

# SAMPLE-EFFICIENT LINEAR REPRESENTATION LEARNING FROM NON-IID NON-ISOTROPIC DATA

**Thomas T.C.K. Zhang**  
University of Pennsylvania  
Philadelphia, PA  
ttz2@seas.upenn.edu

**Leonardo F. Toso**  
Columbia University  
New York, NY  
lt2879@columbia.edu

**James Anderson**  
Columbia University  
New York, NY  
james.anderson@columbia.edu

**Nikolai Matni**  
University of Pennsylvania  
Philadelphia, PA  
nmatni@seas.upenn.edu

## ABSTRACT

A powerful concept behind much of the recent progress in machine learning is the extraction of common features across data from heterogeneous sources or tasks. Intuitively, using all of one’s data to learn a common representation function benefits both computational effort and statistical generalization by leaving a smaller number of parameters to fine-tune on a given task. Toward theoretically grounding these merits, we propose a general setting of recovering linear operators  $M$  from noisy vector measurements  $y = Mx + w$ , where the covariates  $x$  may be both non-i.i.d. and non-isotropic. We demonstrate that existing isotropy-agnostic meta-learning approaches incur biases on the representation update, which causes the scaling of the noise terms to lose favorable dependence on the number of source tasks. This in turn can cause the sample complexity of representation learning to be bottlenecked by the single-task data size. We introduce an adaptation, `De-bias & Feature-Whiten (DFW)`, of the popular alternating minimization-descent (AMD) scheme proposed in Collins et al., (2021) (Collins et al., 2021), and establish linear convergence to the optimal representation with noise level scaling down with the *total* source data size. This leads to generalization bounds on the same order as an oracle empirical risk minimizer. We verify the vital importance of `DFW` on various numerical simulations. In particular, we show that vanilla alternating-minimization descent fails catastrophically even for iid, but mildly non-isotropic data. Our analysis unifies and generalizes prior work, and provides a flexible framework for a wider range of applications, such as in controls and dynamical systems.

## 1 INTRODUCTION

A unifying paradigm belying recent exciting progress in machine learning is learning a common feature space or *representation* for downstream tasks from heterogeneous sources. This forms the core of fields such as meta-learning, transfer learning, and federated learning. A shared theme across these fields is the scarcity of data for a specific task out of many, such that designing individual models for each task is both computationally and statistically inefficient, impractical, or impossible. Under the assumption that these tasks are similar in some way, a natural alternative approach is to use data across many tasks to learn a common component, such that fine-tuning to a given task involves fitting a much smaller model that acts on the common component. Over the last few years, significant attention has been given to framing this problem setting theoretically, providing provable benefits of learning over multiple tasks in the context of linear regression (Bullins et al., 2019; Du et al., 2020; Tripuraneni et al., 2021; Collins et al., 2021; Thekumparampil et al., 2021; Saunshi et al., 2021) and in identification/control of linear dynamical systems (Modi et al., 2021; Chen et al., 2023; Zhang et al., 2023). These works study the problem of *linear representation learning*, where the data for each task is generated noisily from an unknown shared latent subspace, and the goal is to efficiently recover a representation of the latent space  $\hat{\Phi}$  from data across different task distributions.

For example, in the linear regression setting, one may have data of the form

$$y_i^{(t)} = \theta^{(t)\top} \Phi x_i^{(t)} + \text{noise}, \quad y_i^{(t)} \in \mathbb{R}, x_i^{(t)} \in \mathbb{R}^{d_x}, \Phi \in \mathbb{R}^{r \times d_x},$$

with  $i = 1, \dots, N$  iid data points from  $t = 1, \dots, T$  task distributions. Since the representation  $\Phi$  is shared across all tasks, one may expect the generalization error of an approximate representation  $\hat{\Phi}$  fit on  $TN$  data points to scale as  $\frac{d_x r}{TN}$ , where  $d_x r$  is the number of parameters determining the representation. This is indeed the flavor of statistical guarantees from prior work (Du et al., 2020; Tripuraneni et al., 2021; Thekumparampil et al., 2021; Zhang et al., 2023), which concretely demonstrates the benefit of using data across different tasks.

However, existing work, especially beyond the scalar measurement setting, is limited in one or more important components of their analysis. For example, it is common to assume that the covariates  $x_i^{(t)}$  are isotropic across all tasks. Furthermore, statistical analyses often assume access to an empirical risk minimizer, even though the linear representation learning problem is non-convex and ill-posed (Maurer et al., 2016; Du et al., 2020; Zhang et al., 2023). Our paper addresses these problems under a unified framework of *linear operator recovery*, i.e. recovering linear operators  $M \in \mathbb{R}^{d_y \times d_x}$  from (noisy) vector measurements  $y = Mx + w$ , where the covariates  $x$  may not be independent or isotropic. This setting subsumes the scalar measurement setting, and encompasses many fundamental control and dynamical systems problems, such as linear system identification and imitation learning. In particular, the data in these settings are incompatible with the common distributional assumptions (e.g., independence, isotropy) made in prior work.

**Contributions:** Toward this end, our main contributions are as follows:

- We demonstrate that naive implementation of local methods for linear representation learning fail catastrophically even when the data is iid but mildly non-isotropic. We identify the source of the failure as interaction between terms incurring biases in the representation gradient, which do not scale down with the number of tasks.
- We address these issues by introducing two practical algorithmic adjustments, De-bias & Feature-Whiten (DFW), which provably mitigate the identified issues. We then show that DFW is necessary for gradient-based methods to benefit from the total size of the source dataset.
- We numerically show our theoretical guarantees are predictive of the efficacy of our proposed algorithm, and of the key importance of individual aspects of our algorithmic framework.

Our main result can be summarized by the following descent guarantee for our proposed algorithm.

**Theorem 1.1 (main result, informal)** *Let  $\hat{\Phi}$  be the current estimate of the representation, and  $\Phi_*$  the optimal representation. Running one iteration of DFW yields the following improvement*

$$\text{dist}(\hat{\Phi}_+, \Phi_*) \leq \rho \cdot \text{dist}(\hat{\Phi}, \Phi_*) + \frac{C}{\sqrt{\# \text{ tasks} \times \# \text{ data per task}}}, \quad \rho \in (0, 1), C > 0.$$

Critically, the second term of the right hand side scales jointly in the number of tasks and datapoints per task, whereas naively implementing other methods may be bottlenecked by a term that scales solely with the amount of data for a single task, which leads to suboptimal sample-efficiency.

## 1.1 RELATED WORK

**Multi-task linear regression:** Directly related to our work are results demonstrating the benefits of multi-task learning for linear regression (Maurer et al., 2016; Bullins et al., 2019; Du et al., 2020; Tripuraneni et al., 2021; Collins et al., 2021; Thekumparampil et al., 2021), under the assumption of a shared but unknown linear feature representation. In particular, our proposed algorithm is adapted from the alternating optimization scheme independently in Nayer and Vaswani (2022) and Collins et al. (2021), and extends these results to the vector measurement setting and introduces algorithmic modifications to extend its applicability to non-iid and non-isotropic covariates. We also highlight that in the isotropic linear regression setting, Thekumparampil et al. (2021) provide an alternating minimization scheme that results in near minimax-optimal representation learning. However, the representation update step simultaneously accesses data across tasks, which we avoid in this work due to motivating applications, e.g. distributed learning, that impose locality or data privacy constraints.

**Meta/multi-task RL:** There is a wealth of literature in reinforcement learning that seeks empirically to solve different tasks with shared parameters (Teh et al., 2017; Hessel et al., 2018; Singh et al., 2020; Deisenroth et al., 2014). In parallel, there is a body of theoretical work which studies the sample efficiency of representation learning for RL (Lu et al., 2021; Cheng et al., 2022; Maurer et al., 2015). This line of work considers MDP settings, and thus the specific results are often stated with incompatible assumptions (such as bounded states/cost functions and discrete action spaces), and are suboptimal when instantiated in our setting.

**System identification and control:** Multi-task learning has gained recent attention in controls, e.g. for adaptive control over similar dynamics (Harrison et al., 2018; Richards et al., 2021; Shi et al., 2021; Muthirayan et al., 2022), imitation learning for linear systems (Zhang et al., 2023; Guo et al., 2023), and notably linear system identification (Li et al., 2022; Wang et al., 2022; Xin et al., 2023; Modi et al., 2021; Faradonbeh and Modi, 2022; Chen et al., 2023). In many of these works (Li et al., 2022; Wang et al., 2022; Xin et al., 2023), task similarity is quantified by a generic norm closeness of the dynamics matrices, and thus the benefit of multiple tasks extends only to a radius around optimality. Under the existence of a shared representation, our work provides an efficient algorithm and statistical analysis to establish convergence to per-task optimality.

## 2 PROBLEM FORMULATION

**Notation:** the Euclidean norm of a vector  $x$  is denoted  $\|x\|$ . The spectral and Frobenius norms of a matrix  $A$  are denoted  $\|A\|$  and  $\|A\|_F$ , respectively. For symmetric matrices  $A, B$ ,  $A \preceq B$  denotes  $B - A$  is positive semidefinite. The largest/smallest singular and eigenvalues of a matrix  $A$  are denoted  $\sigma_{\max}(A)$ ,  $\sigma_{\min}(A)$ , and  $\lambda_{\max}(A)$ ,  $\lambda_{\min}(A)$ , respectively. The condition number of a matrix  $A$  is denoted  $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$ . Define the indexing shorthand  $[n] := \{1, \dots, n\}$ . We use big-O notation  $\mathcal{O}(\cdot)$ ,  $\Theta(\cdot)$ ,  $\Omega(\cdot)$  to omit universal numerical factors, and  $\tilde{\mathcal{O}}(\cdot)$ ,  $\tilde{\Theta}(\cdot)$ ,  $\tilde{\Omega}(\cdot)$  to additionally omit polylog factors in the argument.

**Regression Model.** Let a covariate sequence (also denoted a *trajectory*) be an indexed set  $\{x_i\}_{i \geq 1} \subset \mathbb{R}^{d_x}$ . We denote a distribution  $\mathbb{P}_x$  over covariate sequences, which we assume to have bounded second moments for all  $i \geq 1$ , i.e.  $\mathbb{E}[x_i x_i^\top]$  is finite for all  $i \geq 1$ . Defining the filtration  $\{\mathcal{F}_i\}_{i \geq 0}$  where  $\mathcal{F}_i := \sigma(\{x_k\}_{k=1}^{i+1}, \{w_k\}_{k=1}^i)$  is the  $\sigma$ -algebra generated by the covariates up to  $i + 1$  and noise up to  $i$ , we assume that  $\{w_i\}_{i \geq 1}$  is a  $\sigma_w^2$ -subgaussian martingale difference sequence (MDS):

$$\mathbb{E}[v^\top w_i \mid \mathcal{F}_{i-1}] = 0, \quad \mathbb{E}[\exp(\lambda v^\top w_i) \mid \mathcal{F}_{i-1}] \leq \exp\left(\lambda^2 \|v\|^2 \sigma_w^2\right) \text{ a.s. } \forall \lambda \in \mathbb{R}, v \in \mathbb{R}^{d_y}.$$

Assuming a ground truth operator  $M_\star \in \mathbb{R}^{d_y \times d_x}$ , our observation model is given by

$$y_i = M_\star x_i + w_i, \quad i \geq 1,$$

for  $y_i$  the *labels*, and  $w_i$  the *label noise*. We further define  $\Sigma_{x,N} := \frac{1}{N} \sum_{i=1}^N \mathbb{E}[x_i x_i^\top]$ . When the marginal distributions of  $x_i$ ,  $i \geq 1$  are identical, we denote  $\Sigma_x \equiv \Sigma_{x,N}$ .

**Multi-Task Operator Recovery.** We consider the following instantiation of the above linear operator regression model over multiple tasks. In particular, we consider heterogeneous data  $\{(x_i^{(t)}, y_i^{(t)})\}_{i=1, t=1}^{N,T}$ , consisting of trajectories of length  $N$ , generated independently across  $t = 1, \dots, T$  task distributions. For notational convenience, we assume that the length of trajectories  $N$  is the same across training tasks. For each task  $t$ , the observation model is

$$y_i^{(t)} = M_\star^{(t)} x_i^{(t)} + w_i^{(t)}, \quad (1)$$

where  $M_\star^{(t)} = F_\star^{(t)} \Phi_\star$  admits a decomposition into a ground-truth representation  $\Phi_\star \in \mathbb{R}^{r \times d_x}$  common across all tasks  $t \in [T]$  and a task-specific weight matrix  $F_\star^{(t)} \in \mathbb{R}^{d_y \times r}$ ,  $r \leq d_x$ . We denote the joint distribution over covariates and observations  $\{x_i^{(t)}, y_i^{(t)}\}_{i \geq 1}$  by  $\mathbb{P}_{x,y}^{(t)}$ . We assume that the representation  $\Phi_\star$  is normalized to have orthonormal rows to prevent boundedness issues, since  $F_\star^{(t)'} = F_\star^{(t)} Q^{-1}$ ,  $\Phi_\star' = Q \Phi_\star$  are valid decompositions for any invertible  $Q \in \mathbb{R}^{r \times r}$ . To measure closeness of an approximate representation  $\hat{\Phi}$  to optimality, we define a subspace metric.

**Definition 2.1 (Subspace Distance (Stewart and Sun, 1990; Collins et al., 2021))** Let  $\Phi, \Phi_\star \in \mathbb{R}^{r \times d_x}$  be matrices whose rows are orthonormal. Furthermore, let  $\Phi_{\star, \perp} \in \mathbb{R}^{(d_x - r) \times d_x}$  be a matrix such that  $[\Phi_\star^\top \ \Phi_{\star, \perp}^\top]$  is an orthogonal matrix. Define the distance between the subspaces spanned by the rows of  $\Phi$  and  $\Phi_\star$  by

$$\text{dist}(\Phi, \Phi_\star) := \|\Phi \Phi_{\star, \perp}^\top\|_2 \quad (2)$$

In particular, the subspace distance quantitatively captures the alignment between two subspaces, interpolating smoothly between 0 (occurring iff  $\text{span}(\Phi_\star) = \text{span}(\Phi)$ ) and 1 (occurring iff  $\text{span}(\Phi_\star) \perp \text{span}(\Phi)$ ). We define the task-specific stacked vector notation by capital letters, e.g.,

$$X^{(t)} = \begin{bmatrix} x_1^{(t)} & \cdots & x_i^{(t)} & \cdots & x_N^{(t)} \end{bmatrix}^\top \in \mathbb{R}^{N \times d_x}.$$

The goal of multi-task operator recovery is to estimate  $\{F_\star^{(t)}\}_{t=1}^T$  and  $\Phi_\star$  from data collected across multiple tasks  $\{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^N, t = 1, \dots, T$ . Some prior works (Maurer et al., 2016; Du et al., 2020; Zhang et al., 2023) assume access to an empirical risk minimization oracle, i.e. access to

$$\{\hat{F}^{(t)}\}_{t=1}^T, \hat{\Phi} \in \underset{\{F^{(t)}, \Phi\}}{\text{argmin}} \sum_{t=1}^T \sum_{i=1}^N \left\| y_i^{(t)} - F^{(t)} \Phi x_i^{(t)} \right\|^2,$$

focusing on the statistical generalization properties of an ERM solution. However, the above optimization is non-convex even in the linear setting, and thus it is imperative to design and analyze efficient algorithms for recovering optimal matrices  $\{F_\star^{(t)}\}_{t=1}^T$  and  $\Phi_\star$ . To address this problem in the linear regression setting, various works, e.g. FedRep (Collins et al., 2021), AltGD-Min (Nayer and Vaswani, 2022), propose an alternating minimization-descent scheme, where on a fresh data batch, the weights  $\{\hat{F}^{(t)}\}$  are computed on local data via least-squares, and subsequently an estimate of the representation gradient is then computed with respect to local data and aggregated across tasks to perform gradient descent. This algorithmic framework is intuitive, and thus forms a reasonable starting point toward a provably sample-efficient algorithm in our setting.

### 3 SAMPLE-EFFICIENT LINEAR REPRESENTATION LEARNING

We begin by describing the vanilla alternating minimization-descent scheme proposed in Collins et al. (2021). We show that in our setting with label noise and non-isotropy, interaction terms arise in the representation gradient, which cause biases to form that do not scale down with the number of tasks  $T$ . In §3.2, we propose alterations to the scheme to remove these biases, which we then show in §3.3 lead to fast convergence rates that allow us to recover near-oracle ERM generalization bounds.

