k)} \sum_{h=1}^{H} 4 \operatorname{Tr}\Big((\mathbf{Z}^{(h)})^\top \mathbf{X}^{(h)} (\Delta^{(h)})^\top \Big) - \left\| \mathbf{X}^{(h)} (\Delta^h(h))^\top \right\|_F^2 \\
&\leq \sup_{\{\Delta^{(h)}\}_{h=1}^H \in \Theta_2(n_x,2k)} \sum_{h=1}^{H} 4 \operatorname{Tr}\Big((\mathbf{Z}^{(h)})^\top \mathbf{X}^{(h)} (\Delta^{(h)})^\top \Big) - \left\| \mathbf{X}^{(h)} (\Delta^h(h))^\top \right\|_F^2 \\
&= \sup_{\Phi^\top \in \mathcal{O}_{n_x,2k}} \sum_{h=1}^{H} \sup_{F_h \in \mathbb{R}^{n_u \times 2k}} \left[4 \operatorname{Tr}\Big((\mathbf{Z}^{(h)})^\top \mathbf{X}^{(h)} \Phi^\top F_h^\top \Big) - \left\| \mathbf{X}^{(h)} \Phi^\top F_h^\top \right\|_F^2 \right] \\
&= 4 \sup_{\Phi^\top \in \mathcal{O}_{n_x,2k}} \sum_{h=1}^{H} \left\| (\Phi (\mathbf{X}^{(h)})^\top \mathbf{X}^{(h)} \Phi^\top)^{-1/2} \Phi (\mathbf{X}^{(h)})^\top \mathbf{Z}^{(h)} \right\|_F^2.
\end{aligned}$$

The last equality above follows from Fact B.1. We now derive an upper bound on the sum for a fixed $\Phi^\top \in \mathcal{O}_{n_x,2k}$. To do this, we invoke Proposition B.1 with the trajectories within each task concatenated in sequence: $V^h \leftarrow 0.9 N_1 T \Phi \Sigma_x^{(h)} \Phi^\top$, $x_t^h \leftarrow \Phi x_i^{(h)}[t]$, and $\eta_t^h \leftarrow z_i^{(h)}[t]$. Note that since $\Phi (\mathbf{X}^{(h)})^\top \mathbf{X}^{(h)} \Phi^\top \succeq V^h$, we then have that:

$$(\Phi (\mathbf{X}^{(h)})^\top \mathbf{X}^{(h)} \Phi^\top)^{-1} \preceq 2 (\Phi (\mathbf{X}^{(h)})^\top \mathbf{X}^{(h)} \Phi^\top + V^h)^{-1}.$$

Consequently, with probability at least $1-\delta'$,

$$
\begin{aligned}
&\sum_{h=1}^{H}\left\|(\Phi(\mathbf{X}^{(h)})^\top\mathbf{X}^{(h)}\Phi^\top)^{-1/2}\Phi(\mathbf{X}^{(h)})^\top\mathbf{Z}^{(h)}\right\|_F^2\\
&\quad\lesssim \sum_{h=1}^{H}\left\|(\Phi(\mathbf{X}^{(h)})^\top\mathbf{X}^{(h)}\Phi^\top+V^h)^{-1/2}\Phi(\mathbf{X}^{(h)})^\top\mathbf{Z}^{(h)}\right\|_F^2\\
&\quad\lesssim \sigma^2\left[\sum_{h=1}^{H}\log\det\left(\left(\Phi\left(\mathbf{X}^{(h)}\right)^\top\mathbf{X}^{(h)}\Phi^\top+V^h\right)^{n_u/2}\left(V^h\right)^{-n_u/2}\right)+\log(1/\delta')\right]\\
&\quad= \frac{n_u}{2}\sigma^2\left[\sum_{h=1}^{H}\log\det\left(\left(\Phi\left(\mathbf{X}^{(h)}\right)^\top\mathbf{X}^{(h)}\Phi^\top+V^h\right)\left(V^h\right)^{-1}\right)+\log(1/\delta')\right]\\
&\quad= \frac{n_u}{2}\sigma^2\left[\sum_{h=1}^{H}\log\det\left(\frac{1}{0.9N_1T}\left(\Phi\left(\mathbf{X}^{(h)}\right)^\top\mathbf{X}^{(h)}\Phi^\top\right)\left(\Phi\Sigma_x^{(h)}\Phi^\top\right)^{-1}+I_{2k}\right)+\log(1/\delta')\right]\\
&\quad\le \frac{n_u}{2}\sigma^2\left[\sum_{h=1}^{H}\log\det\left(\frac{1.1}{0.9}+1\right)I_{2k}+\log(1/\delta')\right]\\
&\quad\lesssim \sigma_z^2(kHn_u+\log(1/\delta')).
\end{aligned}
\tag{24}
$$

This holds for a fixed $\Phi^\top\in\mathcal{O}_{n_x,2k}$, so it remains to union bound over $\mathcal{O}_{n_x,2k}$. Let $\{\Phi_i^\top\}_{i=1}^N$ be an ε-net of $\mathcal{O}_{n_x,2k}$ in the spectral norm of resolution to be determined. For $\Phi^\top\in\mathcal{O}_{n_x,2k}$, let $\Phi_i^\top$ denote the nearest element in the cover. By triangle inequality and $(a+b)^2\le 2(a^2+b^2)$:

$$
\begin{aligned}
&\sum_{h=1}^{H}\left\|(\Phi(\mathbf{X}^{(h)})^\top\mathbf{X}^{(h)}\Phi^\top)^{-1/2}\Phi(\mathbf{X}^{(h)})^\top\mathbf{Z}^{(h)}\right\|_F^2\\
&\quad=\sum_{h=1}^{H}\left\|P_{\mathbf{X}^{(h)}\Phi}\mathbf{Z}^{(h)}\right\|_F^2\\
&\quad\le 2\sum_{h=1}^{H}\left\|(P_{\mathbf{X}^{(h)}\Phi}-P_{\mathbf{X}^{(h)}\Phi_i})\mathbf{Z}^{(h)}\right\|_F^2+2\sum_{h=1}^{H}\left\|P_{\mathbf{X}^{(h)}\Phi_i}\mathbf{Z}^{(h)}\right\|_F^2\\
&\quad\le 2\sum_{h=1}^{H}\left\|P_{\mathbf{X}^{(h)}\Phi}-P_{\mathbf{X}^{(h)}\Phi_i}\right\|_2^2\left\|\mathbf{Z}^{(h)}\right\|_F^2+2\sum_{h=1}^{H}\left\|P_{\mathbf{X}^{(h)}\Phi_i}\mathbf{Z}^{(h)}\right\|_F^2.
\end{aligned}
$$

By Lemma B.2,

$$
\begin{aligned}
\left\|P_{\mathbf{X}^{(h)}\Phi}-P_{\mathbf{X}^{(h)}\Phi_i}\right\| &\le \left\|(\Phi(\mathbf{X}^{(h)})^\top\mathbf{X}^{(h)}\Phi^\top)^{-1}\Phi(\mathbf{X}^{(h)})^\top\right\|\left\|\mathbf{X}^{(h)}(\Phi-\Phi_i)\right\|\\
&\le \frac{1.1}{0.9}\frac{\lambda_{\max}(\Sigma_x^{(h)})}{\lambda_{\min}(\Sigma_x^{(h)})}\varepsilon\lesssim\frac{\overline{\lambda}}{\underline{\lambda}}\varepsilon.
\end{aligned}
$$

Next, we need to bound $\sum_{h=1}^{H}\left\|\mathbf{Z}^{(h)}\right\|_F^2$. Because each $z_i^{(h)}[t]$ is i.i.d. $\mathcal{N}(0,\sigma_z^2 I)$, this sum is distributed as a $\sigma_z^2\psi$, where ψ is a χ^2 random variable with HN_1Tn_u degrees of freedom. Hence by standard χ^2 tail bounds, with probability at least $1-\delta/20$, $\sum_{h=1}^{H}\left\|\mathbf{Z}^{(h)}\right\|_F^2\lesssim\sigma_z^2(HN_1Tn_u+\log(1/\delta))$. This prompts us to select $\varepsilon \lesssim \frac{k}{N_1 T\bar{\lambda}/\underline{\lambda}}$, from which we conclude:

$$\sum_{h=1}^{H} \left\|P_{\mathbf{X}^{(h)}\Phi} - P_{\mathbf{X}^{(h)}\Phi_i}\right\|^2 \left\|\mathbf{Z}^{(h)}\right\|_F^2 \lesssim \sigma_z^2 k H n_u.$$