#### 3.1 PERILS OF (VANILLA) GRADIENT DESCENT ON THE REPRESENTATION

We begin with a summary of the main components of an alternating minimization-descent method analogous to FedRep (Collins et al., 2021) and AltGD-Min (Nayer and Vaswani, 2022). During each optimization round, a new data batch is sampled for each task:  $\{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^N, t \in [T]$ . We then compute task-specific weights  $\hat{F}^{(t)}$  on the corresponding dataset, keeping the current representation estimate  $\hat{\Phi}$  fixed. For example,  $\hat{F}^{(t)}$  may be the least-squares weights conditioned on  $\hat{\Phi}$  (Collins et al., 2021). Define  $z_i^{(t)} := \hat{\Phi} x_i^{(t)}$ , and the empirical covariance matrices  $\hat{\Sigma}_z^{(t)} := \frac{1}{N} X^{(t)\top} X^{(t)}, \hat{\Sigma}_z^{(t)} := \frac{1}{N} Z^{(t)\top} Z^{(t)}$ . The least squares solution  $\hat{F}^{(t)}$  is given by the convex quadratic minimization  $\hat{F}^{(t)} = \underset{F}{\text{argmin}} \sum_{i=1}^N \|y_i^{(t)} - F z_i^{(t)}\|^2$ . For each task, we then fix the weight matrix  $\hat{F}^{(t)}$  and perform a descent step with respect to the representation conditioned on the local data. The resulting representations are averaged across tasks to form the new representation. When the descent direction is the gradient, the update rule is given by

$$\bar{\Phi}_+^{(t)} = \hat{\Phi} - \frac{\eta}{2N} \nabla_{\Phi} \left\| \sum_{i=1}^N y_i^{(t)} - \hat{F}^{(t)} \hat{\Phi} x_i^{(t)} \right\|^2, \quad \bar{\Phi}_+ = \frac{1}{T} \sum_{t=1}^T \bar{\Phi}_+^{(t)} \quad (3)$$

where  $\eta > 0$  is a given step size. We normalize  $\bar{\Phi}_+$  to have orthonormal rows, e.g. by (thin/reduced) QR decomposition (Trefethen and Bau, 2022), to produce the final output  $\hat{\Phi}_+$ , i.e.  $\bar{\Phi}_+ = R\hat{\Phi}_+$ ,  $R \in \mathbb{R}^{r \times r}$ , leading to

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \left( \hat{F}^{(t)} \hat{\Phi} - F_\star^{(t)} \Phi_\star \right) \hat{\Sigma}_x^{(t)} - \frac{\eta}{NT} \sum_{t=1}^T \hat{F}^{(t)\top} W^{(t)\top} X^{(t)}. \quad (4)$$

As in Collins et al. (2021), we right-multiply both sides of (4) by  $\Phi_{\star,\perp}^\top$ , recalling  $\|\Phi\Phi_{\star,\perp}^\top\|_2 =: \text{dist}(\Phi, \Phi_\star)$ . Crucially, Collins et al. (2021) assume  $x_i^{(t)}$  has mean 0 and identity covariance, and  $w_i^{(t)} \equiv 0$  across  $i, t$ . Therefore, the label noise terms  $\hat{F}^{(t)\top} W^{(t)\top} X^{(t)}$  disappear, and the sample covariance for each task  $\hat{\Sigma}_x^{(t)}$  concentrates to identity, such that  $\Phi_\star \hat{\Sigma}_x^{(t)} \Phi_{\star,\perp}^\top \approx 0$ . Under appropriate choice of  $\eta$  and bounding the effect of the orthonormalization factor  $R$ , linear convergence to the optimal representation can be established. However, two issues arise when label noise  $w_i^{(t)}$  is introduced and when  $x_i^{(t)}$  has non-identity covariance.

1. When *label noise*  $w_i^{(t)}$  is present, since  $\hat{F}^{(t)}$  is computed on  $Y^{(t)}, X^{(t)}$ , the gradient noise term is generally biased:  $\frac{1}{NT} \mathbb{E}[\hat{F}^{(t)} W^{(t)\top} X^{(t)}] \neq 0$ . Even in the simple case that all task distributions  $\mathbb{P}_{x,y}^{(t)}$  are identical,  $\frac{\eta}{NT} \sum_{t=1}^T \hat{F}^{(t)\top} W^{(t)\top} X^{(t)}$  concentrates to its bias, and thus for large  $T$  the size of noise term is bottlenecked at  $\frac{\eta}{NT} \mathbb{E} \left[ \left\| \hat{F}^{(t)\top} W^{(t)\top} X^{(t)} \right\| \right]$ . This critically causes the noise term to lose scaling in the *number of tasks*  $T$ , even when the tasks are identical.
2. When  $x_i^{(t)}$  has *non-identity covariance*, the decomposition into a contraction and covariance concentration term no longer holds, since generally  $\Phi_\star \mathbb{E}[\hat{\Sigma}_x^{(t)}] \Phi_{\star,\perp}^\top \neq 0$ . This causes a term whose norm otherwise concentrates around 0 in the isotropic case to scale with  $\lambda_{\max}(\hat{\Sigma}_x^{(t)}) - \lambda_{\min}(\hat{\Sigma}_x^{(t)})$  in the worst case.

This motivates modifying the representation update beyond following the vanilla stochastic gradient.

### 3.2 A TASK-EFFICIENT ALGORITHM: DE-BIAS & FEATURE-WHITEN

In the previous section, we identified two fundamental issues: 1. the bias introduced by computing the least squares weights and representation update on the same data batch, and 2. the nuisance term introduced by non-identity second moments of the covariates  $x_i^{(t)}$ . Toward addressing the first issue, we introduce a “de-biasing” step, where each agent computes the least squares weights  $\hat{F}^{(t)}$  and the representation update on *independent* batches of data, e.g. disjoint subsets of trajectories. To address the second issue, we introduce a “feature-whitening” adaptation (LeCun et al., 2002), where the gradient estimate sent by each agent is pre-conditioned by its inverse sample covariance matrix. Combining these two adaptations, the representation update becomes

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \left( \hat{F}^{(t)} \hat{\Phi} - F_\star^{(t)} \Phi_\star \right) - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W^{(t)\top} X^{(t)} \left( \hat{\Sigma}_x^{(t)} \right)^{-1}, \quad (5)$$

where we assume  $\{\hat{F}^{(t)}\}$  are computed on independent data using the aforementioned batching strategy. When  $x_i^{(t)}, w_i^{(t)}, t = 1, \dots, T$ , are all mutually independent, then the first two terms of the update form the contraction, and the last term is an average of *zero-mean* least-squares-error-like terms over tasks, which can be studied using standard tools (Abbasi-Yadkori et al., 2011; Abbasi-Yadkori, 2013). This culminates in convergence rates that scale favorably with the number of tasks (§3.3). To operationalize our proposed adaptations, let  $D^{(t)} = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^N, t \in [T]$ , be a dataset available to each agent. For the weights de-biasing step, we sub-sample trajectories  $\mathcal{N}_1 \subset [N]$ . For each agent, we compute least-squares weights from  $\mathcal{N}_1$ . We then sub-sample trajectories  $\mathcal{N}_2 \subset [N] \setminus \mathcal{N}_1$ , and compute the task-conditioned representation gradients from  $\mathcal{N}_2$ .

$$\hat{F}^{(t)} = \underset{F}{\operatorname{argmin}} \sum_{i \in \mathcal{N}_1} \left\| y_i^{(t)} - F z_i^{(t)} \right\|^2, \quad \hat{\mathcal{G}}_{\mathcal{N}_2}^{(t)} = \nabla_{\Phi} \frac{1}{2} \left\| \sum_{i \in \mathcal{N}_2} y_i^{(t)} - \hat{F}^{(t)} \hat{\Phi} x_i^{(t)} \right\|^2. \quad (6)$$

**Algorithm 1** De-biased & Feature-whitened (DFW) Alt. Minimization-Descent

---

```

1: Input: step sizes  $\{\eta_k\}_{k \geq 1}$ , batch sizes  $\{N_k\}_{k \geq 1}$ , initial estimate  $\hat{\Phi}_0$ .
2: for  $k = 1, \dots, K$  do
3:   for  $t \in [T]$  (in parallel) do
4:     Obtain samples  $\{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_k}$ .
5:     Partition trajectories  $[N_k] = \mathcal{N}_{k,1} \sqcup \mathcal{N}_{k,2}$ .
6:     Compute  $\hat{F}_k^{(t)}$ , e.g. via least squares on  $\mathcal{N}_{k,1}$  (6).
7:     Compute task-conditioned representation gradient  $\hat{\mathcal{G}}_{\mathcal{N}_{k,2}}^{(t)}$  on  $\mathcal{N}_{k,2}$  (6).
8:     Compute task-conditioned representation update  $\bar{\Phi}_k^{(t)}$  (7).
9:   end for
10:   $\hat{\Phi}_{k, \_} \leftarrow \text{thin\_QR} \left( \frac{1}{T} \sum_{t=1}^T \bar{\Phi}_k^{(t)} \right)$ .
11: end for
12: return Representation estimate  $\hat{\Phi}_K$ .

```

---

Lastly, each agent updates its local representation via a feature-whitened gradient step to yield  $\bar{\Phi}_+^{(t)}$ . The global representation update is computed by averaging the updated task-conditioned representations  $\bar{\Phi}_+^{(t)}$  and performing orthonormalization:

$$\bar{\Phi}_+^{(t)} := \hat{\Phi} - \eta \hat{\mathcal{G}}_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1}, \quad R \hat{\Phi}_+ = \frac{1}{T} \sum_{t=1}^T \bar{\Phi}_+^{(t)}, \quad (7)$$

We summarize the full algorithm in Algorithm 1. The above de-biasing and feature whitening steps ensure that the expectation of the representation update is a contraction (with high probability):

$$\mathbb{E} \left[ \text{dist}(\hat{\Phi}_+, \Phi_*) \right] = \mathbb{E} \left[ \left\| R^{-1} \left( I_{d_x} - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \right\| \right] \text{dist}(\hat{\Phi}, \Phi_*), \quad (8)$$

where the task and trajectory-wise independence ensures that the variance of the gradient scales inversely in  $NT$ .

### 3.3 ALGORITHM GUARANTEES

We present our main result in the form of convergence guarantees for Algorithm 1. We begin by defining a standard measure of dependency along covariate sequences via  $\beta$ -mixing.

**Definition 3.1 ( $\beta$ -mixing)** Let  $\{x_i\}_{i \geq 1}$  be a  $\mathbb{R}^d$ -valued discrete-time stochastic process adapted to filtration  $\{\mathcal{F}_i\}_{i=1}^\infty$ . We denote the stationary distribution  $\nu_\infty$ . We define the  $\beta$ -mixing coefficient

$$\beta(k) := \sup_{i \geq 1} \mathbb{E}_{\{x_\ell\}_{\ell=1}^i} \left[ \left\| \mathbb{P}_{x_{i+k}}(\cdot | \mathcal{F}_i) - \nu_\infty \right\|_{\text{TV}} \right], \quad (9)$$

where  $\|\cdot\|_{\text{TV}}$  denotes the total variation distance between probability measures.

Intuitively, the  $\beta$ -mixing coefficient measures how quickly on average a process mixes to its stationary distribution along any sample path. To instantiate our bounds, we make the following assumptions on the covariates.

**Assumption 3.1 (Subgaussian covariates, geometric mixing)** We assume the marginal distributions of  $x_i^{(t)}$  to be identical with zero-mean, covariance  $\Sigma_x^{(t)}$ , and to  $\gamma^2$ -subgaussian across all  $i, t$ :

$$\mathbb{E}[x_i^{(t)}] = 0, \quad \mathbb{E} \left[ \exp \left( \lambda v^\top x_i^{(t)} \right) \right] \leq \exp \left( \lambda^2 \|v\|^2 \gamma^2 \right) \quad \text{a.s. } \forall \lambda \in \mathbb{R}, v \in \mathbb{R}^{d_x}.$$

Furthermore, we assume the process  $\{x_i^{(t)}\}_{i \geq 1}$  is geometrically  $\beta$ -mixing with  $\beta^{(t)}(k) \leq \Gamma^{(t)} \mu^{(t)k}$ ,  $\Gamma^{(t)} > 0$ ,  $\mu^{(t)} \in [0, 1)$ , for each task  $t \in [T]$ . Lastly, we define  $\tau_{\text{mix}}^{(t)} := \left( \frac{\log(\Gamma^{(t)} N / \delta)}{\log(1/\mu^{(t)})} \vee 1 \right)$ .

Notably, these assumptions subsume fundamental problems in learning over (stable) dynamical systems, in particular linear system identification and imitation learning, where non-iid and non-isotropy across tasks are unavoidable. We discuss these applications in depth in Appendix B. As in prior work, our final convergence rates depend on a notion of task diversity.

**Definition 3.2 (Task diversity)** We define the quantities

$$\lambda_{\min}^{\mathbf{F}} := \lambda_{\min} \left( \frac{1}{T} \sum_{t=1}^T F_{\star}^{(t)\top} F_{\star}^{(t)} \right), \quad \lambda_{\max}^{\mathbf{F}} := \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T F_{\star}^{(t)\top} F_{\star}^{(t)} \right). \quad (10)$$

We now present our main result.

**Theorem 3.1 (Main result)** Let Assumption 3.1 hold. Assume the current representation iterate  $\hat{\Phi}$  satisfies  $\text{dist}(\hat{\Phi}, \Phi_{\star}) < \frac{1}{30} \frac{\lambda_{\min}^{\mathbf{F}}}{\sqrt{\lambda_{\max}^{\mathbf{F}}}} \min_t \kappa(\Sigma_x^{(t)})^{-1}$ , and

$$|\mathcal{N}_1| \geq \max_t \tau_{\text{mix}}^{(t)} \tilde{\Omega} \left( \max \left\{ \gamma^4 (r + \log(T/\delta)), \frac{\lambda_{\max}^{\mathbf{F}}}{(\lambda_{\min}^{\mathbf{F}})^2} \frac{\sigma_w^{(t)2}}{\|F_{\star}^{(t)}\|^2 \lambda_{\min}(\Sigma_x^{(t)})} (d_y + r + \log(T/\delta)) \right\} \right)$$

$$|\mathcal{N}_2| T \geq \max_t \tau_{\text{mix}}^{(t)} \tilde{\Omega} \left( \max \left\{ T \gamma^4 (d_x + \log(T/\delta)), \frac{\lambda_{\max}^{\mathbf{F}}}{(\lambda_{\min}^{\mathbf{F}})^2} \frac{\sigma_w^{(t)2}}{\|F_{\star}^{(t)}\|^2 \lambda_{\min}(\Sigma_x^{(t)})} (d_x + r + \log(T/\delta)) \right\} \right).$$

Given  $\eta \leq \frac{1}{3\lambda_{\max}^{\mathbf{F}}}$ , with probability greater than  $1 - \delta$ , the next iterate outputted by DFW (Algorithm 1) satisfies

$$\text{dist}(\hat{\Phi}_+, \Phi_{\star}) \leq \left(1 - \frac{\eta}{2} \lambda_{\min}^{\mathbf{F}}\right)^{1/2} \text{dist}(\hat{\Phi}, \Phi_{\star}) + \tilde{\mathcal{O}} \left( \sigma_{\text{avg}} \sqrt{\frac{d_x + \log\left(\frac{T}{\delta}\right)}{NT}} \right),$$

where  $\sigma_{\text{avg}} := \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{\sigma_w^{(t)2} \|F_{\star}^{(t)}\|^2}{\lambda_{\min}(\Sigma_x^{(t)})}}$  is the task-averaged noise-level.

**Remark 3.1 (Initialization)** We note that Theorem 3.1 relies on the representation  $\hat{\Phi}$  being sufficiently close to  $\Phi_{\star}$ . We do not address this issue in this paper, and refer to Collins et al. (2021); Tripuraneni et al. (2021); Thekumparampil et al. (2021) for initialization schemes in the iid linear regression setting. Our experiments suggest that initialization is often unnecessary, which mirrors the experimental findings in Thekumparampil et al. (2021, Sec. 6). We leave constructing an initialization scheme for our general setting, or showing it is unnecessary, to future work.