By Lemma B.1, we can then bound the size of the ε-covering of $\mathcal{O}_{n_x,2k}$ by:

$$N \leq \left(\frac{c\sqrt{k}}{k} N_1 T \frac{\bar{\lambda}}{\underline{\lambda}}\right)^{2n_x k}.$$

So now we take $\delta' = \delta/(20N)$, and conclude. In particular, by union bounding over the elements of $\{\Phi_i\}_{i=1}^N$, we have that with probability at least $1 - \frac{\delta}{5}$,

$$\begin{aligned}\sup_{\Phi^\top \in \mathcal{O}_{n_x,2k}} \sum_{h=1}^{H} &\left\|(\Phi(\mathbf{X}^{(h)})^\top \mathbf{X}^{(h)}\Phi^\top)^{-1/2}\Phi(\mathbf{X}^{(h)})^\top \mathbf{Z}^{(h)}\right\|_F^2 \\ &\lesssim \sigma_z^2 k H n_u + n_x k \log\left(N_1 T \frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right).\end{aligned}$$

The probability of $1 - \frac{\delta}{5}$ arises from union bounding over Equation (22), which holds with probability at least $1 - \frac{\delta}{10}$, the bound on $\sum_{h=1}^{H} \left\|\mathbf{Z}^{(h)}\right\|_F^2$ which holds with probability at least $1 - \frac{\delta}{20}$, and the bound in Equation (24), which holds for all $\Phi_i \in \{\Phi_i\}_{i=1}^N$ with probability at least $1 - \frac{\delta}{20}$.

■

Lemma B.5 (Lemma A.7 in Du et al. (2020)) *For any two matrices A_1 and A_2 with the same number of columns which satisfy $A_1^\top A_1 \succeq A_2^\top A_2$, then for any matrix B, we have*

$$A_1^\top P_{A_1 B}^\perp A_1 \succeq A_2^\top P_{A_2 B}^\perp A_2.$$

As a consequence, for any matrices B and B', we have

$$\left\|P_{A_1 B}^\perp A_1 B'\right\|_F^2 \geq \left\|P_{A_2 B}^\perp A_2 B'\right\|_F^2.$$

Lemma B.6 (target controller concentration) *Under the setting and data assumptions of Theorem 3.1, with probability at least $1 - \frac{2\delta}{5}$,*

$$\frac{1}{N_2 T}\left\|\mathbf{X}^{(H+1)}\left(\hat{F}^{(H+1)}\hat{\Phi} - F_\star^{(H+1)}\Phi_\star\right)^\top\right\|_F^2 \lesssim \bar{\sigma}_z^2\left(\frac{k n_x \log\left(N_1 T \frac{\bar{\lambda}}{\underline{\lambda}}\right)}{c N_1 T H} + \frac{k n_u + \log(\frac{1}{\delta})}{N_2 T}\right).$$

Proof By the optimality of $\hat{\Phi}, \hat{F}^{(1)}, \ldots, \hat{F}^{(H)}$ for Problem (1), we know that

$$\hat{F}^{(h)} = \left(\hat{\Phi}\left(\mathbf{X}^{(h)}\right)^\top \mathbf{X}^{(h)}\hat{\Phi}^\top\right)^{-1}\hat{\Phi}\left(\mathbf{X}^{(h)}\right)^\top \mathbf{U}^{(h)}.$$

Therefore, $\mathbf{X}^{(h)}\left(\hat{F}^{(h)}\hat{\Phi}\right)^\top = P_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\mathbf{U}^{(h)} = P_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\left(\mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top + \mathbf{Z}^{(h)}\right)$. Then by applying Lemma B.4, we have that with probability at least $1-\frac{\delta}{5}$,

$$\begin{aligned}
\bar{\sigma}_z^2\left(kmH + kn\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right)\right) &\gtrsim \sum_{h=1}^H \left\|\mathbf{X}^{(h)}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \\
&= \sum_{h=1}^H \left\|P_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\left(\mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top + Z^{(h)}\right) - \mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \\
&= \sum_{h=1}^H \left\|(P_{\mathbf{X}^{(h)}\hat{\Phi}^\top} - I)\mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top + P_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\mathbf{Z}^{(h)}\right\|_F^2 \\
&= \sum_{h=1}^H \left\|P^\perp_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 + \left\|P_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\mathbf{Z}^{(h)}\right\|_F^2 \\
&\geq \sum_{h=1}^H \left\|P^\perp_{\mathbf{X}^{(h)}\hat{\Phi}^\top}\mathbf{X}^{(h)}\left(F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \\
&\overset{(i)}{\geq} 0.9N_1T\sum_{h=1}^H \left\|P^\perp_{\left(\Sigma^{(h)}\right)^{1/2}\hat{\Phi}^\top}\left(\Sigma^{(h)}\right)^{1/2}\left(F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \\
&\overset{(ii)}{\geq} 0.9cN_1T\sum_{h=1}^H \left\|P^\perp_{\left(\Sigma^{(H+1)}\right)^{1/2}\hat{\Phi}^\top}\left(\Sigma^{(H+1)}\right)^{1/2}\left(F_\star^{(h)}\Phi_\star\right)^\top\right\|_F^2 \\
&= 0.9cN_1T\left\|P^\perp_{\left(\Sigma^{(H+1)}\right)^{1/2}\hat{\Phi}^\top}\left(\Sigma^{(H+1)}\right)^{1/2}\Phi_\star^\top\mathbf{F}_\star\right\|_F^2
\end{aligned} \tag{25}$$

where (i) follow from the inequality in Lemma A.3 in addition to Lemma B.5, while (ii) follows from the inequality in (5) (note that we have already conditioned on this event through Lemma B.4) in addition to Lemma B.5. Now, letting $\Phi = \begin{bmatrix}\hat{\Phi}^\top & \hat{\Phi}_\star^\top\end{bmatrix}$, Lemma A.4 gives us $\frac{1}{N_2T}\Phi^\top\left(\mathbf{X}^{(H+1)}\right)^\top\mathbf{X}^{(H+1)}\Phi \preceq 1.1\Phi^\top\Sigma_x^{(H+1)}\Phi$. Combining with Lemma B.5, we have that with probability at least $1-\frac{\delta}{10}$,

$$1.1\left\|P^\perp_{\left(\Sigma^{(H+1)}\right)^{1/2}\hat{\Phi}^\top}\left(\Sigma^{(H+1)}\right)^{1/2}\Phi_\star^\top\mathbf{F}_\star\right\|_F^2 \geq \frac{1}{N_2T}\left\|P^\perp_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\mathbf{F}_\star\right\|_F^2.$$

Plugging this in above, we have that with probability at least $1-\frac{2\delta}{5}$,

$$\bar{\sigma}_z^2\left(kmH + kn\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right)\right) \gtrsim \frac{cN_1}{N_2}\left\|P^\perp_{X^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\mathbf{F}_\star\right\|_F^2.$$

To conclude the proof, note that $\mathbf{X}^{(H+1)}\hat{\Phi}^\top\left(\hat{F}^{(H+1)}\right)^\top = P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{U}^{(h)}$ and $\mathbf{F}_\star^\dagger := \mathbf{F}_\star^\top(\mathbf{F}_\star\mathbf{F}_\star^\top)^{-1}$ such that $\mathbf{F}_\star\mathbf{F}_\star^\dagger = I_k$. Thus

$$\begin{aligned}
&\frac{1}{N_2T}\left\|\mathbf{X}^{(H+1)}\left(\hat{F}^{(H+1)}\hat{\Phi} - F_\star^{(H+1)}\Phi_\star\right)^\top\right\|_F^2 \\
&= \frac{1}{N_2T}\left\|P^\perp_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\left(F_\star^{(H+1)}\Phi_\star\right)^\top + P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \\
&\leq \frac{1}{N_2T}\left\|P^\perp_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\left(F_\star^{(H+1)}\Phi_\star\right)^\top\right\|_F^2 + \frac{1}{N_2T}\left\|P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N_2T}\mathrm{Tr}\Big(P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\Big(F_\star^{(H+1)}\Big)^\top F_\star^{(H+1)}\Phi_\star\Big(\mathbf{X}^{(H+1)}\Big)^\top P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\Big) \\
&\qquad + \frac{1}{N_2T}\left\|P_{X^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \\
&= \frac{1}{N_2T}\mathrm{Tr}\Big(P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\mathbf{F}_\star\mathbf{F}_\star^\dagger\Big(F_\star^{(H+1)}\Big)^\top F_\star^{(H+1)}\Big(\mathbf{F}_\star^\dagger\Big)^\top\mathbf{F}_\star^\top\Phi_\star\Big(\mathbf{X}^{(H+1)}\Big)^\top P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\Big) \\
&\qquad + \frac{1}{N_2T}\left\|P_{X^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \\
&\le \frac{1}{N_2T}\left\|\mathbf{F}_\star^\dagger\Big(F_\star^{(H+1)}\Big)^\top F_\star^{(H+1)}\Big(\mathbf{F}_\star^\dagger\Big)^\top\right\|_2\mathrm{Tr}\Big(P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\mathbf{F}_\star\mathbf{F}_\star^\top\Phi_\star\Big(\mathbf{X}^{(H+1)}\Big)^\top P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\Big) \\
&\qquad + \frac{1}{N_2T}\left\|P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \\
&\lesssim \frac{1}{N_2T}\left\|F_\star^{(H+1)}\Big(\mathbf{F}_\star\mathbf{F}_\star^\top\Big)^{-1}\Big(F_\star^{(H+1)}\Big)^\top\right\|\left\|P^{\perp}_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{X}^{(H+1)}\Phi_\star^\top\mathbf{F}_\star\right\|_F^2 + \frac{1}{N_2T}\left\|P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \\
&\lesssim \frac{1}{cN_1TH}\bar{\sigma}_z^2\left(kn_uH + kn_x\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{1}{N_2T}\left\|P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2
\end{aligned}
$$

where the last line follows from the inequality in (25) and applying Assumption 3.1 on $\left\|F_\star^{(H+1)}(\mathbf{F}_\star\mathbf{F}_\star^\top)^{-1}\Big(F_\star^{(H+1)}\Big)^\top\right\|_2$. As $\hat{\Phi}$ is independent from $\mathbf{X}^{(H+1)}$ and $\mathbf{Z}^{(H+1)}$, the second term may be bounded by applying Proposition B.1 as in Equation (24). In particular, with probability at least $1-\frac{\delta}{5}$,

$$
\left\|P_{\mathbf{X}^{(H+1)}\hat{\Phi}^\top}\mathbf{Z}^{(H+1)}\right\|_F^2 \le \sigma_z^2(kn_u + \log(1/\delta)).
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{N_2T}\left\|\mathbf{X}^{(H+1)}\Big(\hat{F}^{(H+1)}\hat{\Phi} - F_\star^{(H+1)}\Phi_\star\Big)^\top\right\|_F^2 \\
&\lesssim \frac{1}{cN_1TH}\sigma_z^2\left(kn_uH + kn_x\log\left(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\right) + \log\left(\frac{1}{\delta}\right)\right) + \frac{\sigma_z^2kn_u + \sigma_z^2\log(\frac{1}{\delta})}{N_2T} \\
&= \sigma_z^2\left(\frac{kn_uH}{cN_1TH} + \frac{kn_x\log\Big(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\Big)}{cN_1TH} + \frac{\log(1/\delta)}{cN_1TH} + \frac{kn_u + \log(\frac{1}{\delta})}{N_2T}\right) \\
&\lesssim \sigma_z^2\left(\frac{kn_x\log\Big(N_1T\frac{\bar{\lambda}}{\underline{\lambda}}\Big)}{cN_1TH} + \frac{kn_u + \log(\frac{1}{\delta})}{N_2T}\right).
\end{aligned}
$$

■

Theorem 3.1 (target task excess risk bound) *Given $\delta \in (0,1)$, suppose that*

$$
\begin{aligned}
N_1T &\gtrsim \max_{h\in[H]}\mathcal{J}\Big(A + BK^{(h)}\Big)^2\kappa\Big(\Sigma_x^{(h)}\Big)(n_x + \log(H/\delta)), \\
N_2T &\gtrsim \mathcal{J}\Big(A + BK^{(H+1)}\Big)^2\kappa\Big(\Sigma_x^{(H+1)}\Big)(k + \log(1/\delta)).
\end{aligned}
$$

Define $\mathcal{P}_{0:T-1}^{(H+1)}$ as the distribution over target task trajectories $(x^{(H+1)}[0],\cdots,x^{(H+1)}[T-1])$. Then with probability at least $1-\delta$, the excess risk of the learned representation $\hat{\Phi}$ and target task weights $\hat{F}^{(H+1)}$ is *bounded by*

$$\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) := \frac{1}{2T}\mathbb{E}_{\mathcal{P}_{0:T-1}^{(H+1)}}\left[\sum_{t=0}^{T-1}\left\|(F_\star^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi})x^{(H+1)}[t]\right\|^2\right]$$
$$\lesssim \sigma_z^2\left(\frac{kn_x \log\left(N_1 T\frac{\bar{\lambda}}{\underline{\lambda}}\right)}{cN_1TH} + \frac{kn_u + \log(\frac{1}{\delta})}{N_2T}\right). \tag{7}$$

Proof The excess risk may be written as follows:

$$\begin{aligned}
\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) &= \frac{1}{2T}\mathbb{E}_{x[0],w[0],\dots,w[T-1]}\left[\sum_{t=0}^{T-1}\left\|(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi})x[t]\right\|_2^2\right] \\
&= \frac{1}{2T}\mathbb{E}_{x[0],w[0],\dots,w[T-1]}\left[\sum_{t=0}^{T-1}\mathrm{Tr}\left(\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)x[t]x[t]^\top\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\right)\right] \\
&= \frac{1}{2T}\sum_{t=0}^{T-1}\mathrm{Tr}\left(\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\Sigma_x^{(H+1)}\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\right) \\
&= \frac{1}{2}\mathrm{Tr}\left(\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\Sigma_x^{(H+1)}\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\right).
\end{aligned}$$

Then by applying the inequality in Lemma A.4 with $\Phi = \begin{bmatrix}\hat{\Phi} & \Phi_\star\end{bmatrix}$, we have

$$0.9\Phi^\top\Sigma_x^{(H+1)} \preceq \frac{1}{N_2T}\Phi^\top\left(\mathbf{X}^{(H+1)}\right)^\top\mathbf{X}^{(H+1)}\Phi.$$

Using this result this above, we have that

$$\begin{aligned}
\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) &\le \frac{1}{1.8N_2T}\mathrm{Tr}\left(\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\left(\mathbf{X}^{(H+1)}\right)^\top\mathbf{X}^{(H+1)}\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\right) \\
&\le \frac{1}{1.8N_2T}\left\|\left(F^{(H+1)}\Phi_\star - \hat{F}^{(H+1)}\hat{\Phi}\right)\mathbf{X}^{(H+1)}\right\|_F^2.
\end{aligned}$$

The claim then follows by application of Lemma B.6. ■

C Bounding the imitation gap

We recall that the issue with translating a bound on the excess risk of the target task $\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})$ into a bound on the tracking error between the closed-loop learned and expert trajectories comes from the fundamental distribution shift between the expert trajectories seen during training and the trajectories generated by running the learned controller in closed-loop. This has traditionally been a difficult problem to analyze due to the circular nature of requiring the feedback controller error to be small to guarantee small state deviation, which depends on the state deviation being small to begin with.

Toward addressing this issue, recent work by Pfrommer et al. (2022) proposes a theoretical framework for non-linear systems to bound the tracking error by the imitation error via matching the higher-order input/state derivatives $D_x^p\pi(x)$ of the expert controller. For linear systems, matching the Jacobian is sufficient, where we note the Jacobian of a linear controller $u[t] = Kx[t]$ is simply K. Furthermore, a generalization bound on the empirical risk of a learned controller such as Theorem 3.1 implicitly bounds the controller error $\left\|\hat{K} - K^{(H+1)}\right\|$.