A key benefit of having the variance term in Theorem 3.1 scale properly in  $N, T$  is that we may construct representations on fixed datasets whose error scales on the same order as the oracle empirical risk minimizer by running Algorithm 1 on appropriately partitioned subsets of a given dataset.

**Corollary 3.1 (Approximate ERM)** Let the assumptions of Theorem 3.1 hold. Let  $\hat{\Phi}_0$  be an initial representation satisfying  $\text{dist}(\hat{\Phi}_0, \Phi_{\star}) < \nu$ , and define  $\rho := \frac{3}{64} \frac{\lambda_{\min}^{\mathbf{F}}}{\lambda_{\max}^{\mathbf{F}}}$ . Let  $\mathcal{D} := \{\{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^N\}_{t \in [T]}$  be a given dataset. There exists a partition of  $\mathcal{D}$  into independent batches  $\mathcal{B}_1, \dots, \mathcal{B}_K$ , such that iterating DFW on  $\mathcal{B}_k$ ,  $k \in [K]$  yields with probability greater than  $1 - \delta$ :

$$\text{dist}(\hat{\Phi}_K, \Phi_{\star})^2 \leq \tilde{\mathcal{O}} \left( C(\rho) \frac{\max_h \sigma_w^{(t)2} (d_x r + \log(1/\delta))}{NT} \right), \quad (11)$$

where  $C(\rho) > 0$  is a constant depending on the contraction rate  $\rho$ .

We note that the RHS of (11) has the ‘‘correct’’ scaling: noise level  $\sigma_w^{(t)2}$  multiplied by # parameters of the representation, divided by the total amount of data  $NT$ . In particular, given a fine-tuning dataset of size  $N'$  sampled from a task  $T + 1$  that shares the representation  $\Phi_{\star}$ , computing the least

squares weights  $\hat{F}^{(T+1)}$  conditioned on  $\hat{\Phi}_K$  yields a high-probability bound (cf. Appendix A.4.1) on the parameter error

$$\begin{aligned} \left\| \hat{F}^{(T+1)} \hat{\Phi}_K - M_\star^{(T+1)} \right\|_F^2 &\lesssim \text{dist}(\hat{\Phi}_K, \Phi_\star)^2 + \sigma_w^{(T+1)2} \frac{d_y r + \log(1/\delta)}{N'} \\ &\lesssim C(\rho) \frac{\max_t \sigma_w^{(t)2} (d_x r + \log(1/\delta))}{NT} + \frac{\sigma_w^{(T+1)2} (d_y r + \log(1/\delta))}{N'}, \end{aligned}$$

where we omit task-related quantities for clarity. We note that the above parameter recovery bound mirrors that of ERM estimates (Du et al., 2020; Zhang et al., 2023), where we note the latter fine-tuning term scales with  $d_y r$  (the number of parameters in  $F^{(T+1)}$ ) as opposed to  $r$  in the linear regression setting ( $d_y = 1$ ).

## 4 NUMERICAL VALIDATION

We present numerical experiments to demonstrate the key importance of aspects of our proposed algorithm. We consider two scenarios: 1. linear regression with non-isotropic *iid* data, and 2. linear system identification. The linear regression experiments highlight the breakdown of standard approaches when approached with mildly non-isotropic data, thus highlighting the necessity of our proposed `De-biasing & Feature-whitening` steps, and our system identification experiments demonstrate the applicability of our algorithm to sequentially dependent (non-isotropic) data. Additional experiments and experiment details can be found in Appendix C.

### 4.1 LINEAR REGRESSION WITH IID AND NON-ISOTROPIC DATA

We consider the observation model from (1), where we set the operator dimensions and rank as  $d_x = d_y = 50$  and  $r = 7$ . We generate the  $T$  operators using the following steps: 1. a ground truth representation  $\Phi_\star \in \mathbb{R}^{r \times d_x}$  is randomly generated through applying `thin_svd` to a random matrix with values  $\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ , 2. a nominal task weight matrix  $F_0 \in \mathbb{R}^{d_y \times r}$  is generated with elements  $\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ , 3. task-specific weights  $F_\star^{(t)} \in \mathbb{R}^{d_y \times r}$ ,  $t \in [T]$  are generated by applying random rotations to  $F_0$ . A non-isotropic covariance matrix  $\Sigma_x$  shared across all tasks is generated as  $\Sigma_x = \frac{d_x \tilde{\Sigma}_x}{\text{Tr}(\tilde{\Sigma}_x)}$ , where  $\tilde{\Sigma}_x = \frac{1}{2}(U + U^\top)$ ,  $U = 5 \cdot I_{d_x} + V$ , with  $V_{i,j} \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ . We note that by design of  $U$ ,  $\Sigma_x$  is only mildly non-isotropic. Figure 1a compares the performance of Algorithm 1, for a single-task ( $T = 1$ ) and multiple tasks ( $T = 25$ ), as well as standard alternating minimization-descent like `FedRep` (Collins et al., 2021) with  $T = 25$ . For each optimization iteration, we sample  $N = 100$  fresh data per task. The figure highlights that `DFW` is able to make use of all source data to decrease variance and learn a representation to near-optimality. As predicted in §3.1, Figure 1a also shows that vanilla alternating minimization-descent is not able to improve beyond a highly suboptimal bias, despite all tasks sharing the same, rather mildly non-isotropic covariate distribution.

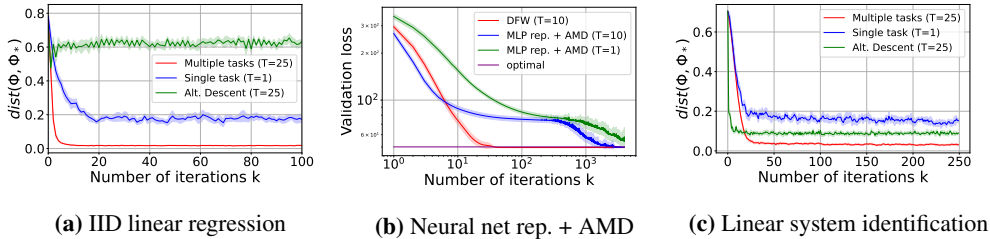
For a second experiment, we consider instead of applying vanilla alternating minimization-descent on a linear representation, we parameterize the representation by a neural network. In particular, we consider a ReLU-activated network with one hidden layer of dimension 64. We keep the same data-generation, with  $N = 100$  fresh samples per optimization iteration per task, and compare `DFW` (on a linear representation) to `AMD` on the neural net representation. Since the subspace distance is no longer defined for the neural net representation, we measure optimality of the learned representations by computing the average least-squares loss with respect to a validation set generated through the nominal weight matrix  $F_0$ . We plot the ‘‘optimal’’ loss as that attained by the ground truth operator  $F_0 \Phi_\star$ . We see in Figure 1b that, despite the much greater feature representation power of the neural net, alternating minimization-descent plateaus for many iterations just like the linear case before finding a non-linear parameter representation that allows descent to optimality, taking almost two orders of magnitude more iterations than `DFW` to reach optimality. This feature learning phase is only exacerbated when there are fewer tasks. We note that in such a streaming data setting, the high iteration complexity of `AMD` translates to a greatly hampered sample-efficiency compared to `DFW`.

### 4.2 LINEAR SYSTEM IDENTIFICATION

We consider a discrete-time linear system identification (sysID) problem, with dynamics

$$x_{i+1} = Ax_i + Bu_i + w_i, \quad i = 0, \dots, N - 1,$$





**Figure 1:** We plot the suboptimality the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 and the multi-task alternating minimization-descent. We observe performance improvement and variance reduction for multi-task DFW as predicted. All curves are plotted as the mean with 95% confidence regions shaded

where  $x_i$  is the state of the system and  $u_i$  is the control input. In contrast to the previous example, the covariates are now additionally non-iid due to correlation over time. In particular, we can instantiate multi-task linear sysID in the form of (1),

$$x_{i+1}^{(t)} = M_\star^{(t)} z_i^{(t)} + w_i^{(t)}, \quad i = 0, \dots, N - 1$$

where  $M_\star^{(t)} := [A^{(t)} \ B^{(t)}] = F^{(t)} \Phi_\star \in \mathbb{R}^{d_x \times d_z}$ . The state-action pair at time instant  $i$  for all tasks  $t \in [T]$  is embedded as  $z_i^{(t)} = [x_i^{(t)\top} \ u_i^{(t)\top}]^\top$ . The process noise  $w_i^{(t)}$  and control action  $u_i^{(t)}$  are assumed to be drawn from Gaussian distributions  $\mathcal{N}(0, \Sigma_w)$  and  $\mathcal{N}(0, \sigma_u^2 I_{d_u})$ , respectively, where  $d_u$  represents the dimension of the control action. We set the state dimension  $d_x = 25$ , control dimension  $d_u = 2$ , latent dimension  $r = 6$ , horizon  $N = 100$ , and input variance  $\sigma_u^2 = 1$ . The generation process of the ground truth system matrices  $M_\star^{(t)}$  follows a similar approach as described in the linear regression problem, with the addition of a normalization step of the nominal weight matrix  $F_0$  to ensure system stability for all tasks  $t \in [T]$ . Furthermore, the process noise covariance  $\Sigma_w$  is parameterized in a similar manner as in the linear regression example, with  $U = 5 \cdot I_{d_x} + 2 \cdot V$ . The initial state  $x_0^{(t)}$  is drawn iid across tasks from the system’s stationary distribution  $\mathcal{N}(0, \Sigma_x^{(t)})$ , which is determined by the solution to the discrete Lyapunov equation  $\Sigma_x^{(t)} = A^{(t)} \Sigma_x^{(t)} (A^{(t)})^\top + \sigma_u^2 B^{(t)} (B^{(t)})^\top + \Sigma_w$ . We note this implies the covariates  $x_i^{(t)}$  are inherently non-isotropic and non-identically distributed across tasks. Figure 1c again demonstrates the advantage of leveraging multi-task data to reduce the error in computing a shared representation across the system matrices  $M_\star^{(t)}$ . In line with our theoretical findings, DFW continues to benefit from multiple tasks, even when the data is sequentially dependent. We see that FedRep remains suboptimal in this non-iid, non-isotropic setting.

## 5 DISCUSSION AND FUTURE WORK

We propose an efficient algorithm to provably recover linear operators across multiple tasks to optimality from non-iid non-isotropic data, recovering near oracle empirical risk minimization rates. We show that the benefit of learning over multiple tasks manifests in a lower noise level in the optimization and smaller sample requirements for individual tasks. These results contribute toward a general understanding of representation learning from an algorithmic and statistical perspective. Some immediate open questions are: whether good initialization of the representation is necessary, and whether the convergence rate of DFW can be optimized e.g., through  $\ell^2$ -regularized weights  $\hat{F}^{(t)}$ . Resolving these questions has important implications for the natural extension of our framework: as emphasized in (Collins et al., 2021), the alternating empirical risk minimization (holding representation fixed) and gradient descent (holding task-specific weights fixed) framework naturally extends to the nonlinear setting. Providing guarantees for nonlinear function classes is an exciting and impactful avenue for future work, which concurrent work is moving toward, e.g. for 2-layer ReLU networks (Collins et al., 2023) and kernel ridge regression (Meunier et al., 2023). It remains to be seen whether a computationally-efficient algorithm can be established for nonlinear meta-learning in the non-iid and non-isotropic data regime, while preserving joint scaling in number of tasks and data.

## REFERENCES

- Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.
- Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *algorithmic learning theory*, pages 235–246. PMLR, 2019.
- 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.
- Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization, 2021.
- Nikunj Saunshi, Arushi Gupta, and Wei Hu. A representation learning perspective on the importance of train-validation splitting in meta-learning, 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.
- Yiting Chen, Ana M Ospina, Fabio Pasqualetti, and Emiliano Dall’Anese. Multi-task system identification of similar linear time-invariant dynamical systems. *arXiv preprint arXiv:2301.01430*, 2023.
- Thomas T Zhang, Katie Kang, Bruce D Lee, Claire Tomlin, Sergey Levine, Stephen Tu, and Nikolai Matni. Multi-task imitation learning for linear dynamical systems. In *Learning for Dynamics and Control Conference*, pages 586–599. PMLR, 2023.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.
- Seyedehsara Nayer and Namrata Vaswani. Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections. *IEEE Transactions on Information Theory*, 69(2):1177–1202, 2022.
- 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.
- Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart, 2018.
- 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.
- 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.
- Rui Lu, Gao Huang, and Simon S. Du. On the power of multitask representation learning in linear mdp, 2021.
- Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning, 2022.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning, 2015.
- 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.
- 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.
- Taosha Guo, Abed AlRahman Al Makdah, Vishaal Krishnan, and Fabio Pasqualetti. Imitation and transfer learning for lqg control. *arXiv preprint arXiv:2303.09002*, 2023.
- Lidong Li, Claudio De Persis, Pietro Tesi, and Nima Monshizadeh. Data-based transfer stabilization in linear systems. *arXiv preprint arXiv:2211.05536*, 2022.
- Han Wang, Leonardo F Toso, and James Anderson. Fedsysid: A federated approach to sample-efficient system identification. *arXiv preprint arXiv:2211.14393*, 2022.
- Lei Xin, Lintao Ye, George Chiu, and Shreyas Sundaram. Learning dynamical systems by leveraging data from similar systems. *arXiv preprint arXiv:2302.04344*, 2023.
- Mohamad Kazem Shirani Faradonbeh and Aditya Modi. Joint learning-based stabilization of multiple unknown linear systems. *IFAC-PapersOnLine*, 55(12):723–728, 2022.
- Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.
- Lloyd N Trefethen and David Bau. *Numerical linear algebra*, volume 181. Siam, 2022.
- Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.
- 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.
- Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. 2013.
- Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer relu neural networks. *arXiv preprint arXiv:2307.06887*, 2023.
- Dimitri Meunier, Zhu Li, Arthur Gretton, and Samory Kpotufe. Nonlinear meta-learning can guarantee faster rates. *arXiv preprint arXiv:2307.10870*, 2023.
- Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *The Journal of Machine Learning Research*, 22(1):1186–1246, 2021.
- Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.
- Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 5005–5014. PMLR, 2018.
- Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, aug 2011. doi: 10.1007/s10208-011-9099-z. URL <https://doi.org/10.1007%2Fs10208-011-9099-z>.
- Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.
- Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123–158, 2023.

- Karl Krauth, Stephen Tu, and Benjamin Recht. Finite-time analysis of approximate policy iteration for the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 32, 2019.
- Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.
- Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.
- 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.
- Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.
- Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *Advances in Neural Information Processing Systems*, 31, 2018.
- Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.
- Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019a.
- Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019b.
- Ingvar Ziemann, Anastasios Tsiamis, Henrik Sandberg, and Nikolai Matni. How are policy gradient methods affected by the limits of control? In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 5992–5999. IEEE, 2022.
- Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *arXiv preprint arXiv:2201.01680*, 2022.
- Bruce D Lee, Ingvar Ziemann, Anastasios Tsiamis, Henrik Sandberg, and Nikolai Matni. The fundamental limitations of learning linear-quadratic regulators. *arXiv preprint arXiv:2303.15637*, 2023.
- Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.
- Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. In *Conference on Learning Theory*, pages 3820–3857. PMLR, 2022.
- 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.
- Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.
- Alexander Goldenshluger and Assaf Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29(2):417–444, 2001.

## CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Related Work	2
<b>2</b>	<b>Problem Formulation</b>	<b>3</b>
<b>3</b>	<b>Sample-Efficient Linear Representation Learning</b>	<b>4</b>
3.1	Perils of (Vanilla) Gradient Descent on the Representation	4
3.2	A Task-Efficient Algorithm: De-bias & Feature-whiten	5
3.3	Algorithm Guarantees	6
<b>4</b>	<b>Numerical Validation</b>	<b>8</b>
4.1	Linear Regression with IID and Non-isotropic Data	8
4.2	Linear System Identification	8
<b>5</b>	<b>Discussion and Future Work</b>	<b>9</b>
<b>A</b>	<b>Theoretical Analysis of DFW (Algorithm 1)</b>	<b>14</b>
A.1	Preliminaries	14
A.2	The IID Setting	16
A.3	The Non-IID Setting	22
A.4	Converting to Sample Complexity Bounds	23
A.4.1	Near-ERM Transfer Learning	24
<b>B</b>	<b>Case Study: Linear Dynamical Systems</b>	<b>24</b>
B.1	Linear System Identification	25
B.2	Imitation Learning	26
<b>C</b>	<b>Additional Numerical Experiments and Details</b>	<b>27</b>
C.1	Linear Regression with IID and Non-isotropic Data	28
C.2	System Identification	28
C.3	Imitation Learning	29

## A THEORETICAL ANALYSIS OF DFW (ALGORITHM 1)

### A.1 PRELIMINARIES

We introduce some preliminary concepts and results that recur throughout our analysis. A fundamental concept in the analysis of least-squares solutions is the self-normalized martingale (Abbasi-Yadkori et al., 2011; Abbasi-Yadkori, 2013).

**Lemma A.1 (cf. Zhang et al. (2023, Lemma B.3))** *Let  $\{x_i\}_{i \geq 1}$  be a  $\mathbb{R}^{d_x}$ -valued process adapted to a filtration  $\{\mathcal{F}_i\}_{i \geq 1}$ . Let  $\{w_i\}_{i \geq 1}$  be a  $\mathbb{R}^{d_y}$ -valued process adapted to  $\{\mathcal{F}_i\}_{i \geq 2}$ . Suppose that  $\{w_i\}_{i \geq 1}$  is a  $\sigma^2$ -subgaussian martingale difference sequence, i.e.,:*

$$\mathbb{E}[w_i \mid \mathcal{F}_i] = 0, \quad (12)$$

$$\mathbb{E}[\exp(\lambda v^\top w_i) \mid \mathcal{F}_i] \leq \exp\left(\frac{\lambda^2 \sigma^2 \|v\|^2}{2}\right) \quad \forall \mathcal{F}_i\text{-measurable } \lambda \in \mathbb{R}, v \in \mathbb{R}^{d_y}. \quad (13)$$

For  $\Lambda \in \mathbb{R}^{d_y \times d_x}$ , let  $\{M_k(\Lambda)\}_{k \geq 1}$  be the  $\mathbb{R}$ -valued process:

$$M_k(\Lambda) = \exp\left(\frac{1}{\sigma} \sum_{i=1}^k \langle \Lambda x_i, w_i \rangle - \frac{1}{2} \sum_{i=1}^k \|\Lambda x_i\|^2\right). \quad (14)$$

Then, the process  $\{M_k(\Lambda)\}_{k \geq 1}$  satisfies  $\mathbb{E}[M_k(\Lambda)] \leq 1$  for all  $k \geq 1$ .

In particular, this implies the following self-normalized martingale inequality that handles multiple matrix-valued self-normalized martingales simultaneously. This can be seen as an instantiation of the Hilbert space variant from Abbasi-Yadkori (2013).

**Proposition A.1 (cf. Zhang et al. (2023, Prop. B.1), Sarkar et al. (2021, Proposition 8.2))** *Fix  $T \in \mathbb{N}_+$ . For  $t \in [T]$ , let  $\{x_i^{(t)}, w_i^{(t)}\}_{i \geq 1}$  be a  $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ -valued process and  $\{\mathcal{F}_i^{(t)}\}_{i \geq 1}$  be a filtration such that  $\{x_i^{(t)}\}_{i \geq 1}$  is adapted to  $\{\mathcal{F}_i^{(t)}\}_{i \geq 1}$ ,  $\{w_i^{(t)}\}_{i \geq 1}$  is adapted to  $\{\mathcal{F}_i^{(t)}\}_{i \geq 2}$ , and  $\{w_i^{(t)}\}_{i \geq 1}$  is a  $\sigma^2$ -subgaussian martingale difference sequence. Suppose that for all  $t_1 \neq t_2$ , the process  $\{x_i^{(t_1)}, w_i^{(t_1)}\}$  is independent of  $\{x_i^{(t_2)}, w_i^{(t_2)}\}$ . Fix (non-random) positive definite matrices  $\{V^{(t)}\}_{t=1}^T$ . For  $k \geq 1$  and  $t \in [T]$ , define  $\hat{\Sigma}_k^{(t)} := \sum_{i=1}^k x_i x_i^\top$ . Then, given any fixed  $N, T \in \mathbb{N}_+$ , with probability at least  $1 - \delta$ :*

$$\sum_{t=1}^T \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_F^2 \leq d_y \sigma^2 \sum_{t=1}^T \log \left( \frac{\det \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)}{\det(V^{(t)})} \right) + 2\sigma^2 \log(1/\delta) \quad (15)$$

Alternatively, in the spectral norm, we have with probability at least  $1 - \delta$ ,

$$\begin{aligned} \sum_{t=1}^T \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_2^2 &\leq 4\sigma^2 \sum_{t=1}^T \log \left( \frac{\det \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)}{\det(V^{(t)})} \right) \\ &+ 13d_y T \sigma^2 + 8\sigma^2 \log(1/\delta). \end{aligned} \quad (16)$$

We note that the above bound also holds for individual tasks  $t \in [T]$  simply by removing the summand over  $t$ . We introduce the following useful two-sided concentration inequality for the sample covariance of iid subgaussian covariates.

**Lemma A.2 (Du et al. (2020, Claim A.1, A.2))** *Let  $x_1, \dots, x_N \in \mathbb{R}^d$  be iid random vectors that satisfy  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^\top] = \Sigma$ , and  $x_i$  is  $\gamma^2$ -subgaussian. Fix  $\delta \in (0, 1)$ . Suppose  $N \gtrsim \gamma^4 (d + \log(1/\delta))$ . Then with probability at least  $1 - \delta$ , the following holds*

$$0.9\Sigma \preceq \frac{1}{N} \sum_{i=1}^N x_i x_i^\top \preceq 1.1\Sigma. \quad (17)$$

Furthermore, for any matrix  $U \in \mathbb{R}^{r \times d_x}$ , as long as  $N \gtrsim \gamma^4 (r + \log(1/\delta))$ , we have with probability at least  $1 - \delta$

$$0.9U\Sigma U^\top \preceq \frac{1}{N} \sum_{i=1}^N Ux_i x_i^\top U^\top \preceq 1.1U\Sigma U^\top. \quad (18)$$

Combining Proposition A.1 and Lemma A.2 yields the following self-normalized martingale bound without the non-random lower bound  $V^{(t)}$ .

**Lemma A.3** Consider the quantities defined in Proposition A.1 and assume  $x_i^{(t)}$  are zero-mean and  $\gamma^2$ -subgaussian. Then, as long as  $N \gtrsim \gamma^4 (d_x + \log(T/\delta))$ , with probability at least  $1 - \delta$ :

$$\begin{aligned} \sum_{t=1}^T \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_F^2 &\leq 2d_y d_x T \sigma^2 + 4\sigma^2 \log(T/\delta), \text{ or} \\ \sum_{t=1}^T \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_2^2 &\leq 8d_x T \sigma^2 + 26d_y T \sigma^2 + 16\sigma^2 \log(T/\delta) \end{aligned}$$

*Proof:* we observe that if  $\hat{\Sigma}_N^{(t)} \succeq V^{(t)}$ , then

$$2\hat{\Sigma}_N^{(t)} \succeq V^{(t)} + \hat{\Sigma}_N^{(t)} \implies \left( \hat{\Sigma}_N^{(t)} \right)^{-1} \preceq 2 \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)^{-1}.$$

This implies

$$\begin{aligned} \sum_{t=1}^T \mathbf{1} \left\{ \hat{\Sigma}_N^{(t)} \succeq V^{(t)} \right\} \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_F^2 \\ \leq 2 \sum_{t=1}^T \mathbf{1} \left\{ \hat{\Sigma}_N^{(t)} \succeq V^{(t)} \right\} \left\| \sum_{i=1}^N w_i^{(t)} x_i^{(t)\top} \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)^{-1/2} \right\|_F^2. \end{aligned}$$

Defining  $\Sigma^{(t)} := \mathbb{E} \left[ x_i^{(t)} x_i^{(t)\top} \right]$ , let us consider for each  $t$  the event:

$$0.9\Sigma^{(t)} \preceq \hat{\Sigma}_N^{(t)} \preceq 1.1\Sigma^{(t)},$$

which by Lemma A.2 occurs with probability at least  $1 - \delta$  as long as  $N \gtrsim \gamma^4 (d_x + \log(1/\delta))$ . Setting  $V^{(t)} := 0.9N\Sigma^{(t)}$  and conditioning on the above event, we observe that by definition  $\hat{\Sigma}_N^{(t)} \succeq V^{(t)}$ , and

$$\begin{aligned} \log \frac{\det \left( V^{(t)} + \hat{\Sigma}_N^{(t)} \right)}{\det(V^{(t)})} &= \log \det \left( I_{d_x} + \hat{\Sigma}_N^{(t)} \left( V^{(t)} \right)^{-1} \right) \\ &= \log \det \left( \left( 1 + \frac{1.1}{0.9} \right) I_{d_x} \right) \\ &\leq d_x. \end{aligned}$$

Plugging this into Proposition A.1 and union-bounding over the desirable event over each  $t \in [T]$ , and adjusting the failure probability  $\delta/T \mapsto \delta$ , we get our desired result.  $\blacksquare$

In order to instantiate our bounds for non-iid covariates, we introduce the notions of  $\beta$ -mixing stationary processes (Kuznetsov and Mohri, 2017; Tu and Recht, 2018).

**Definition A.1 ( $\beta$ -mixing)** Let  $\{x_i\}_{t \geq 1}$  be a  $\mathbb{R}^d$ -valued discrete-time stochastic process adapted to filtration  $\{\mathcal{F}_i\}_{t=1}^\infty$ . We denote the stationary distribution  $\nu_\infty$ . We define the  $\beta$ -mixing coefficient

$$\beta(k) := \sup_{t \geq 1} \mathbb{E}_{\{x_\ell\}_{\ell=1}^t} \left[ \left\| \mathbb{P}_{x_{t+k}}(\cdot | \mathcal{F}_t) - \nu_\infty \right\|_{\text{TV}} \right], \quad (19)$$

where  $\|\cdot\|_{\text{TV}}$  denotes the total variation distance between probability measures.

Intuitively, the  $\beta$ -mixing coefficient measures how quickly on average a process mixes to the stationary distribution along any sample path. To see how  $\beta$ -mixing is instantiated, let  $\{x_i\}_{i=1}^T$  be a sample path from a  $\beta$ -mixing process. Consider the following subsampled paths formed by taking every  $a$ -th covariate of  $\{x_i\}$ :

$$X_{(j)}^T := \{x_i : 1 \leq t \leq T, (t-1 \bmod a) = j-1\}, \quad j = 1, \dots, a. \quad (20)$$

Let the integers  $m_1, \dots, m_a$  and index sets  $I_{(1)}, \dots, I_{(a)}$  denote the sizes and indices of  $X_{(1)}^T, \dots, X_{(a)}^T$ , respectively. Finally, let  $X_\infty^{m_j}$  denote a sequence of  $m_j$  iid draws from the stationary distribution  $\nu_\infty$ . The following is a key lemma in relating a correlated process to iid draws.

**Lemma A.4 (Kuznetsov and Mohri (2017, Proposition 2))** *Let  $g(\cdot)$  be a real-valued Borel-measurable function satisfying  $-M_1 \leq g(\cdot) \leq M_2$  for some  $M_1, M_2 \geq 0$ . Then, for all  $j = 1, \dots, a$ ,*

$$\left| \mathbb{E}[g(X_\infty^{m_j})] - \mathbb{E}[g(X_{(j)}^T)] \right| \leq (M_1 + M_2)m_j\beta(a).$$

In our analysis, we often instantiate Lemma A.4 with  $g(\cdot)$  as an indicator function on a success event. For appropriately selected block length  $a$ , we are thus able to relate simpler iid analysis on  $X_\infty^{m_j}$  to the original process  $X_{(j)}^T$ , accruing an additional factor in the failure probability. Lastly, we introduce a standard matrix concentration inequality.

**Lemma A.5 (Matrix Hoeffding (Tropp, 2011))** *Let  $\{X_t\}_{t=1}^T \subset \mathbb{R}^{d \times d}$  be a sequence of independent, random symmetric matrices, and let  $\{B_t\}_{t=1}^T$  be a sequence of fixed symmetric matrices. Assume each random matrix satisfies*

$$\mathbb{E}[X_t] = 0, \quad X_t^2 \preceq B_t^2 \text{ almost surely.}$$

Then for all  $t \geq 0$ ,

$$\mathbb{P} \left[ \lambda_{\max} \left( \sum_{t=1}^T X_t \right) \geq t \right] \leq d \cdot \exp \left( -\frac{t^2}{8\sigma^2} \right), \quad \sigma^2 := \left\| \sum_{t=1}^T B_t^2 \right\|.$$

In particular, for general rectangular  $\{M_t\}_{t=1}^T \subset \mathbb{R}^{d_1 \times d_2}$ , we may define  $X_t := \begin{bmatrix} 0 & M_t \\ M_t^\top & 0 \end{bmatrix}$  to yield a singular value concentration inequality. Assume each  $M_t$  satisfies

$$\mathbb{E}[M_t] = 0, \quad X_t^2 \preceq B_t^2 \text{ almost surely.}$$

Then for all  $t \geq 0$ ,

$$\mathbb{P} \left[ \sigma_{\max} \left( \sum_{t=1}^T M_t \right) \geq t \right] \leq (d_1 + d_2) \cdot \exp \left( -\frac{t^2}{8\sigma^2} \right), \quad \sigma^2 := \left\| \sum_{t=1}^T B_t^2 \right\|.$$

As hinted by the indexing of the matrices, by leveraging the independence of processes across tasks  $h$ , Lemma A.5 can be used to bound various quantities averaged across tasks, under the important caveat that the matrices are *zero-mean*, which ties back to the necessity of our de-biasing and feature-whitening adjustments.

## A.2 THE IID SETTING

We recall that given the current representation iterate  $\hat{\Phi}$ , an iid draw of a multitask dataset  $\{(x_i^{(t)}, y_i^{(t)})\}_{t=1, i=1}^{T, N}$ ,  $h = 1, \dots, H$ , and DFW trajectory partitions  $\mathcal{N}_1, \mathcal{N}_2$ , the least squares weights  $\hat{F}^{(t)}$  can be written as

$$\begin{aligned} \hat{F}^{(t)} &= \operatorname{argmin}_F \sum_{i \in \mathcal{N}_1} \left\| y_i^{(t)} - F z_i^{(t)} \right\|^2 \\ &= F_\star^{(t)} \Phi_\star X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} + W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \\ &= F_\star^{(t)} \Phi_\star \hat{\Phi}^\top + F_\star^{(t)} \Phi_\star \left( I_{d_x} - \hat{\Phi}^\top \hat{\Phi} \right) X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} + W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1}. \end{aligned} \quad (21)$$



Now recalling the DFW representation update in the iid setting (5), we have

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \left( \hat{F}^{(t)} \hat{\Phi} - F_\star^{(t)} \Phi_\star \right) - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_x^{(t)} \right)^{-1}. \quad (22)$$

Right multiplying the update by  $\Phi_{\star,\perp}^\top$ , we get

$$\begin{aligned} R\hat{\Phi}_+ \Phi_{\star,\perp}^\top &= \hat{\Phi} \Phi_{\star,\perp}^\top - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \left( \hat{F}^{(t)} \hat{\Phi} - F_\star^{(t)} \Phi_\star \right) \Phi_{\star,\perp}^\top - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x,\mathcal{N}_2}^{(t)} \right)^{-1} \Phi_{\star,\perp}^\top \\ &= \left( I_{d_x} - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \hat{\Phi} \Phi_{\star,\perp}^\top - \frac{\eta}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x,\mathcal{N}_2}^{(t)} \right)^{-1} \Phi_{\star,\perp}^\top, \end{aligned}$$

where the last line is composed of a contraction term and a noise term. We start with an analysis of the noise term.