However, the framework described in Pfrommer et al. (2022) crucially relies on assuming a noiseless system, whereas the problem considered in this work is only non-trivial in the presence of process noise. Furthermore, due to the lack of excitatory noise and the generality of non-linear systems, the tracking error bounds in Pfrommer et al. (2022) scale with $\tilde{\mathcal{O}}\left(\frac{\#\text{param}}{N_2\delta'}\right)$, where δ' is the failure probability over a new trajectory. To that end, the main goal of this section is to introduce bounds on the tracking error that scale multiplicatively with the favorable generalization bounds on the excess risk, which improve with the trajectory length T, and also improving the system-agnostic Markov scaling $1/\delta'$ to $\log(1/\delta')$ in our linear systems setting.

Toward proving a bound on the imitation gap, we first introduce the notion of general stochastic incremental stability (δ-SISS). We recall definitions of standard comparison functions (Khalil, 1996): a function $\gamma(x)$ is class $\mathcal{K}$ if it is continuous, strictly increasing, and satisfies $\gamma(0) = 0$. A function $\beta(x, t)$ is class $\mathcal{KL}$ if it is continuous, $\beta(\cdot, t)$ is class $\mathcal{K}$ for each fixed t, and $\beta(x, \cdot)$ is decreasing for each fixed x.

Definition C.1 (δSISS) *Consider the discrete-time control system $x[t+1] = f(x[t], u[t])$ subject to input perturbations $\Delta[t]$ and additive zero-mean process noise $\{w[t]\}_{t\geq 0}$. The system $f(x[t], u[t])$ is* incremental stochastic input-to-state stable *(δ-SISS) if for all initial conditions $x[0], y[0] \in \mathcal{X}$, perturbation sequences $\{\Delta[0]\}_{t\geq 0}$ and noise realizations $\{w[0]\}_{t\geq 0}$ there exists a class $\mathcal{KL}$ function β and a class $\mathcal{K}$ function γ such that the trajectories $x[t+1] = f(x[t], u[t]) + w[t]$ and $y[t+1] = f(y[t], u[t] + \Delta[t]) + w[t]$ satisfy for all $t \in \mathbb{N}$*

$$\|x[t] - y[t]\| \leq \beta(\|x[0] - y[0]\|, t) + \gamma\left(\max_{0\leq k\leq t-1} \|\Delta[k]\|\right). \tag{26}$$

Lemma C.1 (Stable linear dynamical systems are δSISS) *Let (A, B) describe a linear time-invariant system $x[t+1] = Ax[t] + Bu[t]$. Let A be Schur-stable, i.e. $\rho(A) < 1$, where $\rho(\cdot)$ denotes the spectral radius. Recall the constant $\mathcal{J}(A) := \sum_{t\geq 0} \left\|A^t\right\|$. Furthermore, for any given ν such that $\rho(A) < \nu < 1$, define the corresponding constant*

$$\tau(A, \nu) := \sup_{k\in\mathbb{N}} \frac{\left\|A^k\right\|}{\nu^k},$$

which is guaranteed to be finite by Gelfand's formula (Horn and Johnson, 2012, Corollary 5.6.13). The linear system (A, B) is incrementally stochastic input-to-state stable, where we may set

$$\begin{aligned} \beta(x, t) &:= \tau(A, \nu)\nu^t x \\ \gamma(x) &:= \mathcal{J}(A) \|B\| x. \end{aligned}$$

Proof Let us define the nominal and perturbed states x_t and y_t defined by

$$\begin{aligned} x[t+1] &= Ax[t] + Bu[t] + w[t], \quad x[0] = \xi_1 \\ y[t+1] &= Ay[t] + B(u[t] + \Delta[t]) + w[t], \quad y[0] = \xi_2. \end{aligned}$$

We observe that by linearity, we may write

$$x[t] = A^t x[0] + \sum_{k=0}^{t-1} (A^{t-k-1}Bu[k] + A^{t-k-1}w[k])$$

$$y[t] = A^t y[0] + \sum_{k=0}^{t-1} (A^{t-k-1} Bu[k] + A^{t-k-1} B\Delta[k] + A^{t-k-1} w[k]).$$

Therefore, their difference can be written as

$$\begin{aligned} x[t] - y[t] &= A^t (x[0] - y[0]) - \sum_{k=0}^{t-1} A^{t-k-1} B\Delta[k] \\ \implies \|x[t] - y[t]\| &\leq \left\|A^t\right\| \|x[0] - y[0]\| + \left(\sum_{k=0}^{t-1} \left\| A^{t-k-1} \right\| \right) \|B\| \max_{0 \leq k \leq t-1} \|\Delta[k]\| \\ &\leq \tau(A, \nu) \nu^t \|x[0] - y[0]\| + \mathcal{J}(A) \|B\| \max_{0 \leq k \leq t-1} \|\Delta[k]\| \end{aligned}$$

where we can match the functions $\beta(\cdot, t)$ and $\gamma(\cdot)$ to their respective quantities above. Crucially, we observe that due to linearity, the noise terms in $x[t]$ and $y[t]$ cancel out, and therefore δSISS is effectively the same as the standard δISS. ■

By showing stable linear systems are δSISS, we may adapt Corollary A.1 from Pfrommer et al. (2022) to yield the following guarantee with respect to the expert closed-loop system of the target task $A + BK^{(H+1)}$, which allows us to bound the imitation gap in terms of the maximal deviation between the learned and target controller inputs.

Proposition C.1 *Let the target task closed-loop system $(A + BK^{(H+1)}, B)$ be δSISS for $\beta(\cdot, t)$ and $\gamma(\cdot)$ defined in Lemma C.1. For a given test controller K, and realizations of randomness $x[0] \sim \mathcal{N}(0, \Sigma_x^{(H+1)})$, $z[t] \sim \mathcal{N}(0, \sigma_z^2 I)$, $w[t] \sim N(0, \Sigma_w^{(H+1)}), t = 0, \dots, T-1$, we write*

$$\begin{aligned} x_\star[t+1] &= (A + BK^{(H+1)}) x_\star[t] + Bz[t] + w[t], \quad x_\star[0] = x[0] \\ \hat{x}[t+1] &= (A + BK)\hat{x}[t] + Bz[t] + w[t], \quad \hat{x}[0] = x[0]. \end{aligned} \tag{27}$$

In other words, $x_\star[t]$ is the state from rolling out the expert target task controller $K^{(H+1)}$, and $\hat{x}[t]$ is the state from rolling out the test controller K under the same conditions. If K satisfies

$$\left\| K - K^{(H+1)} \right\| \leq \frac{1}{2\mathcal{J}(A + BK^{(H+1)}) \|B\|},$$

then the tracking error satisfies for any $T \geq 1$,

$$\max_{1 \leq t \leq T} \|\hat{x}[t] - x_\star[t]\| \leq \max_{0 \leq t \leq T-1} 2\mathcal{J}(A + BK^{(H+1)}) \|B\| \left\| \left(K - K^{(H+1)} \right) x_\star[t] \right\|.$$

In other words, Proposition C.1 allows us to bound the imitation gap by the excess risk induced by controller K, as long as K is sufficiently close to the expert controller $K^{(H+1)}$ in the spectral norm.

Proof We borrow Proposition 3.1 from Pfrommer et al. (2022), which states a general result for non-linear controllers and non-linear systems, but in the noiseless setting, and specify it to the linear systems setting with process noise. In particular, the result states that for a given state-feedback controller $\pi(x)$ and an expert state-feedback controller $\pi_\star(x)$ that induces a δ-SISS closed-loop system $x_\star[t+1] = f(x_\star[t], \pi_\star(x_\star[t])) + w[t]$ (the result can be extended from δ-ISS with no modification to the proof), then for a given $\varepsilon > 0$, if π

satisfies the following on an expert trajectory generated by realizations of randomness $x[0], w[0], \dots, w[T-1]$:

$$\max_{0 \le t \le T-1} \sup_{\|\delta\| \le \varepsilon} \|\pi_\star(x_\star[t] + \delta) - \pi(x_\star[t] + \delta)\| \le \gamma^{-1}(\varepsilon), \tag{28}$$

where γ^{-1} is the (generalized) inverse of the class-$\mathcal{K}$ function $\gamma(\cdot)$, then the imitation gap is bounded by

$$\max_{0 \le t \le T-1} \|\hat{x}[t] - x_\star[t]\| \le \varepsilon,$$

where $\hat{x}[t]$ are the states generated by running controller π in closed-loop on the same noise realizations as those generating $x_\star[t]$. Now, substituting $\pi(x) := Kx$, $\pi_\star(x) := K^{(H+1)}x$, and $\gamma(x) := \mathcal{J}(A + BK^{(H+1)}) \|B\| \, x$, such that $\gamma^{-1}(x) = \big(\mathcal{J}(A + BK^{(H+1)}) \|B\|\big)^{-1} x$, we observe

$$\begin{aligned}
&\max_{0 \le t \le T-1} \sup_{\|\delta\| \le \varepsilon} \|\pi_\star(x_\star[t] + \delta) - \pi(x_\star[t] + \delta)\| \\
&= \max_{0 \le t \le T-1} \sup_{\|\delta\| \le \varepsilon} \left\|(K - K^{(H+1)})x_\star[t] + (K - K^{(H+1)})\delta\right\| \\
&\le \max_{0 \le t \le T-1} \left\|(K - K^{(H+1)})x_\star[t]\right\| + \left\|K - K^{(H+1)}\right\| \varepsilon.
\end{aligned}$$

Therefore, in order to satisfy (28), it suffices to satisfy

$$\max_{0 \le t \le T-1} \left\|(K - K^{(H+1)})x_\star[t]\right\| + \left\|K - K^{(H+1)}\right\| \varepsilon \le \Big(\mathcal{J}(A + BK^{(H+1)}) \|B\|\Big)^{-1} \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, setting $\varepsilon := \max_t 0.5\big(\mathcal{J}(A + BK^{(H+1)}) \|B\|\big)^{-1} \left\|(K - K^{(H+1)})x_\star[t]\right\|$, it is sufficient for

$$\left\|K - K^{(H+1)}\right\| \le \frac{1}{2\mathcal{J}(A + BK^{(H+1)}) \|B\|},$$

to satisfy the above inequality, which leads to the following bound on the imitation gap

$$\max_{0 \le t \le T-1} \|\hat{x}[t] - x_\star[t]\| \le \varepsilon = \max_{0 \le t \le T-1} 2\mathcal{J}(A + BK^{(H+1)}) \|B\| \left\|Kx_\star[t] - K^{(H+1)}x_\star[t]\right\|.$$

This completes the proof. ■

Before moving on, we discuss a few qualitative properties of Proposition C.1. First off, a bound on the imitation gap is somewhat higher resolution than the LQR-type bounds in Fazel et al. (2018); Mania et al. (2019), where the quantity of interest is the difference in the expected infinite-horizon costs of the two controllers, which does not directly imply a bound on the deviation between the states induced by the two controllers at a given time. Furthermore, we note Proposition C.1 is meaningful *precisely* because there is driving process noise. If the target task is noiseless, and test controller K is stabilizing, then $\|\hat{x}[t] - x_\star[t]\|$ is trivially upper bounded by an exponentially decaying quantity, which for a sufficiently long horizon T will beat out any generalization bound scaling with $\mathrm{poly}(N_1, N_2, H, T)$–however, this noiseless setting is correspondingly uninteresting to study as a statistical problem.

We propose the following key result in the form of a high probability bound on the tracking error.

Theorem C.1 (Full version of Theorem 3.2, (9)) *Let the target task closed-loop system $(A + BK^{(H+1)}, B)$ be δSISS for $\beta(\cdot, t)$ and $\gamma(\cdot)$ as defined in Lemma C.1. Given a generalization bound* *on the excess risk of the learned representation and target task weights* $(\hat{\Phi}, \hat{F}^{(H+1)})$ *(such as Theorem 3.1) of the form: with probability greater than* $1-\delta$

$$\mathrm{ER}\Big(\hat{\Phi}, \hat{F}^{(H+1)}\Big) \le f(N_1, T, H, N_2, \delta),$$

we have the following bound on the tracking error. Assuming we have enough samples such that

$$f(N_1, T, H, N_2, \delta) \le \frac{\lambda_{\min}(\Sigma_x^{(H+1)})}{8\mathcal{J}(A+BK^{(H+1)})^2 \left\|B\right\|^2},$$

then with probability greater than $1-\delta-\delta'$*, for a new target task trajectory sampled with i.i.d. process randomness* $x[0] \sim \Sigma_x^{(H+1)}$*,* $w[t] \sim \Sigma_w^{(H+1)}$*,* $t = 0, \dots, T-1$*, the tracking error satisfies*

$$\max_{1 \le t \le T} \left\|\hat{x}[t] - x_\star[t]\right\|^2 \le 4\mathcal{J}(A+BK^{(H+1)})^2 \left\|B\right\|^2 \left(1 + \frac{8\sqrt{3}}{3}\log\left(\frac{T}{\delta'}\right)\right) \mathrm{ER}(\hat{\Phi}, \hat{W}^{H+1}) \tag{29}$$

$$\lesssim \mathcal{J}(A+BK^{(H+1)})^2 \left\|B\right\|^2 \log\left(\frac{T}{\delta'}\right) f(N_1, T, H, N_2, \delta), \tag{30}$$

where the expert and learned trajectory states $\hat{x}[t]$ *and* $x_\star[t]$ *are as defined in Proposition C.1.*

Proof By Proposition C.1, we have that if the learned controller $\hat{K} = \hat{F}^{(H+1)}\hat{\Phi}$ satisfies

$$\left\|\hat{K} - K^{(H+1)}\right\| \le \frac{1}{2\mathcal{J}(A+BK^{(H+1)}) \left\|B\right\|},$$

where we plugged in the $\beta(\cdot, t)$, $\gamma(\cdot)$ functions derived from Lemma C.1 on the target task expert closed-loop system $(A + BK^{(H+1)}, B)$, then we have for learned and expert trajectories generated by independent instances of randomness $x[0], w[0], \dots, w[T-1]$ the following imitation gap:

$$\max_{1 \le t \le T} \left\|\hat{x}[t] - x_\star[t]\right\| \le \max_{0 \le t \le T-1} 2\mathcal{J}(A+BK^{(H+1)}) \left\|B\right\| \left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|.$$

In order derive sample complexity bounds from this, we observe a generalization bound on the excess risk

$$\begin{aligned}
\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)}) &:= \frac{1}{2}\mathbb{E}_{\{x_\star[t]\}\sim \mathcal{P}^{(H+1)}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|^2\right] \\
&= \frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\mathrm{Tr}\Big((\hat{K} - K^{(H+1)})x[t]x[t]^\top(\hat{K} - K^{(H+1)})^\top\Big)\Big] \\
&= \frac{1}{2}\mathrm{Tr}\Big((\hat{K} - K^{(H+1)})\Sigma_x^{(H+1)}(\hat{K} - K^{(H+1)})^\top\Big) \\
&\le f(N_1, T, H, N_2, \delta) \quad \text{w.p. } \ge 1-\delta
\end{aligned}$$

directly implies a generalization bound on the Frobenius norm deviation between the learned and expert controllers:

$$\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\Big((\hat{K} - K^{(H+1)})\Sigma_x^{(H+1)}(\hat{K} - K^{(H+1)})^\top\Big) \ge \frac{1}{2}\lambda_{\min}(\Sigma_x^{(H+1)})\left\|\hat{K}^{(H+1)} - K^{(H+1)}\right\|_F^2 \\
\implies &\left\|\hat{K}^{(H+1)} - K^{(H+1)}\right\|_F^2 \le \frac{2}{\lambda_{\min}(\Sigma_x^{(H+1)})} f(N_1, T, H, N_2, \delta) \text{ w.p. } \ge 1-\delta.
\end{aligned}$$

Since the Frobenius norm upper bounds the spectral norm, it suffices to have enough samples N_1, N_2, T, H such that

$$f(N_1, T, H, N_2, \delta) \leq \frac{\lambda_{\min}(\Sigma_x^{(H+1)})}{8\mathcal{J}(A + BK^{(H+1)})^2 \left\|B\right\|^2}$$

If the sample complexity requirement is satisfied, then we have with probability greater than $1 - \delta$ that the spectral norm gap is satisfied

$$\left\|\hat{K} - K^{(H+1)}\right\| \leq \frac{1}{4\mathcal{J}(A + BK^{(H+1)})^2 \left\|B\right\|^2},$$

which in turn implies the bound on the tracking error

$$\max_{1 \leq t \leq T} \left\|\hat{x}[t] - x_\star[t]\right\|^2 \leq \max_{0 \leq t \leq T-1} 4\mathcal{J}(A + BK^{(H+1)})^2 \left\|B\right\|^2 \left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|^2. \tag{31}$$

In order to convert the RHS of the above expression, which involves a maximum over time of the imitation error between $\hat{K}$ and $K^{(H+1)}$, into something involving the excess risk of the controller $\hat{K}$, which is the expected value of the imitation error, we again appeal to the Hanson-Wright inequality, adapted specifically for Gaussian quadratic forms in Proposition A.1.