#### BOUNDING THE NOISE TERM

We observe that since  $\hat{F}^{(t)}$  is by construction independent of  $W_{\mathcal{N}_2}^{(t)}, X_{\mathcal{N}_2}^{(t)}$ , by the independence of  $x_i^{(t)}$  across  $t$  and  $i$ , and the noise independence  $w_i^{(t)} \perp x_i^{(t)}$ , we find

$$\mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x,\mathcal{N}_2}^{(t)} \right)^{-1} \right] = \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \hat{F}^{(t)\top} \right] \mathbb{E} \left[ W_{\mathcal{N}_2}^{(t)} \right] \mathbb{E} \left[ X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x,\mathcal{N}_2}^{(t)} \right)^{-1} \right] = 0.$$

Therefore, we set up for an application of Lemma A.5. Toward doing so, we prove the following two ingredients: 1. a high probability bound on  $\|\hat{F}^{(t)}\|$ , 2. a high probability bound on the least-squares noise-esque term  $\|\hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x,\mathcal{N}_2}^{(t)} \right)^{-1}\|$ . We then condition on these two high-probability events to instantiate the almost-sure boundedness in Lemma A.5. We start with the analysis of  $\hat{F}^{(t)}$ . By (21), we have

$$\|\hat{F}^{(t)}\| \leq \|F_\star^{(t)}\| + \|F_\star^{(t)}\| \left\| \Phi_\star \mathcal{P}_\Phi^\perp X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z,\mathcal{N}_1}^{(t)} \right)^{-1} \right\| + \left\| W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z,\mathcal{N}_1}^{(t)} \right)^{-1} \right\|.$$

**Lemma A.6** *Let  $|\mathcal{N}_1| := N_1 \gtrsim \gamma^4 (r + \log(1/\delta))$ . Then, with probability greater than  $1 - \delta$ , we have*

$$\left\| \Phi_\star \mathcal{P}_\Phi^\perp X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z,\mathcal{N}_1}^{(t)} \right)^{-1} \right\| \leq \frac{5}{4} \text{dist}(\hat{\Phi}, \Phi_\star) \kappa \left( \Sigma_x^{(t)} \right), \quad (23)$$

$$\left\| W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z,\mathcal{N}_1}^{(t)} \right)^{-1} \right\| \lesssim \sigma_w^{(t)} \sqrt{\frac{d_y + r + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(t)}) N_1}}. \quad (24)$$

*Proof:* we begin with the bound (23). We observe that by definition  $\frac{1}{N_1} X_{\mathcal{N}_1}^{(t)\top} X_{\mathcal{N}_1}^{(t)} = \hat{\Sigma}_{x,\mathcal{N}_1}^{(t)}$ , and  $\Phi_\star \mathcal{P}_\Phi^\perp, \hat{\Phi}$  are  $r \times d_x$  matrices. Therefore, we invoke Lemma A.2 twice to find: as long as  $N_1 \gtrsim \gamma^4 (r + \log(1/\delta))$ , with probability at least  $1 - \delta$ , the following bounds hold simultaneously:

$$\begin{aligned} \frac{1}{N_1} \left\| \Phi_\star \mathcal{P}_\Phi^\perp X_{\mathcal{N}_1}^{(t)\top} \right\|^2 &\leq 1.1 \left\| \Phi_\star \mathcal{P}_\Phi^\perp \left( \Sigma_x^{(t)} \right)^{1/2} \right\|^2 \leq 1.1 \text{dist}(\hat{\Phi}, \Phi_\star)^2 \lambda_{\max} \left( \Sigma_x^{(t)} \right) \\ \frac{1}{N_1} \left\| Z_{\mathcal{N}_1}^{(t)} \right\|^2 &\leq 1.1 \left\| \Sigma_z^{(t)} \right\| \leq 1.1 \lambda_{\max} \left( \Sigma_x^{(t)} \right) \\ \left\| \left( \hat{\Sigma}_{z,\mathcal{N}_1}^{(t)} \right)^{-1} \right\| &\leq 0.9 \lambda_{\min} \left( \Sigma_z^{(t)} \right)^{-1} \leq 0.9 \lambda_{\min} \left( \Sigma_x^{(t)} \right)^{-1}, \end{aligned}$$

where we recall that  $\Sigma_z^{(t)} = \hat{\Phi} \Sigma_x^{(t)} \hat{\Phi}^\top$ . Therefore, applying Cauchy-Schwarz on the LHS of (23) and the above bounds (converting  $1.1/0.9 < 5/4$ ) yields the desired upper bound on the RHS.

Moving onto (24), we observe that since  $\hat{\Phi}$  is fixed,  $\{w_i^{(t)}, z_i^{(t)}\}_{i \geq 1}$  is an  $\mathbb{R}^{d_y} \times \mathbb{R}^r$ -valued martingale difference sequence. Therefore, we may apply Lemma A.2 and Lemma A.3 to find: as long as  $N_1 \gtrsim \gamma^4 (r + \log(1/\delta))$ , with probability at least  $1 - \delta$ ,

$$\begin{aligned} \left\| W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \right\| &\leq \left\| W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1/2} \right\| \lambda_{\min} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1/2} \\ &\lesssim \sqrt{\frac{\sigma_w^{(t)2} (d_y + r + \log(1/\delta))}{\lambda_{\min}(\Sigma_x^{(t)}) N_1}}, \end{aligned}$$

which establishes the bound (24).  $\blacksquare$

Therefore, by assuming  $\text{dist}(\hat{\Phi}, \Phi_*)$  is sufficiently small, and that  $N_1$  is large enough to offset the noise bound (24), we immediately get the following bound relating  $\|\hat{F}^{(t)}\|$  to  $\|F_*^{(t)}\|$ .

**Lemma A.7** *Assume*

$$N_1 \gtrsim \max \left\{ \gamma^4 (r + \log(1/\delta)), \frac{\sigma_w^{(t)2}}{C^2 \|F_*^{(t)}\|^2 \lambda_{\min}(\Sigma_x^{(t)})} (d_y + r + \log(1/\delta)) \right\},$$

and  $\text{dist}(\hat{\Phi}, \Phi_*)^2 \leq \frac{2}{5} C^2 \kappa \left( \Sigma_x^{(t)} \right)^{-1}$ , for fixed  $C > 0$ . Then with probability at least  $1 - \delta$

$$\|\hat{F}^{(t)}\| \leq (1 + C) \|F_*^{(t)}\|. \quad (25)$$

The factor  $C > 0$  serves as a free parameter which we can determine later—a larger  $C$  implies a relaxed requirement on the initial condition (disappears when  $C \geq \frac{5}{2} \kappa(\Sigma_x^{(t)})$ ) and burn-in requirement on  $N_1$ , but results in a larger subgaussian-parameter bound on the representation update noise term (22), as we demonstrate: given that the success event of Lemma A.7 holds, then by the definition of subgaussianity (Assumption 3.1), we observe that  $\hat{F}^{(t)\top} w_i^{(t)}$  is zero-mean and  $(1 + C)^2 \sigma_w^{(t)2} \|F_*^{(t)}\|^2$ -subgaussian, supported on  $\mathbb{R}^r$ . Therefore, bounding

$$\left\| \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \leq \left\| \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1/2} \right\| \lambda_{\min} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1/2},$$

we invoke Proposition A.1 and Lemma A.2 to get the following bound.

**Lemma A.8** *Let the conditions of Lemma A.7 hold. Then, as long as  $|\mathcal{N}_2| := N_2 \gtrsim \gamma^4 (d_x + \log(1/\delta))$ , with probability at least  $1 - \delta$ ,*

$$\left\| \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \lesssim (1 + C) \sigma_w^{(t)} \|F_*^{(t)}\| \sqrt{\frac{d_x + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(t)}) N_2}}. \quad (26)$$

With a bound on the task-specific noise term in hand, we may now produce the final bound on the (task-averaged) noise term.

**Proposition A.2 (Noise term bound)** *Assume*

$$N_1 \gtrsim \max \left\{ \gamma^4 (r + \log(T/\delta)), \max_t \frac{\sigma_w^{(t)2}}{C^2 \|F_*^{(t)}\|^2 \lambda_{\min}(\Sigma_x^{(t)})} (d_y + r + \log(T/\delta)) \right\},$$

$$N_2 \gtrsim \gamma^4 (d_x + \log(T/\delta)),$$

and  $\text{dist}(\hat{\Phi}, \Phi_*)^2 \leq \max_t \frac{2}{5} C^2 \kappa \left( \Sigma_x^{(t)} \right)^{-1}$ , for fixed  $C > 0$ . Then, with probability at least  $1 - \delta$ ,

$$\left\| \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \lesssim \sigma_{\text{avg}} (1 + C) \sqrt{\frac{d_x + \log(T/\delta)}{T N_2} \log \left( \frac{d_y + r}{\delta} \right)},$$

where  $\sigma_{\text{avg}} := \sqrt{\frac{1}{T} \sum_{t=1}^T \frac{\sigma_w^{(t)2} \|F_*^{(t)}\|^2}{\lambda_{\min}(\Sigma_x^{(t)})}}$  is the task-averaged noise-level.

*Proof of Proposition A.2:* we set up for an application of the Matrix Hoeffding bound (Lemma A.5). By union bounding over the task-specific noise bound Lemma A.8, we have with probability at least  $1 - \delta$ , for all  $t \in [T]$  simultaneously:

$$\left\| \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \lesssim (1 + C) \frac{\sigma_w^{(t)} \|F_\star^{(t)}\|}{\sqrt{\lambda_{\min}(\Sigma_x^{(t)})}} \sqrt{\frac{d_x + \log(T/\delta)}{N_2}},$$

given the assumed burn-in conditions hold. Therefore, by setting

$$\begin{aligned} B^{(t)} &= \mathcal{O}(1) \frac{\sigma_w^{(t)} \|F_\star^{(t)}\|}{\sqrt{\lambda_{\min}(\Sigma_x^{(t)})}} \sqrt{\frac{d_x + \log(T/\delta)}{N_2}} I_{d_x+r} \\ \sigma^2 &= \left\| \sum_{t=1}^T (B^{(t)})^2 \right\| = \left( \sum_{t=1}^T \frac{\sigma_w^{(t)2} \|F_\star^{(t)}\|^2}{\lambda_{\min}(\Sigma_x^{(t)})} \right) (1 + C)^2 \frac{d_x + \log(T/\delta)}{N_2} \\ &= T \sigma_{\text{avg}}^2 (1 + C)^2 \frac{d_x + \log(T/\delta)}{N_2} \end{aligned}$$

we invoke Lemma A.5 and invert

$$(d_x + r) \cdot \exp\left(\frac{-t^2}{8\sigma^2}\right) \leq \delta$$

to set

$$t \approx \sqrt{T} \sigma_{\text{avg}} (1 + C) \sqrt{\frac{d_x + \log(T/\delta)}{N_2} \log\left(\frac{d_y + r}{\delta}\right)}.$$

The resulting Hoeffding bound yields

$$\begin{aligned} \mathbb{P} \left[ \left\| \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \geq \sqrt{T} \sigma_{\text{avg}} (1 + C) \sqrt{\frac{d_x + \log(T/\delta)}{N_2} \log\left(\frac{d_y + r}{\delta}\right)} \right] &\leq \delta \\ \iff \mathbb{P} \left[ \left\| \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} W_{\mathcal{N}_2}^{(t)\top} X_{\mathcal{N}_2}^{(t)} \left( \hat{\Sigma}_{x, \mathcal{N}_2}^{(t)} \right)^{-1} \right\| \geq \sigma_{\text{avg}} (1 + C) \sqrt{\frac{d_x + \log(T/\delta)}{TN_2} \log\left(\frac{d_y + r}{\delta}\right)} \right] &\leq \delta \end{aligned}$$

We have bounded the noise term on the DFW representation update, demonstrating critically the bound on the noise term benefits from a scaling with the number of tasks  $T$ . The task-relevant quantity  $\sigma_{\text{avg}}$  quantifies that the “noise level” of the problem is an *average* over the noise-levels of each task. We note that our application of Matrix Hoeffding is rather crude, and the above bound can likely be improved in terms of  $\text{polylog}(1/\delta)$  factors with stronger moment bounds on the matrix-valued self-normalized martingale terms, but this is out of the scope of this paper.

Returning to the choice of  $C$ ,  $C \approx 1$  implies no further system/task-specific dependence beyond the terms in  $K^{(1:T)}$ ; however, this may translate into a stringent requirement on the burn-in  $N_1$  and the subspace distance  $\text{dist}(\hat{\Phi}, \Phi_\star)$ . On the other hand,  $C \gtrsim \sqrt{\kappa(\Sigma_x^{(t)})}$  relaxes the burn-in and potentially renders the subspace distance requirement trivial, but manifests a condition number in the noise bound. We note that as we expect  $\text{dist}(\hat{\Phi}, \Phi_\star)$  to decrease geometrically with iterations of DFW, the subspace distance requirement is only relevant for the first few iterations. In general, this intuitively captures the cost of ill-conditioned data distributions. We now move on to bounding the contraction term.

#### BOUNDING THE CONTRACTION TERM

Let us define,

$$\begin{aligned} \Delta^{(t)} &:= \Phi_\star \left( I_{d_x} - \hat{\Phi}^\top \hat{\Phi} \right) X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \\ E^{(t)} &:= W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1}, \end{aligned}$$

such that we may write (21) as  $\hat{F}^{(t)} = F_\star^{(t)} \Phi_\star \hat{\Phi}^\top + F_\star^{(t)} \Delta^{(t)} + E^{(t)}$ . We expand

$$\begin{aligned} \hat{F}^{(t)\top} \hat{F}^{(t)} &= \hat{\Phi} \Phi_\star^\top F_\star^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top + \Delta^{(t)\top} \Delta^{(t)} + E^{(t)\top} E^{(t)} \\ &\quad + \text{Sym}(\Delta^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(E^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(\Delta^{(t)\top} E^{(t)}), \end{aligned}$$

where  $\text{Sym}(A) := A + A^\top$ . We will make repeated use of the following matrix Cauchy-Schwarz-type lemma.

**Lemma A.9** *Let  $A_t, B_t$  be real-valued matrices for  $t = 1, \dots, T$ . Then,*

$$\left\| \sum_{t=1}^T A_t B_t \right\| \leq \left\| \sum_{t=1}^T A_t A_t^\top \right\|^{1/2} \left\| \sum_{t=1}^T B_t^\top B_t \right\|^{1/2}.$$

Now taking the average over tasks  $T$ , we then may write

$$\begin{aligned} &\lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \\ &\leq \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \right) + \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} \Delta^{(t)} \right) + \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T E^{(t)\top} E^{(t)} \right) \\ &\quad + \left\| \text{Sym} \left( \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right) \right\| + \left\| \text{Sym} \left( \frac{1}{T} \sum_{t=1}^T E^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right) \right\| + \left\| \text{Sym} \left( \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} E^{(t)} \right) \right\|. \end{aligned}$$

We observe that

$$\begin{aligned} \|\text{Sym}(A)\| &= \max_{\|u\|, \|v\|=1} x^\top A v + x^\top A^\top v \\ &\leq 2 \|A\|, \end{aligned}$$

and thus applying the above fact and Lemma A.9 on the cross terms yields

$$\begin{aligned} &\lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \\ &\leq \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \right) + \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} \Delta^{(t)} \right) + \lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T E^{(t)\top} E^{(t)} \right) \\ &\quad + 2 \left\| \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} \Delta^{(t)} \right\|^{1/2} \left\| \frac{1}{T} \sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \right\|^{1/2} + 2 \left\| \frac{1}{T} \sum_{t=1}^T E^{(t)\top} E^{(t)} \right\|^{1/2} \left\| \frac{1}{T} \sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \right\|^{1/2} \\ &\quad + 2 \left\| \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} \Delta^{(t)} \right\|^{1/2} \left\| \frac{1}{T} \sum_{t=1}^T E^{(t)\top} E^{(t)} \right\|^{1/2}. \end{aligned}$$

Using Lemma A.6, we get the following upper bound on  $\lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right)$ .