We recall the tracking error is with probability greater than $1 - \delta$ bounded by

$$\max_{1 \leq t \leq T} \left\|\hat{x}[t] - x_\star[t]\right\|^2 \leq \max_{0 \leq t \leq T-1} \frac{4\tau^2 \left\|B\right\|_2^2}{(1-\nu)^2} \left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|^2.$$

In order to bound the RHS of the above inequality by a multiplicative factor of the excess risk $\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})$, we apply Proposition A.1, setting

$$R := \left(\hat{K} - K^{(H+1)}\right)\left(\Sigma_x^{(H+1)}\right)^{1/2}, \quad z_t := \left(\Sigma_x^{(H+1)}\right)^{-1/2} x_\star[t]. \tag{32}$$

By an application of the union bound, we have

$$\begin{aligned}
\mathbb{P}\left[\max_{t \leq T-1} z_t^\top R^\top R z_t \geq (1+\varepsilon)\left\|R\right\|_F^2\right] &\leq \sum_{t=0}^{T-1} \mathbb{P}\left[z_t^\top R^\top R z_t \geq (1+\varepsilon)\left\|R\right\|_F^2\right] \\
&\leq T \exp\left(-\frac{1}{\sqrt{3}} \min\left\{\frac{9\varepsilon^2}{64}, \frac{3\varepsilon}{8}\right\} \frac{\left\|R\right\|_F^2}{\left\|R\right\|^2}\right).
\end{aligned}$$

Setting the last line less than the desired failure probability $\delta' \in (0,1)$, we get

$$\min\left\{\frac{9\varepsilon^2}{64}, \frac{3\varepsilon}{8}\right\} \geq \sqrt{3}\frac{\left\|R\right\|^2}{\left\|R\right\|_F^2} \log\left(\frac{T}{\delta'}\right).$$

Since $\left\|R\right\| \leq \left\|R\right\|_F$, it suffices to choose

$$\varepsilon = \frac{8\sqrt{3}}{3} \log\left(\frac{T}{\delta'}\right),$$

such that with probability greater than $1 - \delta'$

$$\max_{0 \leq t \leq T-1} \left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|^2 \leq \left(1 + \frac{8\sqrt{3}}{3}\log\left(\frac{T}{\delta'}\right)\right) \left\|\left(\hat{K} - K^{(H+1)}\right)\left(\Sigma_x^{(H+1)}\right)^{1/2}\right\|_F^2$$

$$= \left(1 + \frac{8\sqrt{3}}{3}\log\left(\frac{T}{\delta'}\right)\right)\mathrm{ER}\left(\hat{\Phi}, \hat{F}^{(H+1)}\right).$$

Plugging this back into the tracking error bound and union bounding over the generalization bound event on $\mathrm{ER}\left(\hat{\phi}, \hat{F}^{(H+1)}\right)$ of probability $1-\delta$ and the concentration event of probability $1-\delta'$ yields the desired result.

■

The high probability bound introduced in Theorem C.1 provides a very high resolution control over the deviation between closed-loop learned and expert states, where we control the maximum deviation of states over time by the *expected time-averaged* excess risk of the learned controller, accruing only a $\log(T)$ factor in the process. Furthermore, our high-probability bounds are multiplicative with respect to only $\log(1/\delta)$ rather than the naive Markov's inequality factor $1/\delta$, which is immediately higher-resolution than the corresponding in-expectation bound (for example derived from Proposition C.2), on which only Markov's inequality can be applied without more detailed analysis. However, this being said, in much of existing literature in control and reinforcement learning, the performance of a policy is evaluated as an expected cost/reward over the trajectories it generates, for example LQG cost (Fazel et al., 2018), or expected reward in online imitation learning (Ross et al., 2011). This therefore motivates understanding whether the tools we develop directly imply bounds in-expectation on general cost functions evaluated on the learned trajectory distribution versus the expert trajectory distribution. Intuitively, if we treat the tracking error between two trajectories as a metric, if the cost function varies continuously (e.g. Lipschitz) with respect to this metric, our generalization bounds should imply some sort of *distributional distance* between the closed-loop learned and expert trajectory distributions. This is made explicit in the following result.

Theorem C.2 (Full version of Theorem 3.2, (10)) *Let us denote stacked trajectory vectors* $\vec{x}_{1:T} = \left[x[1]^\top \quad \cdots \quad x[T]^\top\right]^\top \in \mathbb{R}^{n_x T}$, *and denote* $\vec{x}^\star_{1:T} \sim \mathcal{P}^\star_{1:T}$ *and* $\hat{\vec{x}}_{1:T} \sim \hat{\mathcal{P}}_{1:T}$ *as the distributions of closed-loop expert and learned trajectories generated by* $K^{(H+1)}$ *and* $\hat{K}$, *respectively. Let* h *be any cost function that is* L*-Lipschitz with respect to the metric between trajectories* $d(\vec{x}_{1:T}, \vec{y}_{1:T}) = \max_{1\le t\le T}\|x[t]-y[t]\|$, *i.e.,*

$$|h(x) - h(y)| \le Ld(x,y), \ \forall x, y \in \mathcal{X}$$

Assume the sample requirements of Theorem C.1 are satisfied for given δ. *Then with probability greater than* $1-\delta$, *the following bound holds on the gap between the expected costs of expert and learned trajectories:*

$$\left|\mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right] - \mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})]\right| \le 2\sqrt{3}L\mathcal{J}(A_{\mathsf{cl}})\|B\|\sqrt{1+\log(T)}\sqrt{\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})}$$

where $A_{\mathsf{cl}} := A + BK^{(H+1)}$.

Remark C.1 *We note that since* $\frac{1}{T}\sum_{t=1}^T\|x[t]-y[t]\| \le \max_{1\le t\le T}\|x[t]-y[t]\|$, *the above result holds with minimal modification for* $d(\vec{x}_{1:T}, \vec{y}_{1:T}) = \frac{1}{T}\sum_{t=1}^T\|x[t]-y[t]\|$.

Before proceeding with the proof of Theorem C.2, the following are valid examples of $h(\cdot)$:

- $h(\vec{x}_{1:T}) = \max_{1\le t\le T}\left\|Q^{1/2}x[t]\right\|$,

- $h(\vec{x}_{1:T}) = \max_{1 \leq t \leq T} \|x[t] - x_{\text{goal}}[t]\| + \lambda \|Rx[t]\|$.

Proof We appeal to an application of Kantorovich-Rubinstein duality to establish general in-expectation guarantees. Let us define the metric

$$d(\vec{x}_{1:T}, \vec{y}_{1:T}) = \max_{1 \leq t \leq T} \|x[t] - y[t]\|.$$

Define $\Gamma(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T})$ as the set of all couplings between the distributions $\mathcal{P}^\star_{1:T}$ and $\hat{\mathcal{P}}_{1:T}$. In particular, trajectories $\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}$ following the same instances of $w[t]$ and $z[t]$ considered in (27) is one such coupling. The Wasserstein 1-distance between $\mathcal{P}^\star_{1:T}$ and $\hat{\mathcal{P}}_{1:T}$ is defined as

$$\begin{aligned}
\mathcal{W}_1\left(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T}\right) &= \inf_{\gamma \in \Gamma(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T})} \mathbb{E}_{(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}) \sim \gamma}\left[d\left(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}\right)\right] \\
&= \inf_{\gamma \in \Gamma(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T})} \mathbb{E}_{(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}) \sim \gamma}\left[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|\right].
\end{aligned}$$

In particular, defining the coupling (27) as $\overline{\gamma}$, we immediately have

$$\mathcal{W}_1\left(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T}\right) \leq \mathbb{E}_{(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}) \sim \overline{\gamma}}\left[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|\right].$$

Kantorovich-Rubinstein duality (cf. Villani (2021)) provides the following dual characterization of the Wasserstein distance:

$$\frac{1}{L} \sup_{\|h\|_{\text{Lip}} \leq L} \mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right] - \mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})] = \mathcal{W}_1\left(\mathcal{P}^\star_{1:T}, \hat{\mathcal{P}}_{1:T}\right),$$

where

$$\|h\|_{\text{Lip}} = \sup_{x, y \in \mathcal{X}} \frac{|h(x) - h(y)|}{d(x, y)}.$$

Therefore, given any cost function h that is L-Lipschitz continuous with respect to $d(x, y)$, the gap between the expected closed-loop costs of the learned and expert controllers, we can chain inequalities to yield

$$\mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right] - \mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})] \leq L\mathbb{E}_{(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}) \sim \overline{\gamma}}\left[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|\right].$$

It remains to provide a bound on $\mathbb{E}_{(\vec{x}^\star_{1:T}, \hat{\vec{x}}_{1:T}) \sim \overline{\gamma}}[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|]$. Recalling the bound (31), this reduces to bounding the imitation error along expert trajectories $x_\star[t]$. In order to convert a bound on $\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|$ to $\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|$, we apply Jensen's inequality to yield

$$\mathbb{E}\left[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|\right]^2 \leq \mathbb{E}\left[\max_{1 \leq t \leq T} \|x_\star[t] - \hat{x}[t]\|^2\right]$$

The rest follows from an application of the maximal inequality.

Proposition C.2 *Given a test controller K, we have the following maximal-type inequality between the excess maximal risk and the excess average risk over the expert target task data:*

$$\mathbb{E}\left[\max_{0 \leq t \leq T-1} \left\|(K - K^{(H+1)})x_\star[t]\right\|^2\right] \leq 3(1 + \log(T))\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \left\|(K - K^{(H+1)})x_\star[t]\right\|^2\right].$$

Proof We recall that the population distribution of $x_\star[t]$ is $\mathcal{N}(0, \Sigma_x^{(H+1)})$. Recalling the definitions in (32), we can re-write

$$\left\|(K - K^{(H+1)})x_\star[t]\right\|^2 = z_t^\top R^\top R z_t,$$

where $z_t \sim \mathcal{N}(0, I)$. Therefore, what we need to show is a maximal inequality on

$$\mathbb{E}\left[\max_{0 \le t \le T-1} z_t^\top R^\top R z_t\right].$$

Toward establishing such an inequality, we recall the following bound on the moment-generating function of $z_t^\top R^\top R z_t$ from (14) (substituting λ for λ^2):

$$\mathbb{E}\left[\exp\left(\lambda \left\|Rz\right\|^2\right)\right] \le \exp\left(3\lambda \left\|R\right\|_F^2\right), \quad |\lambda| \le \frac{1}{3\left\|R\right\|^2}.$$

Now, we leverage the bound on the MGF: for $\lambda > 0$, we write

$$\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\max_{t=0,\dots,T-1} z_t^\top R^\top R z_t\right]\right) &\le \mathbb{E}\left[\exp\left(\lambda \max_t z_t^\top R^\top R z_t\right)\right] && \text{Jensen's} \\
&= \mathbb{E}\left[\max_t \exp\left(\lambda z_t^\top R^\top R z_t\right)\right] && \text{monotonicity} \\
&\le \sum_{t=0}^{T-1} \mathbb{E}\left[\exp\left(\lambda z_t^\top R^\top R z_t\right)\right]
\end{aligned}$$

We note that in the last line, we are taking the expectation of each z_t over its population distribution, such that $z_t \sim \mathcal{N}(0, I)$. Thus,

$$\exp\left(\lambda \mathbb{E}\left[\max_{t=0,\dots,T-1} z_t^\top R^\top R z_t\right]\right) \le \sum_{t=0}^{T-1} \exp\left(3\lambda \left\|R\right\|_F^2\right) = T \exp\left(3\lambda \left\|R\right\|_F^2\right), \quad \lambda \le \frac{1}{3\left\|R\right\|^2}.$$

Taking the logarithm of both sides, re-arranging, and taking the infimum over λ, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{t=0,\dots,T-1} z_t^\top R^\top R z_t\right] &\le \inf_{\lambda \in \left[0, \frac{1}{3\|R\|^2}\right]} \frac{\log(T)}{\lambda} + 3\left\|R\right\|_F^2 \\
&= 3\left\|R\right\|^2 \log(T) + 3\left\|R\right\|_F^2 \\
&\le 3(1 + \log(T))\left\|R\right\|_F^2.
\end{aligned}$$

Substituting back $R = (K - K^{(H+1)})\left(\Sigma_x^{(H+1)}\right)^{1/2}$, and observing

$$\begin{aligned}
\left\|R\right\|_F^2 = \left\|(K - K^{(H+1)})\left(\Sigma_x^{(H+1)}\right)^{1/2}\right\|_F^2 &= \mathrm{Tr}\left((K - K^{(H+1)})\Sigma_x^{(H+1)}(K - K^{(H+1)})^\top\right) \\
&= \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|Kx_\star[t] - K^{(H+1)}x_\star[t]\right\|^2\right],
\end{aligned}$$

the proof of Proposition C.2 is complete. ∎

With the maximal inequality in hand, and defining $A_{\mathsf{cl}} := A + BK^{(H+1)}$, we may now bound:

$$
\begin{aligned}
&\mathbb{E}_{\hat{\mathcal{P}}_{1:T}}\left[h(\hat{\vec{x}}_{1:T})\right] - \mathbb{E}_{\mathcal{P}^\star_{1:T}}[h(\vec{x}^\star_{1:T})] \\
&\leq L\mathbb{E}_{(\vec{x}^\star_{1:T},\hat{\vec{x}}_{1:T})\sim\bar{\gamma}}\left[\max_{1\leq t\leq T} \|x_\star[t] - \hat{x}[t]\|\right] \\
&\leq L\sqrt{\mathbb{E}_{(\vec{x}^\star_{1:T},\hat{\vec{x}}_{1:T})\sim\bar{\gamma}}\left[\max_{1\leq t\leq T} \|x_\star[t] - \hat{x}[t]\|^2\right]} \\
&\leq L\sqrt{\mathbb{E}_{\mathcal{P}^\star_{0:T-1}}\left[\max_{0\leq t\leq T-1} 4\mathcal{J}(A_{\mathsf{cl}})^2 \|B\|^2 \left\|\left(\hat{K} - K^{(H+1)}\right)x_\star[t]\right\|^2\right]} \\
&\leq 2\sqrt{3}L\mathcal{J}(A_{\mathsf{cl}}) \|B\| \sqrt{1+\log(T)}\sqrt{\mathbb{E}_{\mathcal{P}^\star_{0:T-1}}\left[\frac{1}{T}\sum_{t=0}^{T-1} \left\|(K - K^{(H+1)})x_\star[t]\right\|^2\right]} \\
&\leq 2\sqrt{3}L\mathcal{J}(A_{\mathsf{cl}}) \|B\| \sqrt{1+\log(T)}\sqrt{\mathrm{ER}(\hat{\Phi}, \hat{F}^{(H+1)})}.
\end{aligned}
$$

■

C.1 Tightness of Dependence on Spectral Radius in Theorem 3.2

Consider the following scalar LTI system

$$x[t+1] = ax[t] + u[t] + w[t], \quad w[t] \sim \mathcal{N}(0,1),$$

where $a > 0$ without loss of generality. We are given the expert controller $u^\star[t] = k^\star x[t]$ that stabilizes the above system, such that we have

$$0 < a + k^\star < 1.$$

Let's say our learned controller $\hat{k}$ attains an $\varepsilon > 0$ error with respect to $k^\star$: $\hat{k} = k^\star + \varepsilon$, and $a + \hat{k} < 1$. We recall a couple of facts: the stationary distribution induced by $k^\star$ can be derived by solving the scalar Lyapunov equation

$$\sigma^2 = (a+k^\star)^2\sigma^2 + 1 \iff \sigma^2 = \frac{1}{1-(a+k^\star)^2}.$$

Therefore, akin to the setting considered in our paper, we assume the initial state distribution is the expert stationary distribution. We now study the tracking between the expert and learned states evolving in the coupled manner analogous to (27)

$$
\begin{aligned}
x_\star[t+1] &= (a+k^\star)x_\star[t] + w[t], \quad x_0 \sim \mathcal{N}\left(0, \frac{1}{1-(a+k^\star)^2}\right) \\
\hat{x}[t+1] &= (a+k^\star+\varepsilon)\hat{x}[t] + w[t], \quad x_0 \sim \mathcal{N}\left(0, \frac{1}{1-(a+k^\star)^2}\right).
\end{aligned}
$$