**Lemma A.10** *Let*

$$N_1 \gtrsim \max \left\{ \gamma^4 (r + \log(T/\delta)), \frac{\lambda_{\max}^{\mathbf{F}} - 1}{T} \sum_{t=1}^T \frac{\sigma_w^{(t)2} (d_y + r + \log(T/\delta))}{\lambda_{\min}(\Sigma_x^{(t)})} \right\},$$

and  $\text{dist}(\hat{\Phi}, \Phi_\star) \leq \frac{1}{5} \frac{\sqrt{\lambda_{\max}^{\mathbf{F}}}}{T} \sum_{t=1}^T \kappa(\Sigma_x^{(t)})^{-1}$ . Then, with probability at least  $1 - \delta$ , we have

$$\lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \leq 3\lambda_{\max}^{\mathbf{F}}.$$

*Proof:* we see that it suffices to establish  $\left\| \frac{1}{T} \sum_{t=1}^T \Delta^{(t)\top} \Delta^{(t)} \right\| \lesssim \left\| \frac{1}{T} \sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \right\| =: \lambda_{\max}^{\mathbf{F}}$  and  $\left\| \frac{1}{T} \sum_{t=1}^T E^{(t)\top} E^{(t)} \right\| \lesssim \lambda_{\max}^{\mathbf{F}}$  in order to establish

$$\lambda_{\max} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \lesssim \lambda_{\max}^{\mathbf{F}}.$$

We recall that

$$\begin{aligned} \Delta^{(t)} &:= \Phi_\star \left( I_{d_x} - \hat{\Phi}^\top \hat{\Phi} \right) X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \\ E^{(t)} &:= W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1}, \end{aligned}$$

which by Lemma A.6 admit the following high-probability bounds: let  $|\mathcal{N}_1| := N_1 \gtrsim \gamma^4 (r + \log(1/\delta))$ . Then, with probability greater than  $1 - \delta$ , we have

$$\begin{aligned} \left\| \Phi_\star \mathcal{P}_{\hat{\Phi}}^\perp X_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \right\| &\leq \frac{5}{4} \text{dist}(\hat{\Phi}, \Phi_\star) \kappa \left( \Sigma_x^{(t)} \right) \\ \left\| W_{\mathcal{N}_1}^{(t)\top} Z_{\mathcal{N}_1}^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \right\| &\lesssim \sigma_w^{(t)} \sqrt{\frac{d_y + r + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(t)}) N_1}}. \end{aligned}$$

Therefore, inverting the RHS' for a prescribed numerical constant factor of  $\lambda_{\max}^{\mathbf{F}}$  yields the result. ■

Lemma A.10 informs the maximum we may set the step-size. To now upper bound the contraction rate, we lower bound  $\lambda_{\min}(\hat{F}^{(t)\top} \hat{F}^{(t)})$ . We observe that the diagonal terms  $\Delta^{(t)\top} \Delta^{(t)}$ ,  $E^{(t)\top} E^{(t)}$  are pd, and thus can be ignored. We then observe that by Weyl's inequality (Horn and Johnson, 2012), we have

$$\begin{aligned} \lambda_{\min}(\hat{F}^{(t)\top} \hat{F}^{(t)}) &\geq \lambda_{\min} \left( \hat{\Phi} \Phi_\star^\top F_\star^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right) \\ &\quad - \lambda_{\max} \left( \text{Sym}(\Delta^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(E^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(\Delta^{(t)\top} E^{(t)}) \right). \end{aligned}$$

Just as in the upper bound of  $\lambda_{\max}(\hat{F}^{(t)\top} \hat{F}^{(t)})$ , it suffices to upper bound the cross terms. Therefore, following the same proof as Lemma A.10, we have the analogous result.

**Lemma A.11** *Let*

$$N_1 \gtrsim \max \left\{ \gamma^4 (r + \log(1/\delta)), \frac{\lambda_{\max}^{\mathbf{F}}}{(\lambda_{\min}^{\mathbf{F}})^2} \frac{1}{T} \sum_{t=1}^T \frac{\sigma_w^{(t)2} (d_y + r + \log(1/\delta))}{\lambda_{\min}(\Sigma_x^{(t)})} \right\},$$

and  $\text{dist}(\hat{\Phi}, \Phi_\star) \leq \min \left\{ \frac{1}{30} \frac{\lambda_{\min}^{\mathbf{F}}}{\sqrt{\lambda_{\max}^{\mathbf{F}}}} \frac{1}{T} \sum_{t=1}^T \kappa \left( \Sigma_x^{(t)} \right)^{-1}, 1/2 \right\}$ . Then, with probability at least  $1 - \delta$ , we have

$$\lambda_{\min} \left( \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right) \geq \frac{1}{2} \lambda_{\min}^{\mathbf{F}}.$$

*Proof:* as in Lemma A.10, we invert the bounds from Lemma A.6 for our desired factors of  $\lambda_{\min}^{\mathbf{F}}$ . By our proposed burn-in, we have with probability at least  $1 - \delta$

$$\lambda_{\max} \left( \text{Sym}(\Delta^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(E^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top) + \text{Sym}(\Delta^{(t)\top} E^{(t)}) \right) \leq \frac{1}{4} \lambda_{\min}^{\mathbf{F}}.$$

The only additional factor we account for here is the lower bound on  $\lambda_{\min} \left( \hat{\Phi} \Phi_\star^\top F_\star^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right)$ . We have

$$\begin{aligned} \lambda_{\min} \left( \hat{\Phi} \Phi_\star^\top F_\star^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right) &= \min_{\|x\|=1} x^\top \hat{\Phi} \Phi_\star^\top F_\star^{(t)\top} F_\star^{(t)} \Phi_\star \hat{\Phi}^\top x \\ &\geq \lambda_{\min} \left( F_\star^{(t)\top} F_\star^{(t)} \right) \min_{\|x\|=1} x^\top \hat{\Phi} \Phi_\star^\top \Phi_\star \hat{\Phi}^\top x \\ &= \lambda_{\min} \left( F_\star^{(t)\top} F_\star^{(t)} \right) \sigma_{\min}^2 \left( \Phi_\star \hat{\Phi}^\top \right). \end{aligned}$$

To further lower bound  $\sigma_{\min}^2(\Phi_* \hat{\Phi}^\top)$ , we observe that

$$\begin{aligned} \hat{\Phi} \hat{\Phi}^\top &= \hat{\Phi} (\Phi_*^\top \Phi_* + \Phi_{*,\perp}^\top \Phi_{*,\perp}) \hat{\Phi}^\top \\ \implies 1 &= \lambda_{\max}(\hat{\Phi} \hat{\Phi}^\top) \leq \lambda_{\min}(\hat{\Phi} \Phi_*^\top \Phi_* \hat{\Phi}^\top) + \lambda_{\max}(\hat{\Phi} \Phi_{*,\perp}^\top \Phi_{*,\perp} \hat{\Phi}^\top) \quad \text{Weyl's inequality} \\ \implies \sigma_{\min}^2(\Phi_* \hat{\Phi}^\top) &\geq 1 - \|\Phi_{*,\perp} \hat{\Phi}^\top\|^2 =: 1 - \text{dist}(\hat{\Phi}, \Phi_*)^2. \end{aligned}$$

Under our assumption that  $\text{dist}(\hat{\Phi}, \Phi_*) \leq \frac{1}{2}$ , we can piece together this bound with the bound on the cross terms to yield with probability at least  $1 - \delta$

$$\lambda_{\min}(\hat{F}^{(t)\top} \hat{F}^{(t)}) \geq \left(\frac{3}{4} - \frac{1}{4}\right) \lambda_{\min}(F_*^{(t)\top} F_*^{(t)}) = \frac{1}{2} \lambda_{\min}(F_*^{(t)\top} F_*^{(t)}),$$

which completes the proof.  $\blacksquare$

Piecing together Lemma A.10 and Lemma A.11, we get the following bound on the contraction factor:

**Lemma A.12** *Let*

$$N_1 \gtrsim \max \left\{ \gamma^4 (r + \log(1/\delta)), \frac{\lambda_{\max}^{\mathbf{F}}}{(\lambda_{\min}^{\mathbf{F}})^2} \frac{1}{T} \sum_{t=1}^T \frac{\sigma_w^{(t)2} (d_y + r + \log(1/\delta))}{\lambda_{\min}(\Sigma_x^{(t)})} \right\},$$

and  $\text{dist}(\hat{\Phi}, \Phi_*) \leq \min\{\frac{1}{30} \frac{\lambda_{\min}^{\mathbf{F}}}{\sqrt{\lambda_{\max}^{\mathbf{F}}}} \frac{1}{T} \sum_{t=1}^T \kappa(\Sigma_x^{(t)})^{-1}, 1/2\}$ . Then, for step-size satisfying  $\eta \leq \frac{1}{3\lambda_{\max}^{\mathbf{F}}}$ , with probability at least  $1 - \delta$ , we have

$$\left\| I_{d_x} - \eta \frac{1}{T} \sum_{t=1}^T \hat{F}^{(t)\top} \hat{F}^{(t)} \right\| \leq \left(1 - \frac{\eta}{2} \lambda_{\min}^{\mathbf{F}}\right).$$

### A.3 THE NON-IID SETTING

To extend our analysis to the non-iid setting, we first instantiate our covariates as  $\beta$ -mixing stationary processes (Yu, 1994; Kuznetsov and Mohri, 2017).

**Assumption A.1 (Geometric mixing)** *For each  $h$ , assume the process  $\{x_i^{(t)}\}_{t \geq 1}$  is a mean-zero stationary  $\beta$ -mixing process, with stationary covariance  $\Sigma_x^{(t)}$  and  $\beta(k) := \Gamma \mu^k$ .*

We note that exact stationarity is unnecessary as long as the marginal distributions converge to stationarity sufficiently fast; however, we assume exact stationarity for convenience. We now invoke the blocking technique on each trajectory, where each trajectory is subsampled into  $a$  trajectories of length  $m$  (we assume  $N = ma$  for notational convenience). We may then apply the analysis of the iid setting on a deflated dataset of  $Tm$  data points drawn from the respective stationary distributions to yield:

**Proposition A.3** *Let  $x_i^{(t)} \sim \nu_\infty^{(t)}$  for each  $t \in [T]$ ,  $i \in [N]$ . Assume  $|\mathcal{N}_1| \gtrsim \max \left\{ \gamma^2, \frac{\sigma_w^{(t)2}}{\lambda_{\min}(\Sigma_x^{(t)})} \frac{\|F_*\|^2}{\sigma_{\min}(F_*^{(t)})^4} \right\} (\max\{d_y, r\} + \log(H/\delta))$ , and  $\|\hat{\Phi} \Phi_{*,\perp}^\top\| \leq \min \left\{ \frac{3}{16} \frac{1}{\kappa(\Sigma_x^{(t)})+1} \frac{\sigma_{\min}(F_*^{(t)})^2}{\|F_*\|}, 1/2 \right\}$ . Let  $\eta \leq \min_t 1/4 \|F_*^{(t)}\|^2$ . There exists numerical constant  $C > 0$  such that the following holds: with probability at least  $1 - \delta$ ,*

$$\begin{aligned} \|\hat{\Phi}_+ \Phi_{*,\perp}^\top\| &\leq \left(1 - \frac{3\eta}{16} \frac{1}{T} \sum_{t=1}^T \sigma_{\min}(F_*^{(t)})^2\right)^{1/2} \|\hat{\Phi} \Phi_{*,\perp}^\top\| \\ &\quad + \max_t C \frac{\eta \sigma_w^{(t)} \|F_*^{(t)}\|}{\sqrt{|\mathcal{N}_2| m}} \sqrt{\frac{d_x + \log(H/\delta)}{\lambda_{\min}(\Sigma_x^{(t)})} \log\left(\frac{d_y + r}{\delta}\right)}. \end{aligned}$$

Now applying Lemma A.4, setting  $g(\cdot)$  as the indicator function for the burn-in requirement and the final descent bound, we have for all  $j = 1, \dots, a$ .

$$\left| \mathbb{E} \left[ g \left( \left\{ X_\infty^{(t), Nm} \right\}_{t \in [T]} \right) \right] - \mathbb{E} \left[ g \left( \left\{ X_{(j)}^{(t), NT} \right\}_{t \in [T]} \right) \right] \right| \leq m\beta(a) \leq \delta'$$

Setting  $\delta' = \delta/a$  and union bounding over each  $j = 1, \dots, a$ , we may invert  $N\beta(a) = \delta$  to find  $a := \log\left(\frac{\Gamma N}{\delta}\right) / \log\left(\frac{1}{\mu}\right)$ . This yields the final descent guarantee.

**Theorem A.1** *Let Assumption A.1 hold for the processes  $\{x_i^{(t)}\}_{i=1}^T$  for each  $i \in [N]$ ,  $t \in [T]$ . Define  $m := \frac{\log(1/\mu)T}{\log(\frac{\Gamma T}{\delta})}$ . Assume  $|\mathcal{N}_1| m \gtrsim \max \left\{ \gamma^2, \frac{\sigma_w^{(t)2}}{\lambda_{\min}(\Sigma_x^{(t)})} \frac{\|F_\star\|^2}{\sigma_{\min}(F_\star^{(t)})^4} \right\} (\max\{d_y, r\} + \log(H/\delta))$ , and  $\|\hat{\Phi}_\star^\top\| \leq \min \left\{ \frac{3}{16} \frac{1}{\kappa(\Sigma_x^{(t)})+1} \frac{\sigma_{\min}(F_\star^{(t)})^2}{\|F_\star\|}, 1/2 \right\}$ . Let  $\eta \leq \min_t 1/4 \|F_\star^{(t)}\|^2$ . There exists numerical constant  $C > 0$  such that the following holds: with probability at least  $1 - \delta$ ,*

$$\begin{aligned} \|\hat{\Phi}_\star^\top\| &\leq \left( 1 - \frac{3\eta}{16} \frac{1}{T} \sum_{t=1}^T \sigma_{\min}(F_\star^{(t)})^2 \right)^{1/2} \|\hat{\Phi}_\star^\top\| \\ &\quad + \max_t C \frac{\eta \sigma_w^{(t)} \|F_\star^{(t)}\|}{\sqrt{|\mathcal{N}_2| m}} \sqrt{\frac{d_x + \log(T/\delta)}{\lambda_{\min}(\Sigma_x^{(t)})} \log\left(\frac{d_y + r}{\delta}\right)} \\ &= \left( 1 - \frac{3\eta}{16} \frac{1}{T} \sum_{t=1}^T \sigma_{\min}(F_\star^{(t)})^2 \right)^{1/2} \|\hat{\Phi}_\star^\top\| \\ &\quad + \tilde{\mathcal{O}} \left( \max_t \frac{\eta \sigma_w^{(t)} \|F_\star^{(t)}\|}{\sqrt{|\mathcal{N}_2| T}} \sqrt{\frac{d_x + \log(T/\delta)}{\lambda_{\min}(\Sigma_x^{(t)})} \log\left(\frac{d_y + r}{\delta}\right)} \right). \end{aligned}$$

#### A.4 CONVERTING TO SAMPLE COMPLEXITY BOUNDS

To highlight the importance of the task scaling  $T$  in our descent guarantees, we demonstrate how to convert general descent lemmas to sample complexity guarantees.

**Lemma A.13** *For a sequence of positive integers  $\{M_k\}_{k \geq 1} \subset \mathbb{N}$ , define  $\{d_k\}_{k \geq 1} \subset \mathbb{R}_+$  as a sequence of non-negative real numbers dependent on  $\{M_k\}$  that satisfy the relation*

$$d_{k+1} \leq \rho \cdot d_k + \frac{C}{M_k},$$

for some  $\rho \in (0, 1)$  and  $C > 0$ . Let  $d_0 = \tau$ . Given a positive integer  $M$ , we may partition  $M = \sum_{k=1}^K M_k$ , where

$$K := \left\lceil \frac{1}{2} \log \left( \frac{2}{1+\rho} \right)^{-1} \frac{M\tau^2}{C} \left( \frac{1-\rho}{2} \right)^3 + 1 \right\rceil,$$

such that the following guarantee holds on  $d_K$ :

$$d_K \leq \tau \sqrt{\frac{2C}{M} \left( \frac{2}{1-\rho} \right)^3}.$$

The proof of Lemma A.13 follows by setting each  $M_k$  such that  $\rho \cdot d_k + \frac{C}{M_k} = \left(\frac{1+\rho}{2}\right) d_k$ , and setting  $K$  as the maximal  $K$  such that  $\sum_{k=1}^K M_k \leq M$ . Evaluating  $d_K \leq \tau \left(\frac{1+\rho}{2}\right)^K$  yields the result. For convenience, we do not consider burn-in times  $M_k \geq M_0 \forall k$  or pseudo-linear dependence  $\frac{C \text{polylog}(M_k)}{M_k}$ . However, these will only lead to inflating  $d_K$  by a  $\text{polylog}(M)$  factor.