In particular, we show the following lower bound on the tracking error

Proposition C.3 *Given the proposed scalar system, the tracking error satisfies the following lower bound for sufficiently large T,*

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\le t\le T} |x_\star[t] - \hat{x}[t]|^2\right] \\
&\gtrsim \frac{1}{(1-(a+k^\star))^2} \underbrace{\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \left|(\hat{k}-k^\star)x_\star[t]\right|^2\right]}_{\text{excess risk of } \hat{k} \text{ on expert data}} =: \frac{1}{(1-(a+k^\star))^2}\mathrm{ER}(\hat{k}).
\end{aligned}$$

We note that instantiating the in-expectation upper bound using the maximal inequality in Proposition C.2 yields

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le t\le T} |x_\star[t] - \hat{x}[t]|^2\right] &\le 12\mathcal{J}(a+k^\star)^2 \|B\|^2 (1+\log(T))\mathrm{ER}(\hat{k}) \\
&= \frac{12}{(1-(a+k^\star))^2}(1+\log(T))\mathrm{ER}(\hat{k}).
\end{aligned}$$

Therefore, Proposition C.3 states that up to a log-factor in horizon length T, the polynomial dependence on the expert system's spectral radius $a+k^\star$ matches that in the upper bound.

Proof We can immediately compute the excess risk in closed form

$$\begin{aligned}
\mathrm{ER}(\hat{k}) := \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\left|(\hat{k}-k^\star)x_\star[t]\right|^2\right] &= (\hat{k}-k^\star)\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} x_\star[t]^2\right] \\
&= \frac{\varepsilon^2}{1-(a+k^\star)^2} \quad \text{by stationarity.}
\end{aligned}$$

Therefore, in addition to the spectral radius showing up in the excess risk, we need to show the tracking error accrues another factor of the spectral radius on top of that. We observe

$$\begin{aligned}
&\max_{1\le t\le T} |x_\star[t]-\hat{x}[t]|^2 \ge |x_\star[T]-\hat{x}[T]|^2 \\
\implies &\mathbb{E}\left[\max_{1\le t\le T} |x_\star[t]-\hat{x}[t]|^2\right] \ge \max_{t\le T}\mathbb{E}\left[|x_\star[T]-\hat{x}[T]|^2\right] \\
&\qquad = \max_{t\le T}\mathbb{E}\left[x_\star[T]^2\right] + \mathbb{E}\left[\hat{x}[T]^2\right] - 2\mathbb{E}[x_\star[T]\hat{x}[T]] \\
&\qquad = \max_{t\le T} \frac{1}{1-(a+k^\star)^2} + \frac{(a+\hat{k})^{2T}}{1-(a+k^\star)^2} + \sum_{t=0}^{T-1}(a+\hat{k})^{2(T-t)} \\
&\qquad\quad - \frac{2(a+k^\star)^T(a+\hat{k})^T}{1-(a+k^\star)^2} - 2\sum_{t=0}^{T-1}(a+k^\star)^{T-t}(a+\hat{k})^{T-t},
\end{aligned}$$

where the summations come from the fact that the initial condition is drawn from the *stationary distribution induced by the expert*, not the learned controller, in conjunction with the formula $x[t] = a^t x_0 + \sum_{k=0}^{t-1} a^{t-1-k} w[t]$ for system $x_{t+1} = ax[t] + w[t]$. Since in the limit we have

$$\lim_{T\to\infty} \mathbb{E}\left[|x_\star[T]-\hat{x}[T]|^2\right] = \frac{1}{1-(a+k^\star)^2} + \frac{1}{1-(a+\hat{k})^2} - \frac{2}{1-(a+k^\star)(a+\hat{k})},$$

we simply take T large enough such that

$$\mathbb{E}\Big[|x_\star[T] - \hat{x}[T]|^2\Big] \geq \frac{1}{2}\left(\frac{1}{1-(a+k^\star)^2} + \frac{1}{1-(a+\hat{k})^2} - \frac{2}{1-(a+k^\star)(a+\hat{k})}\right).$$

We now bound

$$\begin{aligned}
&\frac{1}{1-(a+k^\star)^2} + \frac{1}{1-(a+\hat{k})^2} - \frac{2}{1-(a+k^\star)(a+\hat{k})} \\
&= \sum_{t=0}^{\infty} (a+k^\star)^{2t} + (a+\hat{k})^{2t} - 2(a+k^\star)^t(a+\hat{k})^t \\
&= \sum_{t=0}^{\infty} ((a+\hat{k})^t - (a+k^\star)^t)^2 \\
&= \sum_{t=0}^{\infty} \Big((\hat{k}-k^\star)\Big((a+k^\star)^{t-1} + (a+k^\star)^{t-2}(a+\hat{k}) + \cdots + (a+k^\star)(a+\hat{k})^{t-2} + (a+\hat{k})^{t-1}\Big)\Big)^2 \\
&\geq \varepsilon^2 \sum_{k=0}^{\infty} \big(t(a+k^\star)^{t-1}\big)^2 \\
&= \varepsilon^2 \frac{1+(a+k^\star)^2}{(1-(a+k^\star)^2)^3},
\end{aligned}$$

where we used the algebraic identity

$$u^t - v^t = (u-v)(u^{t-1} + u^{t-2}v + \cdots + uv^{t-2} + v^{t-1}),$$

and the inequality (assuming $u \leq v$)

$$u^{t-1} + u^{t-2}v + \cdots + uv^{t-2} + v^{t-1} \geq nu^{t-1},$$

setting $u := a + k^\star$ and $v := a + \hat{k} = a + k^\star + \varepsilon$. Now recalling that the empirical risk by stationarity is given by

$$\mathrm{ER}(\hat{k}) = \frac{\varepsilon^2}{1-(a+k^\star)^2},$$

we can immediately infer

$$\begin{aligned}
\mathbb{E}\Big[|x_\star[T] - \hat{x}[T]|^2\Big] &\geq 0.5\varepsilon^2 \frac{1+(a+k^\star)^2}{(1-(a+k^\star)^2)^3} \\
&= 0.5\mathrm{ER}(\hat{k}) \frac{1+(a+k^\star)^2}{(1-(a+k^\star)^2)^2} \\
&> \frac{0.5}{(1-(a+k^\star)^2)^2} \mathrm{ER}(\hat{k}).
\end{aligned}$$

In short, we have established a lower bound on the expected imitation gap:

$$\begin{aligned}
\mathbb{E}\Big[\max_{0\leq t\leq T-1} |x_\star[t] - \hat{x}[t]|^2\Big] &\geq \max_{t \leq T-1} \mathbb{E}\Big[|x_\star[T-1] - \hat{x}[T-1]|^2\Big] \\
&\gtrsim \frac{1}{(1-(a+k^\star)^2)^2} \mathrm{ER}(\hat{k}).
\end{aligned}$$

We now claim that $\frac{1}{(1-(a+k^\star)^2)^2}$ matches the dependence on the spectral radius in the upper bound. We recall the upper bound on the tracking error:

$$\mathbb{E}\left[\max_{0\le t\le T-1}|x_\star[t]-\hat{x}[t]|^2\right]\le\frac{12}{(1-(a+k^\star))^2}(1+\log(T))\mathrm{ER}(\hat{k}).$$

Comparing $\frac{1}{(1-(a+k^\star)^2)^2}$ to $\frac{1}{(1-(a+k^\star))^2}$, we get

$$\frac{\frac{1}{(1-(a+k^\star))^2}}{\frac{1}{(1-(a+k^\star)^2)^2}}=\frac{(1-(a+k^\star))^2(1+(a+k^\star))^2}{(1-(a+k^\star))^2}=(1+(a+k^\star))^2\ge 1,$$

which essentially states the dependence on the spectral radius in the lower and upper bounds match up to a constant factor:

$$\frac{0.5}{(1-(a+k^\star))^2}\mathrm{ER}(\hat{k})\le\mathbb{E}\left[\max_{0\le t\le T-1}|x_\star[t]-\hat{x}[t]|^2\right]\le\frac{12}{(1-(a+k^\star))^2}(1+\log(T))\mathrm{ER}(\hat{k}),$$

which completes the result. ■