In essence, Lemma A.13 demonstrates how a fixed offline dataset of size  $M$  can be partitioned into independent blocks of increasing size such that the final iterate satisfies an approximate ERM bound scaling as  $\frac{1}{\sqrt{M}}$ , inflated by a function of the contraction rate  $\rho$ . Instantiating Lemma A.13 with the problem parameters of Theorem 3.1 yields Corollary 3.1.

#### A.4.1 NEAR-ERM TRANSFER LEARNING

An important consequence of Lemma A.13 (thus Corollary 3.1) is that near-ERM parameter recovery bounds can be extracted. In particular, given some  $h \in [H + 1]$ , for a given representation  $\hat{\Phi}$ , and the least squares weights  $\hat{F}^{(t)}$  computed with respect to some independent dataset of size  $NT$ ,

$$\begin{aligned}
\left\| \hat{M}^{(t)} - M_\star^{(t)} \right\|_F^2 &= \left\| \hat{F}^{(t)} \hat{\Phi} - F_\star^{(t)} \Phi_\star \right\|_F^2 \\
&= \left\| \hat{F}^{(t)} \hat{\Phi} \begin{bmatrix} \hat{\Phi}^\top & \hat{\Phi}^\perp \end{bmatrix} - F_\star^{(t)} \Phi_\star \begin{bmatrix} \Phi_\star^\top & \Phi_\star^\perp \end{bmatrix} \right\|_F^2 \\
&= \left\| \begin{bmatrix} \hat{F}^{(t)} - F_\star^{(t)} \Phi_\star \hat{\Phi}^\top & -F_\star^{(t)} \Phi_\star \hat{\Phi}^\perp \end{bmatrix} \right\|_F^2 \\
&= \left\| \hat{F}^{(t)} - F_\star^{(t)} \Phi_\star \hat{\Phi}^\top \right\|_F^2 + \left\| F_\star^{(t)} \Phi_\star \hat{\Phi}^\perp \right\|_F^2 \\
&\leq 2 \left\| F_\star^{(t)} \Phi_\star \left( I_{d_x} - \hat{\Phi}^\top \hat{\Phi} \right) X^{(t)\top} Z^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \right\|_F^2 \\
&\quad + 2 \left\| W^{(t)\top} Z^{(t)} \left( \hat{\Sigma}_{z, \mathcal{N}_1}^{(t)} \right)^{-1} \right\|_F^2 + \left\| F_\star^{(t)} \right\|_F^2 \text{dist}(\hat{\Phi}, \Phi_\star)^2 \\
&\lesssim \left\| F_\star^{(t)} \right\|_F^2 \kappa \left( \Sigma_x^{(t)} \right) \text{dist}(\hat{\Phi}, \Phi_\star)^2 + \sigma_w^{(t)2} \frac{d_y r + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(t)}) NT} \quad \text{w.p. } \geq 1 - \delta,
\end{aligned}$$

where the last line follows from applying Lemma A.6. We observe that the parameter error nicely decomposes into a term quadratic in  $\text{dist}(\hat{\Phi}, \Phi_\star)$  and least squares fine-tuning error scaling with  $\frac{1}{NT}$ . For a fixed dataset of size  $NT$ , one can crudely set aside  $\Theta(N)$  samples for each task, and use the rest of the  $\Theta(N)$  samples to compute  $\hat{\Phi}$ . Invoking Corollary 3.1 and using the set-aside  $\Theta(N)$  samples to compute  $\hat{F}^{(t)}$  conditioned on  $\hat{\Phi}$ , we recover the near-ERM high probability generalization bound on the parameter error

$$\left\| \hat{M}^{(t)} - M_\star^{(t)} \right\|_F^2 \leq \tilde{O} \left( \left\| F_\star^{(t)} \right\|_F^2 \kappa \left( \Sigma_x^{(t)} \right) C(\rho) \frac{\max_t \sigma_w^{(t)2} d_x r}{NT} + \frac{\sigma_w^{(t)2} d_y r}{\lambda_{\min}(\Sigma_x^{(t)}) N} \right).$$

## B CASE STUDY: LINEAR DYNAMICAL SYSTEMS

To understand the importance of permitting non-isotropy and sequential dependence in multi-task data, we consider the fundamental setting of linear systems, which has served as a staple testbed for statistical and algorithmic analysis in recent years, since it lends itself to non-trivial yet tractable *continuous* reinforcement learning problems (see e.g., (Hu et al., 2023; Krauth et al., 2019; Tu and Recht, 2018; 2019; Recht, 2019; Fazel et al., 2018)), as well as (online) statistical learning problems with temporally dependent covariates (Abbasi-Yadkori et al., 2011; Abbasi-Yadkori, 2013; Simchowitz and Foster, 2020; Dean et al., 2018; 2020; Agarwal et al., 2019a;b; Ziemann et al., 2022; Ziemann and Sandberg, 2022; Lee et al., 2023; Simchowitz et al., 2018) (see (Tsiamis et al., 2022) for a tutorial and literature review). In particular for our purposes, the dependence of contiguous covariates in a linear system is intricately connected to its *stability properties* (Simchowitz et al., 2018; Jedra and Proutiere, 2020; Tu et al., 2022), such that we may instantiate the guarantees of DFW for non-iid data in an interpretable manner.

The standard state-space linear system set-up admits the form

$$\begin{aligned}
s[t + 1] &= A^{(h)} s[t] + B^{(h)} u[t] + w[t] \\
w[t] &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w^{(h)}), \quad s[0] \sim \mathcal{N}(0, \Sigma_0^{(h)}),
\end{aligned} \tag{27}$$

where we preemptively index possibly task-specific quantities. We consider the following two common linear system settings: system identification and imitation learning.



## B.1 LINEAR SYSTEM IDENTIFICATION

In linear system identification, the aim is to estimate the system matrices  $(A^{(h)}, B^{(h)})$  given state and input measurements  $s_t, u_t$ . In particular, we may cast the sysID problem as the following regression:

$$s[t+1] = [A^{(h)} \quad B^{(h)}] \begin{bmatrix} s[t] \\ u[t] \end{bmatrix} + w[t].$$

It is customary to consider exploratory signals that are iid zero-mean Gaussian random vectors  $u[t] \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u^{(h)})$  (Simchowit et al., 2018; Simchowit and Foster, 2020; Tsiamis et al., 2022). In the stable system case,  $\rho(A^{(h)}) < 1$ , we can therefore evaluate the covariance of the *stationary* distribution of states  $s[t]$  induced by exploratory signal  $u[t] \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u^{(h)})$  by plugging in (27) into the following equation

$$\begin{aligned} \mathbb{E}_{u,w}[s[t]s[t]^\top] &= \mathbb{E}_{u,w}[s[t+1]s[t+1]^\top] \\ &= A^{(h)}\mathbb{E}_{u,w}[s[t]s[t]^\top]A^{(h)\top} + B^{(h)}\mathbb{E}[u[t]u[t]^\top]B^{(h)\top} + \mathbb{E}[w[t]w[t]^\top] \\ &= A^{(h)}\mathbb{E}_{u,w}[s[t]s[t]^\top]A^{(h)\top} + B^{(h)}\Sigma_u^{(h)}B^{(h)\top} + \Sigma_w^{(h)} \end{aligned}$$

Therefore, evaluating the stationary state covariance  $\Sigma_s^{(h)} := \mathbb{E}[s[\infty]s[\infty]^\top]$  amounts to solving the Discrete Lyapunov Equation (dLYAP):

$$\Sigma_s^{(h)} := A^{(h)}\Sigma_s^{(h)}A^{(h)\top} + B^{(h)}\Sigma_u^{(h)}B^{(h)\top} + \Sigma_w^{(h)}.$$

In the notation introduced earlier in the paper, casting  $y[t] \leftarrow s[t+1]$ ,  $x[t] \leftarrow \begin{bmatrix} s[t] \\ u[t] \end{bmatrix}$ ,  $M_\star^{(h)} \leftarrow [A^{(h)} \quad B^{(h)}]$ , we may instantiate multi-task linear system identification as a non-iid, non-isotropic linear operator recovery problem.

**Definition B.1** *Let the initial state covariance be the stationary covariance  $\Sigma_0^{(h)} = \Sigma_s^{(h)}$ , such that the covariance of the marginal covariate distribution satisfies*

$$\Sigma_x^{(h)} := \mathbb{E}[x[t]x[t]^\top] = \begin{bmatrix} \Sigma_s^{(h)} & 0 \\ 0 & \Sigma_u^{(h)} \end{bmatrix}, \text{ for all } t \geq 0.$$

We make the above standard definition for the initial state distribution for convenience, as it ensures the marginal distributions of each state are identical. We note, however, given a different initial state distribution, the marginal state distribution converges exponentially quickly to stationarity, thus accumulating only a negligible factor to the final rates. We then make the following system assumptions to instantiate our representation learning guarantees.

**Assumption B.1** *We assume that for any task  $h$  the following hold:*

1. *The operators share a rowspace  $M_\star^{(h)} := [A^{(h)} \quad B^{(h)}] = F_\star^{(h)}\Phi_\star$ ,  $F_\star^{(h)} \in \mathbb{R}^{d_s \times r}$ ,  $\Phi_\star \in \mathbb{R}^{r \times (d_s + d_u)}$ .*
2. *The state matrices have uniformly bounded spectral radii  $\rho(A^{(h)}) < \mu < 1$ . Subsequently, we assume there exists a constant  $\Gamma' > 0$  that satisfies*

$$\|A^{(h)k}\|_2 \leq \Gamma' \mu^k, \text{ for all } k \geq 0.$$

*The existence of a uniform  $\Gamma'$  is guaranteed by Gelfand's Formula (Horn and Johnson, 2012), and quantitative bounds may be found in, e.g., Goldenshluger and Zeevi (2001); Tu and Recht (2018).*

The first assumption is satisfied, for example, when  $A^{(h)} = P_\star^{(h)}U_\star$  and  $B^{(h)} = Q_\star^{(h)}V_\star$  individually admit low-rank decompositions. The second assumption translates to a quantitative bound on the mixing time of the covariates  $x[t]$  by adapting a result from Tu and Recht (2018).

**Proposition B.1 (Adapted from Tu and Recht (2018, Prop. 3.1))** For each  $h$ , let the dynamics for the linear system evolve as described in (27). Let Assumption B.1 hold with constants  $\Gamma'$ ,  $\rho$ . Define  $\mathbb{P}_{s[k] \sim \nu_k} [\cdot | s_0 = s]$  as the conditional distribution of states  $s[k]$  given initial condition  $s_0 = s$ . We have for any  $k \geq 0$  and initial state distribution  $\nu_0$ ,

$$\mathbb{E}_{s \sim \nu_0} [\|\mathbb{P}_{s[k]}[\cdot | s_0 = s] - \mathbb{P}_{s[k]}\|_{\text{tv}}] \leq \frac{\Gamma'}{2} \sqrt{\mathbb{E}_{\nu_0} [\|s[0]\|^2] + \frac{\|\Sigma^{-1}\|_*}{1 - \mu^2} \cdot \mu^k}, \quad (28)$$

where  $\|\cdot\|_*$  indicates the nuclear norm (Horn and Johnson, 2012), and  $\Sigma := B^{(h)} \Sigma_u^{(h)} B^{(h)\top} + \Sigma_w^{(h)}$ .

We note that by the independence of control inputs  $u[t]$ , we have trivially that the total variation distance between the conditional and marginal distributions of covariates  $x[t]$  is the same as that of the states  $s[t]$ .

$$\|\mathbb{P}_{s[k]}[\cdot | s_0 = s] - \mathbb{P}_{s[k]}\|_{\text{tv}} = \|\mathbb{P}_{x[k]}[\cdot | s_0 = s] - \mathbb{P}_{x[k]}\|_{\text{tv}}$$

Since by construction the marginal distribution of states is identically  $\mathcal{N}(0, \Sigma_s^{(h)})$ , applying Proposition B.1 to  $s[t]$ ,  $s[t+k]$  for any  $t, k$ , we get the following quantitative bound on the mixing-time of the covariates  $x[t] = [s[t]^\top \quad u[t]^\top]^\top$ .

**Lemma B.1** Following Definition B.1 and Assumption B.1, the covariate process  $\{x^{(h)}[t]\}_{t \geq 0}$  is a mean-zero, stationary, geometrically  $\beta$ -mixing process with covariance  $\Sigma_x^{(h)} = \begin{bmatrix} \Sigma_s^{(h)} & 0 \\ 0 & \Sigma_u^{(h)} \end{bmatrix}$ ,

where  $\Sigma_s^{(h)} = \text{dlyap}(A^{(h)}, B^{(h)} \Sigma_u^{(h)} B^{(h)} + \Sigma_w^{(h)})$ , and mixing-time bounded by

$$\beta(k) = \Gamma \mu^k, \quad \text{where} \quad \Gamma := \frac{\Gamma'}{2} \sqrt{\text{Tr}(\Sigma_s^{(h)}) + \frac{\|\Sigma^{-1}\|_*}{1 - \mu^2}}, \quad \Sigma := B^{(h)} \Sigma_u^{(h)} B^{(h)\top} + \Sigma_w^{(h)}. \quad (29)$$

Thus, instantiating Lemma B.1 in Theorem A.1 gives us guarantees of DFW applied to multi-task linear system identification.

## B.2 IMITATION LEARNING

In linear (state-feedback) imitation learning (IL), the aim is to estimate linear state-feedback controllers  $K^{(h)} \in \mathbb{R}^{d_u \times d_x}$  from (noisy) state-input pairs  $\{(s[t], u[t])\}_{t \geq 0}$  induced by unknown expert controllers  $K_\star^{(h)}$ . In particular, we assume the expert control inputs are generated as

$$u[t] = K_\star^{(h)} s[t] + z[t], \quad z[t] \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_z^{(h)}),$$

which we observe lends itself naturally as a linear regression, casting  $y[t] \leftarrow u[t]$ ,  $x[t] \leftarrow s[t]$ ,  $M_\star^{(h)} \leftarrow K_\star^{(h)}$ . Plugging the expert control inputs into the dynamics (27) yields that the states/covariates evolve as

$$\begin{aligned} s[t+1] &= A^{(h)} s[t] + B^{(h)} \left( K_\star^{(h)} s[t] + z[t] \right) + w[t] \\ &= \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right) s[t] + B z[t] + w[t]. \end{aligned}$$

We make the natural assumption that the expert controller  $K_\star^{(h)}$  stabilizes the system, i.e. the spectral radius of the closed-loop dynamics has spectral radius strictly less than 1:  $\rho \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right) < 1$ . As such, similar to the linear sysID setting, we may plug the above dynamics into the stationarity equation to yield the stationary covariance:

$$\begin{aligned} \mathbb{E}[s[t]s[t]^\top] &= \mathbb{E}[s[t+1]s[t+1]^\top] \\ &= \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right) \mathbb{E}[s[t]s[t]^\top] \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right)^\top + B^{(h)} \Sigma_z^{(h)} B^{(h)\top} + \Sigma_w^{(h)} \\ \implies \Sigma_s^{(h)} &= \text{dlyap} \left( A^{(h)} + B^{(h)} K_\star^{(h)}, B^{(h)} \Sigma_z^{(h)} B^{(h)\top} + \Sigma_w^{(h)} \right). \end{aligned}$$

Analogously to linear sysID, we make the following assumptions.

**Assumption B.2** We assume that for any task  $h$  the following hold:

1. The initial state covariance is set to the stationary covariance  $\Sigma_0^{(h)} = \Sigma_s^{(h)}$ , such that the marginal covariate distributions satisfy

$$\mathbb{E} [x[t]x[t]^\top] = \Sigma_s^{(h)} =: \Sigma_x^{(h)}, \text{ for all } t \geq 0.$$

2. The controllers share a rowspace  $M_\star^{(h)} \equiv K_\star^{(h)} = F_\star^{(t)} \Phi_\star$ ,  $F_\star^{(t)} \in \mathbb{R}^{d_u \times r}$ ,  $\Phi_\star \in \mathbb{R}^{r \times d_s}$ .
3. The closed-loop dynamics have uniformly bounded spectral radii  $\rho \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right) < \mu < 1$ . Subsequently, we assume there exists a constant  $\Gamma' > 0$  that satisfies

$$\left\| \left( A^{(h)} + B^{(h)} K_\star^{(h)} \right)^k \right\|_2 \leq \Gamma' \mu^k.$$

The existence of uniform  $\Gamma'$  is guaranteed by Gelfand's Formula (Horn and Johnson, 2012).

By using a result almost identical to Proposition B.1, we yield the following quantitative bound on the mixing time of covariates generated by stabilizing expert controllers.

**Lemma B.2** Following Assumption B.2, the covariate process  $\{x^{(h)}[t]\}_{t \geq 0}$  is a mean-zero, stationary, geometrically  $\beta$ -mixing process with covariance  $\Sigma_x^{(h)} = \Sigma_s^{(h)}$ , where  $\Sigma_s^{(h)} = \text{dlyap} \left( A^{(h)} + B^{(h)} K_\star^{(h)}, B^{(h)} \Sigma_z^{(h)} B^{(h)\top} + \Sigma_w^{(h)} \right)$ , and mixing-time bounded by

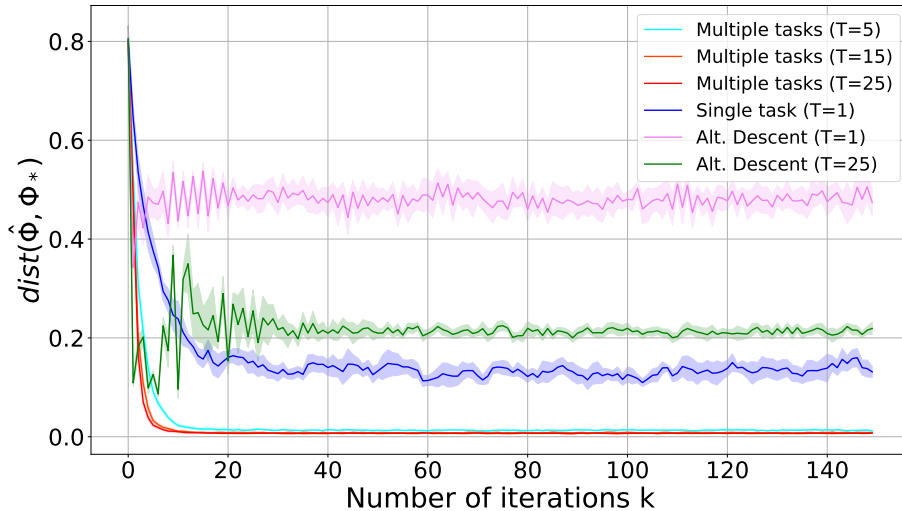
$$\beta(k) = \Gamma \mu^k, \text{ where} \quad \Gamma := \frac{\Gamma'}{2} \sqrt{\text{Tr} \left( \Sigma_s^{(h)} \right) + \frac{\|\Sigma^{-1}\|_*}{1 - \mu^2}}, \quad \Sigma := B^{(h)} \Sigma_z^{(h)} B^{(h)\top} + \Sigma_w^{(h)}. \quad (30)$$

Thus, instantiating Lemma B.1 in Theorem A.1 gives us guarantees of DFW applied to multi-task linear imitation learning.

## C ADDITIONAL NUMERICAL EXPERIMENTS AND DETAILS

We present additional numerical experiments to demonstrate the effectiveness of DFW (Algorithm 1) and provide a more detailed explanation of the task-generating process for constructing random operators in linear regression and system identification examples. Furthermore, we introduce an additional setting, imitation learning, to illustrate the advantages of collaborative learning across tasks in learning a linear quadratic regulator by leveraging expert data to compute a shared common representation across all tasks. In this latter setting, we also emphasize the importance of feature whitening when dealing with non-i.i.d. and non-isotropic data.

- **Random rotation:** For all the numerical experiments presented in this paper, the application of a random rotation around the identity is employed for both task-specific weight generation and the initialization of the representation. This random rotation is defined as  $R_{\text{rot}} = \exp(\tilde{L})$ , where  $\tilde{L} = \frac{L - L^\top}{2}$  and  $L = \gamma S$ . Here,  $S$  is a random matrix with entries drawn from a standard normal distribution,  $d_l$  is the corresponding dimension of the high-dimensional latent space, and  $\gamma$  corresponds to the scale of the rotation. We set  $\gamma = 0.01$  for generating different task weights and  $\gamma = 1$  for initializing the representation.
- **Step-sizes:** The step-size  $\eta$  used to update the common representation is carefully selected to ensure a fair comparison between Algorithm 1 and the vanilla alternating minimization-descent approach employed in FedRep Collins et al. (2021). For example, to obtain the results depicted in Figure ??, we set  $\eta = 7, 5 \times 10^{-3}$ , while for the alternating minimization-descent approach, which demonstrates better performance with a smaller step-size, we set  $\eta = 5 \times 10^{-5}$  to achieve the results presented in Figure ?. In Figure 1a, both the single-task and multi-task implementations of Algorithm 1 adopt  $\eta = 7, 5 \times 10^{-3}$ , whereas the vanilla



**Figure 2:** We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 and the multi-task FedRep for the IID linear regression with random covariance. We observe performance improvement and variance reduction for multi-task DFW as predicted.

alternating minimization-descent approach uses  $\eta = 7.5 \times 10^{-3}$  for a fair comparison. Similarly, in Figure 1c, both the single-task and multi-task versions of Algorithm 1 use  $\eta = 1 \times 10^{-1}$ , while the vanilla alternating minimization-descent approach utilizes  $\eta = 2 \times 10^{-3}$ .

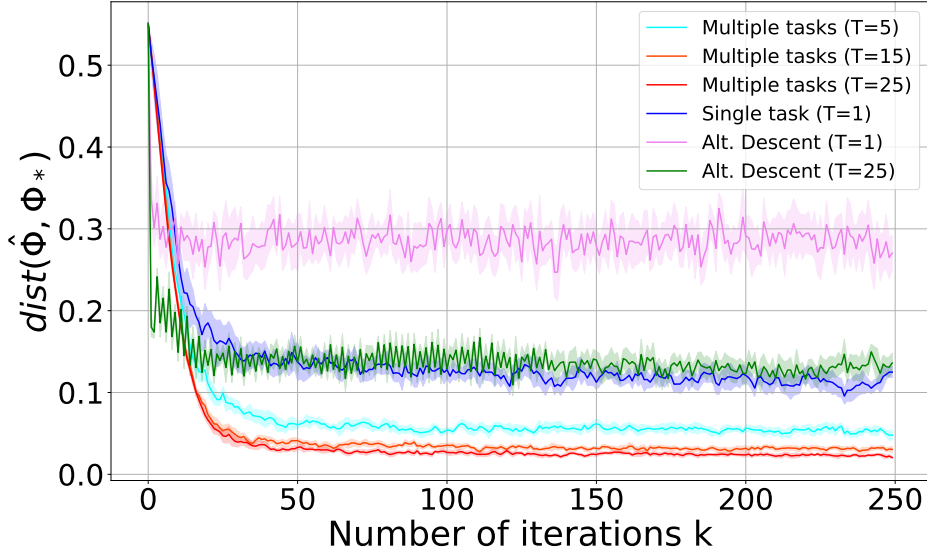
### C.1 LINEAR REGRESSION WITH IID AND NON-ISOTROPIC DATA

Continuing our experiments for the linear regression problem, this time with different random linear operators as illustrated in Figure 1c, we present the results for an extended range of tasks using Algorithm 1 and the alternating minimization-descent approach (FedRep Collins et al. (2021)). In this analysis, we utilize the same specific parameters as discussed in §4. Additionally, we set the step-size  $\eta = 7.5 \times 10^{-3}$  for both the single-task and multi-task implementations of Algorithm 1, and  $\eta = 7.5 \times 10^{-5}$  for both the single-task and multi-task alternating minimization-descent.

Figure 2 presents a comparison of the performance between Algorithm 1 and the vanilla alternating minimization approach in both single and multi-task settings. In line with our theoretical results, the figure demonstrates that as the number of tasks  $T$  increases, the error between the current representation and the ground truth representation significantly diminishes. In the specific case of linear regression with iid and non-isotropic data, this figure emphasizes that a small number of tasks ( $T = 5$ ), is sufficient to achieve a low error in computing a shared representation across the tasks. Furthermore, the depicted figure reveals that while the multi-task alternating descent algorithm outperforms the single-task case, it is worth noting that this algorithm remains sub-optimal and is unable to surpass the limitation imposed by the presence of bias in the non-isotropic data. Despite its improved performance, the multi-task alternating descent algorithm still encounters challenges in overcoming the inherent noise barrier.

### C.2 SYSTEM IDENTIFICATION

Building upon the results presented in §4, we conduct an extended experiment involving a larger range of tasks while maintaining the parameters specified in §4.2. Specifically, we generate distinct random



**Figure 3:** We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 multi-task FedRep for the linear system identification with random covariance. We observe performance improvement and variance reduction for multi-task DFW as predicted.

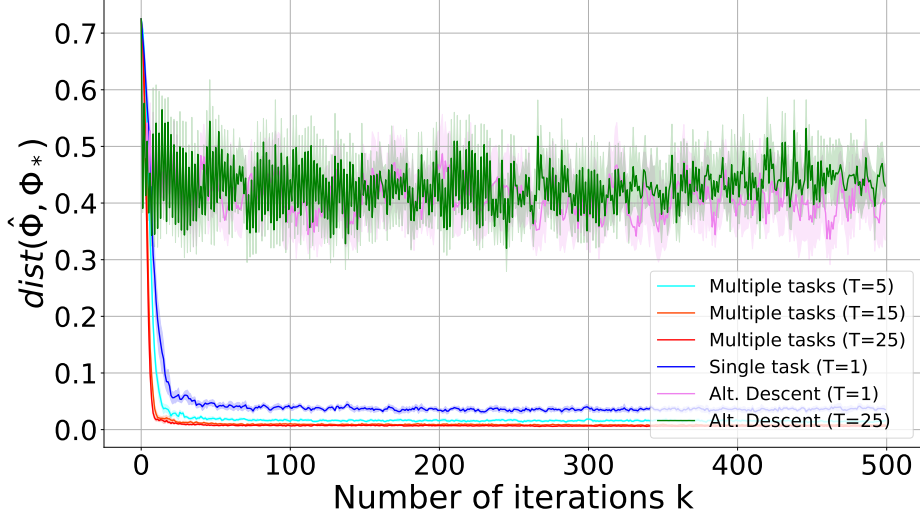
operators different from those utilized to obtain the results illustrated in Figure 1c. In this current analysis, we present the outcomes for the expanded range of tasks using Algorithm 1 and compare them to the single-task and multi-task vanilla alternating minimization-descent algorithms. The step-size  $\eta$  is set to  $1 \times 10^{-1}$  for both the single-task and multi-task implementations of Algorithm 1, while for the single-task and multi-task vanilla alternating minimization-descent algorithms, we set  $\eta$  to  $2 \times 10^{-3}$ .

In alignment with our main theoretical findings, Figure 3 provides compelling evidence regarding the advantages of the proposed algorithm (Algorithm 1) compared to the vanilla alternating descent approach when computing a shared representation for all tasks. Consistent with the trend observed in Figure 2 for the linear regression problem, Figure 3 illustrates a significant reduction in the error between the current representation and the ground truth representation as the number of tasks increases. Additionally, it is noteworthy that while the multi-task alternating descent outperforms the single-task scenario, the single-task variant of Algorithm 1 achieves even better results. This observation underscores the importance of incorporating de-biasing and feature-whitening techniques when dealing with non-iid and non-isotropic data.

### C.3 IMITATION LEARNING

Our focus now turns to the problem of learning a linear quadratic regulator (LQR) controller, denoted as  $K^{(T+1)} = F_*^{(T+1)} \Phi_*$ , by imitating the behavior of  $T$  expert controllers  $K^{(1)}, K^{(2)}, \dots, K^{(T)}$ . These controllers share a common low-rank representation and can be decomposed into the form  $K^{(t)} = F_*^{(t)} \Phi_*$ , where  $F_*^{(t)}$  represents the task-specific weight and  $\Phi_*$  corresponds to the common representation across all tasks. To achieve this, we exploit Algorithm 1 to compute a shared low-rank representation for all tasks by leveraging data obtained from the expert controllers. Within this context, we consider a discrete-time linear time-invariant dynamical system as follows:

$$x^{(t)}[i+1] = Ax^{(t)}[i] + Bu^{(t)}[i], \quad i = 0, 1, \dots, N-1,$$



**Figure 4:** We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 and the multi-task FedRep for the imitation learning with random covariance. We observe performance improvement and variance reduction for multi-task DFW as predicted.

with  $n_x = 4$  states and  $n_u = 4$  inputs, for all  $t \in [T + 1]$ , where  $u^{(t)}[i] = K^{(t)}x^{(t)}[i] + z^{(t)}[i]$ , with  $z^{(t)}[i] \sim \mathcal{N}(0, I_{n_u})$  being the input noise. In our current setting, rather than directly observing the state, we obtain a high-dimensional observation derived from an injective linear function of the state. Specifically, we assume that  $y^{(t)}[i] = Gx^{(t)}[i] + w^{(t)}[i]$ , where  $G \in \mathbb{R}^{25 \times 4}$  represents the high-dimensional linear mapping. The injective linear mapping matrix  $G$  is generated by applying a `thin_svd` operation to a random matrix with values drawn from a normal distribution  $\mathcal{N}(0, 1)$ . This process ensures injectiveness with a high probability.

For this aforementioned multi-task imitation learning setting, we adopt a scheme in which we gather observations of the form  $\{\{y^{(t)}[i], u^{(t)}[i]\}_{i=0}^{N-1}\}_{t=1}^T$  from the initial  $T$  expert controllers to learn the controller  $K^{(T+1)}$ . These observations are obtained by following the dynamics:

$$y^{(t)}[i] = (\tilde{A} + \tilde{B}\tilde{K}^{(t)})y[i] + \tilde{B}z^{(t)}[i] + w^{(t)}[i]$$

with  $\tilde{A} = GAG^\dagger$ ,  $\tilde{B} = GB$ ,  $\tilde{K}^{(t)} = K^{(t)}G^\dagger$ , and process noise  $w^{(t)}[i] \sim \mathcal{N}(0, \Sigma_w)$ .

The collection of stabilizing LQR controllers  $K^{(1)}, K^{(2)}, \dots, K^{(T+1)}$  is generated by assigning different cost matrices, namely  $R = \frac{1}{4}I_{n_u}$  and  $Q^{(t)} = \alpha^{(t)}I_{n_x}$ , where  $\alpha^{(t)} \in \text{logspace}(0, 3, H)$ . These matrices are then utilized to solve the Discrete Algebraic Riccati Equation (DARE):  $P^{(t)} = A^\top P^{(t)}A + A^\top P^{(t)}B(B^\top P^{(t)}B + R)^{-1}B^\top P^{(t)}A + Q^{(t)}$ , and compute  $K^{(t)} = -(B^\top P^{(t)}B + R)^{-1}B^\top P^{(t)}A$ , for all  $t \in [T + 1]$ . Moreover, the system matrices  $A$  and  $B$  are randomly generated, with elements drawn from a uniform distribution. The trajectory length  $N = 75$  remains consistent for all tasks. The shared representation is initialized by applying a random rotation to the true representation, denoted as  $\Phi_* = G^\dagger$ .

Figure 4 presents a comparative analysis between Algorithm 1 and the vanilla alternating minimization-descent approach (FedRep in Collins et al. (2021)) for computing a shared representation across linear quadratic regulators. This shared representation is then utilized to derive the learned controller  $K^{(T+1)}$  in a few-shot learning manner. Consistent with our theoretical findings and in alignment with the trends observed in Figures 2-3, Figure 4 demonstrates a substantial reduction in the error between the current representation and the ground truth representation when leveraging

data from multiple tasks, compared to the single-task scenario in Algorithm 1. Furthermore, this figure underscores the significance of de-biasing and whitening the feature data in overcoming the bias barrier introduced by non-iid and non-isotropic data. In contrast, the vanilla alternating descent algorithm fails to address this challenge adequately and yields sub-optimal solutions.